UNIVERSITY OF CALIFORNIA

Los Angeles

# Human-AI Systems for Video Accessibility

A thesis submitted in partial satisfaction

of the requirements for the degree

Master of Science in Electrical and Computer Engineering

by

Xingyu Liu

2022

© Copyright by
Xingyu Liu
2022

ABSTRACT OF THE THESIS

Human-AI Systems for Video Accessibility

by

Xingyu Liu
Master of Science in Electrical and Computer Engineering
University of California, Los Angeles, 2022
Professor Xiang Chen, Chair

Online video content has proliferated as a key source for information in the form of video lectures, vlogs, reviews, how-to's, and more. A video hosting service, YouTube, is now the second most popular search platform and reaches 81% of internet users under 20. However, most online videos do not provide audio descriptions (AD), presenting serious barriers for blind and visually impaired (BVI) people who may lack access to the visuals. In this thesis work, I present Human-AI systems to make online videos more accessible. Specifically, by (1) understanding what makes videos accessible to blind and visually impaired people; (2) automatically surfacing already accessible video content on platforms like YouTube with computer vision and natural language processing algorithms; and (3) helping video authors efficiently detect and address visual and auditory accessibility issues in their video authoring process.

The thesis of Xingyu Liu is approved.

Bolei Zhou

Yang Zhang

Xiang Chen, Committee Chair

University of California, Los Angeles

2022

*To my mother, Guizhen Du*

*and my father, Zeli Liu*

*for their unconditional love and support*

TABLE OF CONTENTS

**1 Introduction** . . . . . **1**

**2 What Makes Videos Accessible to Blind and Visually Impaired People?** **4**

2.1 Introduction . . . . . 4

2.2 Related Work . . . . . 7

2.2.1 Web accessibility metrics and search . . . . . 7

2.2.2 Video accessibility . . . . . 8

2.2.3 Searching and browsing online videos . . . . . 9

2.3 Formative Interviews and Co-watching Exercise . . . . . 10

2.3.1 Methods . . . . . 10

2.3.2 Results . . . . . 12

2.4 Video accessibility metrics . . . . . 14

2.4.1 Audio-related . . . . . 15

2.4.2 Visual-related . . . . . 18

2.4.3 Audio-visual references . . . . . 19

2.4.4 Research Questions . . . . . 21

2.5 Evaluation: How Metrics Indicate Accessibility Ratings . . . . . 22

2.5.1 BVI Video Accessibility Ratings . . . . . 22

2.5.2 Regression Model . . . . . 24

2.6 Evaluation: A Video Search Interface Augmented with Accessibility Metrics 26

2.6.1 Findings . . . . . 28

2.7 Discussion and Future Work . . . . . 33

2.8 Conclusion . . . . . 35

**3 Understanding How Blind and Visually Impaired People Leverage Accessibility Metrics In Practice . . . . . 37**

3.1 Introduction . . . . . 37

3.2 Related Work . . . . . 39

3.2.1 Studying Information Seeking In-The-Wild . . . . . 39

3.2.2 Searching and Browsing on YouTube . . . . . 40

3.3 Implementation . . . . . 41

3.3.1 Computing Video Accessibility Metrics . . . . . 42

3.3.2 Video Accessibility Metrics Extension . . . . . 43

3.4 Deployment Study . . . . . 46

3.4.1 Participants . . . . . 46

3.4.2 Procedure . . . . . 46

3.4.3 Data . . . . . 48

3.4.4 Limitations . . . . . 48

3.5 Findings . . . . . 49

3.5.1 Overall Usage . . . . . 49

3.5.2 RQ1: Do People Consider Accessibility Metrics in Practice? . . . . . 51

3.5.3 RQ2: What aspects of accessibility are most important to people in practice? . . . . . 55

3.5.4 RQ3: How does knowledge of accessibility metrics impact people's browsing and searching behavior? . . . . . 58

3.6 Discussion . . . . . 62

3.6.1 Content vs. Accessibility . . . 62
3.6.2 Perceivable Accessibility vs. Full Accessibility . . . 63
3.6.3 Exploitation vs. Exploration . . . 64
3.7 Conclusion . . . 65

**4 Identifying Video Accessibility Issues via Cross-modal Grounding . . . 66**
4.1 Introduction . . . 66
4.2 Related Work . . . 68
4.2.1 Authoring AD and CC . . . 68
4.2.2 Assessing Audio and Visual Similarity . . . 69
4.3 CrossA11y Interface . . . 70
4.3.1 Video Pane . . . 70
4.3.2 Video Description Pane . . . 72
4.3.3 Captions Pane . . . 73
4.3.4 Accessible Video Preview . . . 73
4.4 Cross-modal Grounding Pipeline . . . 74
4.4.1 Segmentation . . . 74
4.4.2 Cross-modal Grounding . . . 75
4.4.3 Post-processing . . . 78
4.4.4 Technical Evaluation . . . 79
4.4.5 Limitations . . . 81
4.5 Evaluations: Can CrossA11y Users Efficiently Identify Video Accessibility Issues? . . . 82
4.5.1 Materials . . . 82

4.5.2 Participants . . . . . 83
4.5.3 Procedure . . . . . 83
4.5.4 Findings . . . . . 84
4.6 Case Study: Usage of CrossA11y by Content Creators . . . . . 91
4.6.1 Participants . . . . . 91
4.6.2 Findings . . . . . 91
4.7 Discussion and Future Work . . . . . 93
4.8 Conclusion . . . . . 95

**5 Conclusion** . . . . . **96**

**References** . . . . . **98**

LIST OF FIGURES

2.1 Video search results (A) contain no information about whether or not the video is accessible. People must use trial and error to find an accessible video. Our system calculates *video accessibility metrics* (B, left) informed by BVI formative study participants, and predicts the overall non-visual accessibility of the video (B, right). BVI people using our system can preview the accessibility score and explanation (C) to filter or quickly identify accessible videos from search results. 5

2.2 The accessibility of a video depends on what information is in the audio track (blue, left), what information is in the video track (yellow, right), and what information is redundant between the audio and visual channels (green, middle). In (A) more information is contained in the audio than the visuals, and most of the information in the visual content is also in the audio. In (B) more information is contained in the audio than the visuals, but little of the visual information is explained. In (C) more information is conveyed visually, but some of this information is also covered in the audio. In (D) more of the information is conveyed visually, with little of this information covered in the audio. . . . . 15

3.1 (A) Accessibility metrics that our system computed and displayed onto YouTube. It is displayed as a heading level 4 element next to each video title (heading level 3). Screen reader users can quickly navigate to accessibility metrics with headings. (B) More accessibility information that are related to visual (*e.g.*, visual changes, description of visual objects) are computed upon request, and will be automatically added to the interface. (C) Extension's popup page in which people can turn on/off settings, leave feedback, and download their log data for this study. We highlight parts that are generated by our extension on YouTube in red. . . . . 41

3.2 Participants' Likert-scale ratings (From 1 - not important at all to 5 - very important) to the question, *"Rate the importance of factors that made you decide to click on and watch the video."* Participants rated title, accessibility (default) and percentage of speech, to be the top-3 most important reasons. . . . . 51

3.3 Boxplots of accessibility scores of videos participants click into from different source (search, homepage) in different stage 1, stage 2 and stage 3. We observe that participants clicked into more accessible videos from search (semi-specific) with accessibility metrics (stage 2). However when selecting videos from homepage (open-ended), accessibility of videos stays around the same. . . . . 52

3.4 Histogram of accessibility scores (from 1 - not accessible at all to 7 - very accessible) for all selected videos when not having accessibility metrics available. People were already picking more accessible videos even without our extension. . . . . 58

3.5 Percentage of video foraging approach (page visits) of primary homepage and search users in stage 1, 2, 3 of the study. With accessibility metrics people tend to explore new approaches to find videos. . . . . 61

4.1 Our system (A) identifies accessibility issues by locating *modality asymmetries* (in red) between audio segments and video segments using cross-modal grounding. (B) lets authors address accessibility issues using the CrossA11y interface to write captions and video descriptions, and (C) creates a more accessible video from the authored descriptions. . . . . 67

4.2 In CrossA11y's interface, the *video pane* (A) displays audio and visual timelines with accessibility visualization that allows authors to quickly identify and navigate to accessibility issues. The *video description pane* (F) surfaces inaccessible visual segments and lets authors to add descriptions. The *captions pane* (E) provides time-aligned captions and detected non-speech sound segments for authors to seek within the video and add captions. . . . . 71

4.3 Our computational pipeline includes: (A) Segmentation that segments the video into audio and visual segments, (B) Cross-modal Grounding that finds correspondences between visual and audio segments, and (C) post-processing that filters identified correspondences. . . . . 74

4.4 Example of visual to audio similarity matrix, and visual to text (transcript) similarity matrix. Red dotted lines highlight examples of asymmetric and potentially inaccessible segments. . . . . 75

4.5 Participants' ratings to task load index questions (on a scale of 1-low to 7-high) for their experience adding AD/CC to videos with (1) or without (0) CrossA11y. 84

LIST OF TABLES

2.1 Properties of accessible and inaccessible videos as reported by formative study participants. We include the number of participants who mentioned each property during interviews (#P), and the number of times an issue was mentioned during co-watching exercises (#M). . . . . 12

2.2 7 accessibility heuristics and corresponding metrics, along with their distributions of the 60 video samples we used in regression analysis. Median is colored blue in histograms. Also shows accessible and inaccessible examples of these metrics. . 16

3.1 Demographic and YouTube experience information of our study participants. 10 participants completed our 4-week long deployment. P2, P7 and P10 did not complete the study and are thus not included in this table. . . . . 45

4.1 CrossA11y's experimental test results on a sample of 20 manually labeled videos. CrossA11y generally performs better than random guess and using "gaps in speech" as heuristics, for both detecting visual and auditory accessibility issues. 79

# ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my advisor, Xiang 'Anthony' Chen, for his guidance and support throughout my Master's studies. I would also like to thank my committee members, Yang Zhang and Bolei Zhou, for their valuable feedback and suggestions. None of the projects in this thesis would have been possible without the contribution of my awesome collaborators: Amy Pavel, Dingzeyu Li, Patrick Carrington, and Ruolin Wang. I am grateful to Rajan Vaish, Brian A. Smith and Ruofei Du, who are my advisors during my internships. I would also like to thank my friends who have supported me throughout my studies, including but not limited to, Jiaqi Zou, Shiqi Chou, and my significant other, Qingyi Dong. Last but not least, I would like to thank my parents for their love and support.

# CHAPTER 1

# Introduction

For decades, text-based webpages such as encyclopedic articles, text reviews, how-to instructions, blogs, tourism sites, and news reports were the primary source of information online. Web accessibility guidelines and evaluation techniques then centered around the parsing, navigation, and presentation of text content and the presence of text alternatives for non-text content. Recently web-based video content has proliferated as a new key source for information in the form of explainer videos, lectures, unboxings and reviews, how-to's, vlogs, trip reports, commentary, news and more. A video hosting service, YouTube.com, is now the second most popular search platform [106], the second most used mobile application [5], and reaches 81% of internet users under 25 [106], yet it does not require or provide alternative text descriptions, inline audio descriptions, or extended audio descriptions for the video content (criteria for WCAG 2.1 A, AA, and AAA respectively [19]), presenting potentially serious barriers for blind and visually impaired Internet users who may lack access to the visual content in videos encountered online.

Despite the prevalence of potentially inaccessible video content, blind and visually impaired (BVI) participants in our formative interviews (Chapter 2) watched online videos regularly, finding videos related to their interests via search, recommendation, or external links, and consuming videos for the purposes of entertainment, learning new things, and maintaining social connectedness, similar to studies of the general population [105]. For example, BVI people reported often encountering videos that only conveyed important information via the visual modality (*e.g.*a recipe video where the audio track is just background

music), causing lack of access to a wide range of videos online. Thus, it is important for video sharing platforms and video authors to produce and provide accessible video content.

However, several key challenges remains in making online, user-generated videos more accessible. Unlike traditional media (TV and movies) where professionally produced audio descriptions are becoming increasingly common on streaming services [74], audio description for user-generated videos is exceedingly rare due to many factors including the expertise typically required to create descriptions, the sheer amount of content (around 730,000 hours of videos being uploaded to YouTube everyday), a lack of platform support, and insufficient awareness education.

The goal of my thesis research, is to *make videos and video platforms more accessible with human-AI systems*. Recent advancement in computer vision (CV), natural language processing (NLP) and multimodal machine learning opens up new opportunities to automatically analyze visual and auditory characteristics of videos, and further assist people in identifying and addressing accessibility issues more efficiently. I design, develop and evaluate human-AI systems that help different users (content creators, BVI audience) to improve video accessibility at different stages of video production (authoring, consuming).

These goals are manifested in the following three projects, as described in the remainder of this thesis:

- **Chapter 2** presents an ML-driven interface that leverages visual and auditory features of videos to automatically surface inherently accessible videos for BVI users. We developed a set of automatic video accessibility metrics correlated with BVI people's accessibility ratings (Adjusted $R^2 = 0.642$), and enabled BVI participants to find an accessible video by trying 54% fewer videos and spending 40% less time, compared to the original YouTube interface.

- **Chapter 3** reports on a a 4-week field deployment of a browser extension that displays automated video accessibility metrics for BVI people on YouTube. Our results show

that BVI users considered a video's accessibility only after they first determined if a video met their content needs. Additionally, participants primarily considered audio accessibility metrics which indicate how understandable a video is as-is. Accessibility metrics also enabled participants to explore new approaches to find videos, watch new video content and use videos in more diverse ways.

- **Chapter 4** presents a system that helps video authors efficiently detect and address accessibility issues in videos. Using cross-modal machine learning models, our system automatically measures accessibility of visual and audio segments in a video by checking for modality asymmetries. The system then displays these segments and surfaces visual and audio accessibility issues in a unified interface, making it intuitive to locate, review, script descriptions in-place, and preview the described and captioned video immediately. Participants in our user studies were able to author audio description and closed captions significantly more efficiently (over 40% improvement in both precision and recall) and with lower mental demand.

- **Chapter 5** concludes this thesis by reflecting on my exploration of these projects and suggesting future opportunities.

# CHAPTER 2

# What Makes Videos Accessible to Blind and Visually Impaired People?

## 2.1 Introduction

Recently web-based video content has proliferated as a new key source for information in the form of explainer videos, lectures, unboxings and reviews, how-to's, vlogs, trip reports, commentary, news and more. A video hosting service, YouTube.com, is now the second most popular search platform [106], the second most used mobile application [5], and reaches 81% of internet users under 25 [106], yet it does not require or provide alternative text descriptions, inline audio descriptions, or extended audio descriptions for the video content (criteria for WCAG 2.1 A, AA, and AAA respectively [19]), presenting potentially serious barriers for blind and visually impaired Internet users who may lack access to the visual content in videos encountered online.

Despite the prevalence of potentially inaccessible video content, blind and visually impaired (BVI) participants in our formative interviews watched online videos regularly, finding videos related to their interests via search, recommendation, or external links, and consuming videos for the purposes of entertainment, learning new things, and maintaining social connectedness, similar to studies of the general population [105]. But, BVI participants also cited the ***accessibility*** of a video — or the ability to enjoy and understand the video without additional description of the visual content — as a key criterion for selecting videos. Through interviews and a co-watching exercise, BVI participants identified factors that indi-

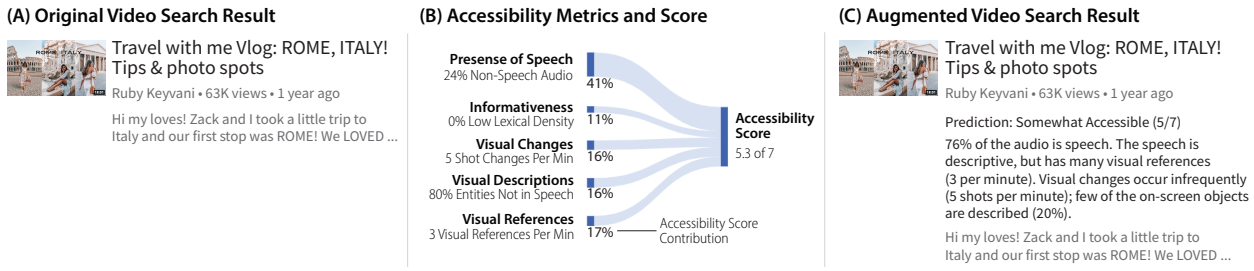



Figure 2.1: Video search results (A) contain no information about whether or not the video is accessible. People must use trial and error to find an accessible video. Our system calculates *video accessibility metrics* (B, left) informed by BVI formative study participants, and predicts the overall non-visual accessibility of the video (B, right). BVI people using our system can preview the accessibility score and explanation (C) to filter or quickly identify accessible videos from search results.

cate whether a video is accessible, including auditory information in the video (*e.g.*, speech vs. silence), visual content in the video (*e.g.*, interview with a single shot of two people talking vs. a movie trailer with rapid shot changes), and moments where visual content was described in audio (*e.g.*, dishwasher repairman explaining each step in detail) or not described in audio (*e.g.*, ingredient amounts displayed using on-screen text and referenced but not described by the presenter). BVI people reported often leaving videos that were unexpectedly inaccessible; and, when exploring new topics or creators, they found accessible videos through trial-and-error, a tedious and time-consuming process that requires clicking on each video, previewing a segment, and guessing whether the rest of the video will be accessible.

To allow BVI people to efficiently surface accessible videos without trial-and-error, we present automated metrics for predicting video accessibility and we exposed these metrics to users during video search (Figure 2.1). To create the metrics, we first defined 7 heuristics to determine a video's accessibility based on the formative study that are related to the video's audio content (*presence of speech*, *informativeness of speech*), visual content (*visual*

*simplicity*, *infrequent scene changes*), and references between the audio and visual content (*describes objects, describes on-screen text*, and *few visual references*); then, we implemented 7 corresponding metrics to assess a video's adherence to each heuristic (Table 2.2).

We collected 180 accessibility ratings for 60 video samples from 14 surveyed BVI participants, then performed a regression analysis suggesting that our metrics are strong indicators of video accessibility as perceived by BVI people (Adjusted $R^2 = 0.642$, $p < 0.001$). We then employed these metrics in the implementation of a proof-of-concept video search interface that displays accessibility metrics and filters videos with respect to predicted accessibility scores (Figure 2.1C). This augmented interface lets users view a a video's accessibility score and metrics-based explanation in the video search results pane alongside typical video information like the title and description, or filter by accessibility score. In a user study, 8 BVI participants performing 3 video search tasks (*e.g.*, assessing capabilities of a new technology, selecting a paper plane tutorial) tried 54% fewer videos and spent 40% less time before making a final selection when using augmented interfaces than they did when using a traditional search interface, and unanimously preferred accessibility metrics-augmented video search interfaces to the traditional approach. Participants reported that they used both the score and the lower level metrics to select a video, and confirmed that the system scores matched their own accessibility assessment after watching their selected video.

In this chapter, we contribute:

- A formative study that finds accessibility to be a key criterion for BVI people when searching for videos, and themes that capture how BVI people evaluate a video's accessibility.
- A set of 7 accessibility heuristics and corresponding automated metrics that correlate with BVI peoples' video accessibility ratings.
- A study with BVI people demonstrating that augmenting a video search interface with our accessibility metrics reduces trial-and-error when selecting videos.

## 2.2 Related Work

We propose metrics to assess video accessibility, and augment a search interface with the metrics to help users find more accessible videos. Our research relates to metrics that quantify web accessibility, prior work on video accessibility, and how people traditionally search, browse, and navigate videos online.

### 2.2.1 Web accessibility metrics and search

A long history of work exists assessing the accessibility of websites by establishing web accessibility guidelines such as the Web Content Accessibility Guidelines (WCAG [19]), and evaluating website adherence to these guidelines through manual or automated methods [92, 58]. Automatically assessing web accessibility could help non-expert developers identify and fix accessibility problems [80], or help web users surface sites that might be accessible [95]; in practice such automated accessibility evaluation results can be challenging for developers to interpret [80], lack sufficient coverage of accessibility issues [93], and inadequately represent user's perceptions of accessibility [95]. Thus, expertise remains important for evaluating and improving on accessibility of websites.

Today, even when websites or applications themselves are accessible (*e.g.*, navigable with a screenreader), a vast majority of the user-generated content hosted on those sites may not be accessible (*e.g.*, videos due to a lack of high-quality captions, or missing alt text). When finding information on large video hosting sites like YouTube, a question becomes what video to choose rather than which website to use. Yet, prior automated metrics do not capture video accessibility [93] and WCAG guidelines provide only high-level guidance [19]. Thus, we study what makes videos accessible to blind and visually impaired users, implement metrics to assess video accessibility based on our findings, and augment a search interface with accessibility information to help users surface accessible videos.

### 2.2.2 Video accessibility

The accessibility of a video is traditionally determined by whether or not it has accompanying audio descriptions – or narrations of "important visual details that can not be understood from the main soundtrack alone" [75] – associated with the video (much as images accessibility is based on the presence of alternative text [19]). For instance, the Section 508 Rehabilitation Act, the Web Content Accessibility Guidelines (WCAG 2.1) [19], and the 21st Century Communications and Video Accessibility Act require, at a minimum, synchronized audio descriptions (i.e. narrations that play alongside the source content, and avoid overlapping important audio [2]) for videos unless the visual content is fully redundant with the audio or text [67]. But, unlike traditional media (TV and movies) where professionally produced audio descriptions are becoming increasingly common on streaming services [74], audio description for user-generated videos is exceedingly rare due to many factors including the expertise typically required to create descriptions, a lack of platform support, and insufficient awareness education. Further, a survey of 91,421 educational videos published by 113 universities found that only 13% of videos provided captions, and none of them provided audio descriptions [8].

Prior work proposed methods to make audio description easier to create through task-specific authoring tools [18, 1], feedback on the content at production-time [70], feedback on audio descriptions [82, 65], and hosting descriptions [41]. Such manual approaches to creating audio descriptions (AD) are time-intensive, and do not yet scale to the endless amount of content people access on video hosting sites today. Work in Computer Vision instead automatically generates captions for visual content in videos [7, 32], but such methods still create inaccurate and unspecific captions in comparison to humans. With few exceptions [79], automatically generated captions are also not specific to AD such that they may not capture content that is contextually important to BVI people (as in the case for alt text [104]). Other research instead used computational assistance to help authors write audio descriptions more efficiently by: using computer vision to detect key visual content [31, 30], deep

learning to provide a computer-drafted summary [31, 107], synthesized voice to convert text to speech [30, 47, 48, 89], and automatic editing to fit human-authored descriptions into the space provided [68]. While these human-in-the-loop techniques help authors improve specific videos, people often search and browse to select videos to watch among a large number of existing videos that do not yet have AD. As in our work, other research also proposes complementary solutions to AD for video accessibility including making the video contrast higher through manipulation [81], and broadly making media players more accessible [63].

But, even when professionals create audio descriptions, the potential accessibility benefit of descriptions for each video depends on factors such as the amount of visual content in the video that can already be "understood from the main soundtrack alone" [2, 6] – or how accessible the visual content in the video already is. Existing guidelines that focus on remediation methods (*e.g.*, WCAG 1.2.5 AA for synchronous audio description, and 1.2.7 AAA for extended description) do not yet distinguish videos that are highly inaccessible (*e.g.*, a silent demonstration of how to fold a paper airplane) from videos that mostly-accessible and already useful (*e.g.*, a video that narrates all demonstrated steps required to fold a paper airplane, but does not narrate a short airplane-flying sequence at the end). Thus, we study what properties make online videos accessible as-is to blind and visually impaired users, and create automated metrics based on these properties. Then, we augment a video search interface with automated metrics to let users efficiently surface more accessible videos.

### 2.2.3 Searching and browsing online videos

Prior work studies how the general population searches, browses, and watches videos online [72, 45, 21, 22, 108]. Despite a common misconception that user-generated videos are watched only for viral content, a recent survey with 12,000 YouTube users finds that 3 of the top 5 reasons users ranked as their primary purpose for watching videos related to seeking information (*e.g.*, #2 teaches me something new, #3 allows me to dig deeper into my interests, and #5 relates to my passions), with entertainment as the second most common

purpose (*e.g.*, #1 helps me to relax, and #4 makes me laugh) [34]. Other work also suggests information seeking, entertainment, along with social connectedness, as key reasons for watching videos online [72, 21, 102, 45]. Given that videos often convey information through visuals, much of the information may be inaccessible to blind and visually impaired users, potentially creating barriers to accessing content of interest. We study what makes videos inaccessible to BVI users and how to help users avoid inaccessible content when seeking information online.

Prior work confirms that many people with visual impairments are active on social media sites where they may encounter videos as part of social interaction (*e.g.*, Twitter [33, 23], Facebook [96], YouTube creators [83], and Snapchat [14]). While such work examines how inaccessible visual content such as videos impacts interactions with others (*e.g.*, ignoring inaccessible videos, or seeking additional information from others), we aim to advance the understanding of: (1) what makes videos inaccessible to blind and visually impaired users, and (2) how blind and visually impaired users search and browse potentially inaccessible online videos as content consumers.

## 2.3 Formative Interviews and Co-watching Exercise

To gain a rich, qualitative understanding of what videos blind and visually impaired people find to be accessible, what factors make those videos accessible, and how they find accessible videos to consume, we conducted semi-structured interviews and a video co-watching exercise.

### 2.3.1 Methods

**Participants:** We used mailing lists and social media to recruit 12 blind and visually impaired participants who consumed videos online. Participants were 19-53 years old and described their visual impairment as blind (9 participants), low vision (1 participant), tun-

nel vision (1 participant), or some light perception (1 participant). All participants used screen readers. All participants watched online user-generated videos daily (9 participants) or weekly (3 participants). We compensated participants $25.

**Interviews:** Interviews were semi-structured and between 43-76 minutes long. Participants were asked what types of online videos they typically watched, how they found the videos that they watched (e.g., via search, recommendation, subscription feed), what accessibility barriers they encountered when searching and browsing videos, and how they navigated such accessibility barriers.

**Co-watching exercise:** We also conducted a video co-watching exercise to elicit participant's lower-level accessibility considerations. Participants watched 3 videos in a random order while sharing their screen. We selected the 3 videos randomly from a set of 12 curated 1-2 minute video clips from YouTube's trending page that represented broad coverage of YouTube categories and amounts of speaking. We asked participants to describe moments when they wanted more information about the video content.

**Analysis:** Two authors of this paper analyzed the interview and co-watching exercise transcripts[1]. The two authors first independently open-coded a subset of the interview transcripts and met frequently to discuss codes until agreement was reached. Then, one author applied codes to the remaining interview transcripts. The interview codes consisted of 7 high-level themes (e.g., video types, accessible video properties, strategies for getting more information) and 70 lower-level codes. The two authors then analyzed participants' information requests from the co-watching transcripts by applying codes for reasons to ask for more information (e.g., missing visual references) and the type of information requested

[1]Transcribed using rev.com

| | Accessible | #P | Inaccessible | #P | #M |
|---|---|---|---|---|---|
| Audio | Presence of speech | 9 | Lacks speech | 10 | 21 |
| | Descriptiveness | 6 | | | |
| Video | Visual simplicity | 4 | Visual complexity | 6 | |
| Audio/video | Described visuals | 4 | Visual references | 4 | 7 |
| | | | Undescribed text | 7 | 2 |
| | | | Undescribed sound | | 25 |

Table 2.1: Properties of accessible and inaccessible videos as reported by formative study participants. We include the number of participants who mentioned each property during interviews (#P), and the number of times an issue was mentioned during co-watching exercises (#M).

(e.g., setting) to all co-watching transcripts independently, and then resolving disagreements.

### 2.3.2 Results

#### 2.3.2.1 How do participants select videos to watch online?

All participants selected videos to watch based on how well the video matched their interests or search need, and the level of accessibility of the video. All participants reported they watched online user-generated videos based on their interests and hobbies, in domains including comedy, sports, gaming, talk shows, reviews, music, and vlogs (similar to the general population [45, 34]). Participants also used online videos for education (e.g., for work, supplementary learning for courses, hobbies, current events), and procedural tasks including: dancing (P1), making a paper plane (P3), repairing a dishwasher (P6), solving a Rubik's cube (P7), programming (P9), music production (P10), knitting (P12), and cooking (P1, P6). They directly selected videos to watch on YouTube based on their interests (via homepage feed, subscription, and searching/browsing for a particular category or topic), and 6 participants also found videos via referral (shared by friends, redirected from social media,

or required by school/company). All 12 participants cited the accessibility of videos, or how much of the video is understandable from the audio alone was a key factor in selecting which videos to consume.

#### 2.3.2.2 What makes a video accessible or inaccessible?

Participants reported that the large majority of videos online did not have audio descriptions, thus they evaluated videos to be accessible or inaccessible based on properties of the original video (Table 2.1, #P):

**Presence of speech.** Participants stated that videos that contained speech for the majority of the time were more accessible than videos where the whole video, or large parts of the video, contained only music or silence. Except for when seeking out music (e.g., P4 often listened to concert recordings), participants found video clips without speech to be uninformative:

> *"The thing that really bugs me too, it's those videos that are only music and no dialogue. Just music, it's so annoying."* – P10

**Descriptiveness.** Participants found videos with descriptive speech to be more accessible, and specifically sought out creators that were more descriptive than others in their speech. For instance, a particularly descriptive streamer could provide more information about a game's visual content than others:

> *"He gives a lot more about the thing rather than pointing at the picture and saying, look at that. Instead he says, here's information about this thing. And to other people maybe that's unnecessary and probably even annoying because it's like I'm seeing it, so why are you telling me? To me it's perfect."* – P8

**Visual complexity.** Participants found that videos that delivered most information verbally and little as visual content (e.g., talking-head style interviews or commentaries)

were more accessible than visually complex videos. Videos were less accessible when they contained a large amount of visual content relative to the amount of time – such as sports highlight reels (P5) and movie trailers (P4) – because the visual content was less likely to be described within the video given time constraints.

**Visual references.** Participants described that speech in the video that referenced visual content (e.g. speaker saying "look at this", or "check it out") would often create inaccessible moments in otherwise accessible videos:

> *"Standup comedians will do a visual joke and will be like, 'Oh yeah, we're just doing this now."* – P9

**Undescribed text.** Participants mentioned that on-screen text was inaccessible when it was not verbally described in the video. Inaccessible on-screen text often included: subtitles (e.g. for a video segment in another language), detailed instructional information (e.g. displaying the amount of salt but not saying it in a recipe video), titles, and other details (e.g. release dates in game trailers).

**Described visuals.** Participants found that descriptions of visual content embedded into verbal explanations such that any necessary visual details are fully explained made videos accessible. Participants cited that embedded descriptions were particularly important for how-to videos including repair (P6) and crafts (P3, P12).

## 2.4 Video accessibility metrics

BVI participants in our formative study cited accessibility as a key criteria when searching, browsing, and selecting videos to watch, and they characterized videos as more accessible or less accessible based on (1) the presence of information in the audio track (presence of speech and descriptiveness), (2) the presence of information in the visual content (visual complexity), and (3) indicators of how well the audio track implies visual information (described visual

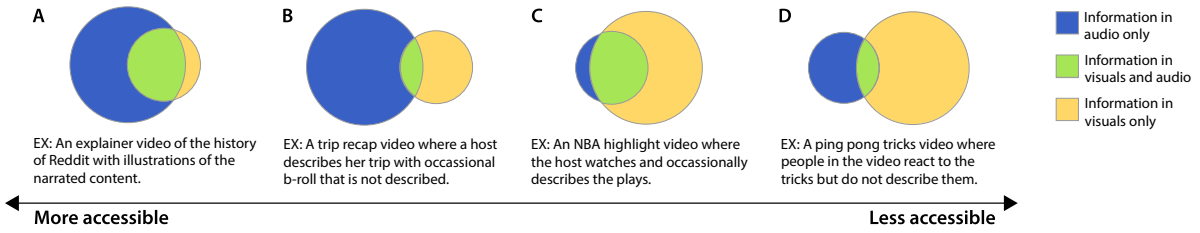



Figure 2.2: The accessibility of a video depends on what information is in the audio track (blue, left), what information is in the video track (yellow, right), and what information is redundant between the audio and visual channels (green, middle). In (A) more information is contained in the audio than the visuals, and most of the information in the visual content is also in the audio. In (B) more information is contained in the audio than the visuals, but little of the visual information is explained. In (C) more information is conveyed visually, but some of this information is also covered in the audio. In (D) more of the information is conveyed visually, with little of this information covered in the audio.

content, undescribed on-screen text, visual reference, and undescribed sounds). Overall, videos that convey more information through audio rather than visuals are more accessible (Figure 2.2B vs. 2.2D). Given an equivalent amount of audio and visual information between two videos, those that convey more of the visual information in the audio are more accessible (Figure 2.2A vs. 2.2B). Based on findings from our formative study, we propose 7 accessibility heuristics (**H**) and 7 corresponding quantitative metrics (**M**) to measure video accessibility (Table 2.2).

### 2.4.1 Audio-related

We propose two heuristics and metrics that are related to the amount of audio information in videos.

| |
|---|
| **H1:** Presence of speech |
| **M1:** Percentage of non-speech duration |

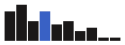
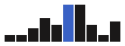
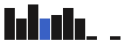
| **Heuristics (H) and Metrics (M)** | **Distribution** | **Median** | **Examples** |
|---|---|---|---|
| *Audio* | | | |
| **H1:** Presence of speech<br>**M1:** % Non-speech |  | 16% | *Accessible*: 0%, explainer video on the history of Reddit [28].<br>*Inaccessible*: 73.6%, demonstrations of Ping Pong trick shots [71]. |
| **H2:** Informative language<br>**M2:** % Low lexical density speech |  | 0% | *Accessible segment*: 0.80, a narrator reads a scripted description [85].<br>"...and **taro root**. **Next, boiled cassava**, **known locally** as **Yuca**."<br>*Inaccessible segment*: 0.27, two people talk about their food as they eat [85].<br>"Oh this is **awesome**. Oh that is **pretty crispy** it's like..." |
| *Visual* | | | |
| **H3:** Infrequent visual changes<br>**M3:** Rate of shot changes |  | 17.43 | *Accessible*: 3.8 shots/min, interview video after a mixed martial arts match [91].<br>*Inaccessible*: 51.6 shots/min, video game trailer [52]. |
| **H4:** Simple visual content<br>**M4:** # Visual entities / min |  | 10.87 | *Accessible*: 2.6 entities/min, late night talk show [51].<br>[audience, crowd, people, performance, reality television]<br>*Inaccessible*: 63.9 entities/min, car advertisement [42].<br>[mountain, car, climbing, dust, dirt road,...] and 45 more. |
| *Audio-visual* | | | |
| **H5:** Description of visual objects<br>**M5:** % Visual entities not in speech |  | 79% | *Accessible*: 60% visual entities not in speech, TikTok food hack reaction video [90].<br>[black, cake, **chocolate**, kitchen]; "...putting **chocolate** on a saltine cracker..."<br>*Inaccessible*: 94% visual entities not in speech, car advertisement [42].<br>[mountain, car, dust, dirt road,...]+45 more; "...to see things from a new perspective..." |
| **H6:** Description of on-screen text<br>**M6:** # Undescribed on-screen text / min |  | 5.16 | *Accessible*: 0 instances/min, customizing fingerboards [59]. No on-screen text.<br>*Inaccessible*: 4.18 instances/min, video game trailer [52]. Release date, producer, platforms, etc.<br>*APEX LEGENDS SEASON 04 COMING FEBRUARY 4!* |
| **H7:** Few visual references<br>**M7:** # Visual references / min |  | 2.46 | *Accessible*: 0 instances/min, story of El Chapo with animated illustrations [86].<br>"...Francisco 'El Chito' Camberos Rivera opened the electronic door..."<br>*Inaccessible*: 13.8 instances/min, TikTok food hack reaction video [90].<br>"...we're gonna pour **this** all on **here** to melt..." |

Table 2.2: 7 accessibility heuristics and corresponding metrics, along with their distributions of the 60 video samples we used in regression analysis. Median is colored blue in histograms. Also shows accessible and inaccessible examples of these metrics.

**Presence of speech** indicates that videos with a larger amount of speech are more likely to be accessible, because BVI viewers gain information from the audio track more often, and fewer portions of the video rely purely on visual content to convey information. We use the metric *percentage of non-speech* to quantify this heuristic (we measure the opposite of this heuristic to keep all metrics' relationship to accessibility consistently negative). For example, a Reddit explainer video [28] which explains how Reddit works in details and keeps on talking (0% non-speech) would be more accessible than ping pong trick shots video [71] in which most of the audio track is just an upbeat background music with occasional verbal

reactions and interjections (73.6% non-speech).

To compute this metric, we retrieve the transcript and audio track of a video, then align the transcript and audio using Gentle forced-aligner [4] to get word-level timing. We consider any gap between words longer than 2 seconds, or about 5 words [3] in length, to be a pause in the speech. We divide the duration of the non-speech pauses over the total duration of the video to find the percentage of non-speech duration.

| |
|---|
| **H2:** Informative language |
| **M2:** Percentage of low lexical density speech |

In addition to the absolute amount of speech in a video, we consider the **descriptiveness** of the speech. Even if speech is present, it is not necessarily informative or descriptive if the speech is vague or implicitly relies on inaccessible visual content (Table 2.2). We use lexical density [44], or the number of lexical words (nouns, verbs, adjective, adverbs) divided by the total number of words, to represent descriptiveness. Comparing transcribed speech segments of equal lengths, a segment with high lexical density (*e.g.*, "boiled cassava, known locally as Yuca") provides more information from the audio alone, than a segment with low lexical density that may be uninformative without the visual content (*e.g.*, "oh that is pretty crispy it's") (Table 2.2).

To calculate *percentage of low lexical density speech*, we calculated the lexical density of transcribed speech within each 10s window (on average, the length of a sentence [43]) of the video (shifted by 0.5s offsets). We calculate the lexical density using NLTK Part of Speech Tagger [55] to identify lexical words for the transcript text within each window. We then divide the total time amount of video segments with lexical density below a threshold of 0.35 (a typical lexical density score for spoken language is 0.45 [44]) over the over the total speech time to find the percentage of low lexical density speech.

### 2.4.2 Visual-related

Participants reported the theme **Visual complexity** in our formative study. We break this into two heuristics and metrics: infrequent visual changes (rate of shot changes), and simple visual content (number of detected entities).

| |
|---|
| **H3:** Infrequent visual changes |
| **M3:** Rate of shot changes |

Participants found that videos were less accessible when they contained a large number of scenes relative to the amount of time, because videos will be more likely to convey information via visual content and the visual content will be less likely to be described given time constraints. For example, videos with few visual changes (e.g. an interview video that only has a scene of two people talking [91], 3.8 shots per min) is usually found more accessible than videos that change shots rapidly (e.g. a video game trailer with complex and fast visual changes [52], 51.6 shots per min). We propose *rate of shot changes* to measure how fast the visual content of the video changes.

We employed a popular shot detection package PySceneDetect[2] to automatically detect the number of shot changes in the video, which compares the HSV colour space difference in content between adjacent frames against a set threshold (default to 30). We divide the number of shots detected by the video duration to get the final score.

| |
|---|
| **H4:** Simple visual content |
| **M4:** Number of detected visual entities per minute |

In addition to scene changes, the number of objects in each frame also affects the level of complexity of the visual content. For example, a talk show video [51] with a simple setup would have less visual content that needs to be described, and a car advertisement [42] may

[2]https://github.com/Breakthrough/PySceneDetect

include a large number of visual objects, which are likely to be inaccessible to BVI users. We propose this metric to capture how many objects are displayed in the visual content.

We used Google's Video Intelligence API[3] to automatically detect entities (objects, locations, activities, animal species, products) in a video. We filtered out any detected entity that has a confidence score lower than 0.9 and count the number of unique entities in the final list. The final number of unique entities is normalized for each video by dividing by the video duration.

### 2.4.3 Audio-visual references

BVI users also reported heuristics related to references between the audio track and the visual content. Participants mentioned that they prefer videos where necessary visual details are described and explained in the speech. They also reported two specific cases where they notice a gap between the visual content and the audio content: the lack of description of texts shown on screen (e.g. subtitles, relevant information, release date), and speech referring to visual content without explaining it (e.g.'Look at this', or 'check it out').

| |
|---|
| **H5:** Description of visual objects |
| **M5:** Percentage of visual entities not in speech |

Based on the theme **Described visuals** in our formative study, BVI users prefer videos where most visual objects are described in the audio. For example, a TikTok food hack reaction video with a very talkative host describing everything she was seeing [90] would be more accessible than a car advertisement in which the speech is just motivating quotes that are completely irrelevant to the visual [42]. We propose this metric to estimate how much of the visual objects are not described or even mentioned in the audio track.

We first detect all visual entities and their timestamps with Google's Video Intelligence

[3]https://cloud.google.com/video-intelligence

API and filter out entities with a low confidence score ($< 0.9$), same as what we did in ***H4***. Then, we find synonyms for these detected entities using NLTK WordNet [61, 62] and check if at least one of their synonyms is mentioned in the transcript. If the entity is not mentioned anywhere in the transcript, we consider it not-in-speech. We compute the final score by dividing the number of visual entities not in speech by the total number of entities detected in this video.

| |
|---|
| **H6:** Description of on-screen text |
| **M6:** Number of detected on-screen text not in speech per minute |

We propose to catch scenarios where on-screen texts are not described in the audio, based on our formative study finding **Undescribed text**. On-screen texts often contain important information including translation of a foreign language, detailed recipe information, etc. For example, in a video game trailer [52], BVI users will completely miss the announcement of its releasing date displayed as on-screen texts without description, which is one of the most important information in this video. Many audio description guidelines explicitly required describers to describe on-screen texts that are not in the original audio track.

To automatically detect non-described on-screen text, we first applied Google's Video Intelligence API's Optical Character Recognition (OCR) function[4] and retrieved a list of on-screen texts with their timestamps. We then determine if each text is covered in the audio track by checking if there is a similar text within the $\pm 10$ seconds period in the transcript. Specifically, we consider two texts segment to be similar if they have a word-wise Levenshtein Distance greater than 0.8. We then normalize the score by dividing it by the duration of video.

| |
|---|
| **H7:** Few visual references |
| **M7:** Number of unresolved reference words per minute |

[4]https://cloud.google.com/video-intelligence/

In our formative study, BVI participants reported **Visual references**, where the speech is referring to visual content without detailed explanations (e.g., "Oh *this* looks very interesting to me", "We're gonna pour *this* all on *here* to melt"). We propose this metric to capture these unexplained visual references.

To automatically compute this, we first establish a set of reference words that we collected from video co-watching exercise in our formative study ("this", "these", "that", "those", "they", "here", "there"). Then, we use AllenNLP's co-reference resolution API[5] to filter out reference words that are already co-referenced, so that all remaining reference words are not mentioned or explained anywhere in the text. We also filtered out a special case for the word "that", because "that" is often used as a conjunction (e.g. He said *that* he was hungry) rather than actually referring to something (e.g. Put *that* in this bowl). We use NLTK POS-tagger[6] to filter out all "that"s with a part-of-speech of conjunction. Finally, we normalized the number of visual reference words by the video duration.

### 2.4.4 Research Questions

Ou formative study identified heuristics for video accessibility that we used to design corresponding automated metrics for assessing these heuristics (Table 2.2). In the following evaluations, we aim to address two research questions:

**R1**: Can our 7 heuristics: *presence of speech, informative language, infrequent visual changes, simple visual content, description of visual objects, description of on screen text,* and *few visual references*, instantiated as 7 corresponding metrics indicate video accessibility as perceived by BVI users? If so, in what proportions?

**R2**: Will augmenting a search interface with our video accessibility metrics improve video search for BVI users?

[5]https://demo.allennlp.org/coreference-resolution

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

## 2.5 Evaluation: How Metrics Indicate Accessibility Ratings

To determine whether and how our metrics correlate to BVI users' perceived accessibility of videos (**R1**), we collected a set of accessibility ratings from BVI users for 60 video clips, then performed a regression analysis.

### 2.5.1 BVI Video Accessibility Ratings

**Materials:** We first manually selected 60 videos from our dataset of videos (Section 3.2) to obtain a broad coverage of YouTube categories and production styles. The sample contains videos from 11 different categories (e.g. sports, how-to & style, comedy). There is no overlap between this dataset and the 12 video we used in our formative study. For each video, we selected a clip with duration between 1 - 3 minutes.

**Participants:** We recruited 14 blind and visually impaired participants to rate their perceived accessibility of our collected videos. Participants were recruited through an email list of BVI participants from prior studies. Participants ranged from 20-53 years old (4 female and 10 male), and described their visual impairments as totally blind (8), light perception (3) and tunnel vision (1). All participants watched online videos regularly (9 daily, 5 weekly). 9 participants have participated in our formative study, and 5 participants were newly recruited.

**Survey design:** To collect BVI users' accessibility ratings of the 60 videos, we sent participants 45-minute online surveys (Google Forms), each of which contained 10 randomly selected videos. To obtain reliable accessibility ratings, we collected 3 participants' ratings for each of the 60 video clips, for a total of 180 video ratings, and 18 surveys (10 participants completed 1 survey, 4 participants completed 2 surveys with different videos). Participants received $20 in cash or gift card per survey. In the survey, we first ask about participants'

| Survey Question | Avg. Ratings (1-7) |
|---|---|
| Q1: Rate the accessibility. |  |
| Q3: Much information is conveyed via audio. |  |
| Q4: Much information is conveyed via visuals. |  |
| Q5: Audio described most of the visuals. |  |
| Q6: Audio was confusing without visuals. |  |

demographic information and their prior experience with online videos. We then ask them to watch the 10 video clips randomly assigned. After each video, we ask participants to rate the accessibility of the video (from 1-very inaccessible, to 7-very accessible), and to rate four additional Likert scale questions designed to assess what factors — informed by our formative study (audio, visual, or audio-visual references) — contributed to their accessibility ratings (Table 2.5.1, full questions in Appendix). To learn if users' perception of accessibility align with our assumptions, and to find out if there are cases of users not aware of what they are missing, we asked two open ended questions: provide reasons for your accessibility rating of this video, provide a 3-5 sentence summary of this video.

**Per-video accessibility ratings:** To obtain the per-video accessibility ratings, we averaged participant ratings (Table 2.5.1). Overall, participants achieved moderate to substantial agreement [49] for accessibility ratings with Cohen's Kappa $\kappa = 0.57$, and per-component ratings with $\kappa_{Q3} = 0.57, \kappa_{Q4} = 0.61, \kappa_{Q5} = 0.54, \kappa_{Q6} = 0.60$. Participants ratings of video accessibility (Q1), significantly correlated with their ratings of audio, visual and audio/visual components of accessibility ($p < 0.001$) with Pearson correlation coefficients of 0.942, 0.949, -0.887, -0.944 for questions about audio (Q3), audio description of visuals (Q5), visuals (Q4), and audio confusing without visuals (Q6), respectively. Thus, we use only the overall accessibility rating for the remainder of this paper. As our goal is to obtain ground truth

accessibility ratings for videos, we removed 5 videos from our dataset with significant disagreement on accessibility ratings between participants (range (max - min) $\geq 5$, where the median range for per-video accessibility ratings was 1) to obtain a final set of accessibility ratings for 55 videos.

**Survey Results:** Overall, participants rated videos in our sample as slightly more accessible than inaccessible with a mean video accessibility rating of 4.64 ($\sigma = 1.79$). While participants achieved an agreement in their ratings for most videos, all 5 high disagreement videos (at least one rating of both "very accessible" and "inaccessible", or both "very inaccessible" and "accessible") shared similar characteristics: (1) a large proportion of the video contained descriptive speech, but (2) most of the speech was communication between the hosts that did not talk about the visual topic of the video (e.g., a video about a dog and a cat meeting where the hosts chat and joke consistently but do not discuss the pet's actions, or a meme reaction video where memes prompt chatty tangents but the hosts do not describe the memes).

### 2.5.2 Regression Model

We fit a linear regression model using video accessibility ratings as the dependent variable, and automatically computed metrics as independent variables. More complex models may have better fits, but a linear regression model allows us to take advantage of its explainability, which informs us about whether and how much our accessibility metrics indicate video accessibility ratings. We can also display this interpretable algorithmic decision process to users. We compute our accessibility metrics (Section 4) on the 55 videos in our final dataset (Table 2.2), then normalize the metrics to fall on a 0-1 scale to build the model. The linear regression model is fitted using the Ordinary Least Squares (OLS) estimator.

**Results:** For the first research question **R1**, we are confident that there is a statistically

| | Accessibility Rating | | |
|---|---|---|---|
| **Metric** | *Initial Model* | *Reduced Model* | *Weight* |
| Const. | $8.22^{***}$ | $8.44^{***}$ | |
| M1: % Non-speech | $-4.85^{***}$ | $-5.03^{***}$ | 40.5% |
| M2: % Low lexical density speech | $-1.58^{*}$ | $-1.35^{*}$ | 10.9% |
| M3: Rate of shot changes | $-1.66^{*}$ | $-1.95^{**}$ | 15.7% |
| M4: # Visual entities / min | $-1.53$ | | |
| M5: % Visual entities not in speech | $-1.78^{*}$ | $-2.00^{*}$ | 16.1% |
| M6: # Detected on-screen text / min | 1.19 | | |
| M7: # Visual references / min | $-2.09^{**}$ | $-2.08^{**}$ | 16.8% |
| $R^2$ | 0.689 | 0.669 | |
| Adjusted $R^2$ | 0.642 | 0.635 | |
| *Note:* | | $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$ | |

significant relationship between a video's accessibility and our proposed 7 metrics. The model fits the data well with an Adjusted $R^2 = 0.642$, $p < 0.001$ ($F = 14.86$). This means that our accessibility metrics contribute to approximately 64% of the variability in the accessibility ratings.

M1 (%speech), M2 (%low lexical density), M3 (rate of shot changes), M5 (%visual entities not in speech) and M7 (#visual references) are tested to have non-zero correlations to accessibility ratings, while there is insufficient evidence for M4 (#visual entities) and M6 (#on-screen texts not in speech) (Table 2.5.2). Thus, the model suggests that M1, M2, M3, M5 and M7 are statistically significant predictors of video accessibility. We hypothesized M4 and M6 are not significant because we were collecting BVI users' perceived accessibility of videos, and M4 and M6 are the only two metrics that are purely based on visual and not accessible to participants. M4 and M6 are not indicative of the perceived accessibility, but could still measure the true accessibility. We further validate this in our user study with BVI participants (section 6), and discuss this difference in the discussion section (section 7).

To measure how much each of our metrics contribute to video accessibility ratings, we first removed the two insignificant metrics in our initial model and re-fitted a reduced model (Table 2.5.2). This model still has high fitness with an Adjusted $R^2 = 0.635$, $p < 0.001$ ($F = 19.81$), and all remaining metrics are statistically significant. Since all the metrics are normalized into 0-1 scale beforehand, the magnitude of their coefficients tells us how much the accessibility rating of a video will change when metrics change, thus showing how much each metric relatively contribute to video accessibility. M1 (%speech) is the most important factor and contributes to over 40.5% of participants' perceived accessibility. M3 (rate of shot changes), M5 (%visual entities not in speech) and M7 (#visual references) are on the similar level contributing around 16% each, and M2 (%low lexical density) about 11%.

## 2.6 Evaluation: A Video Search Interface Augmented with Accessibility Metrics

Our study wanted to find out whether accessibility scores and metrics can improve BVI users' experience browsing and searching videos online (**R2**). We created a proof-of-concept video search interface augmented with video accessibility prediction and metrics, and evaluated this interface with 8 BVI participants who watch YouTube videos regularly.

**Materials:** We designed three video searching tasks: (A) Find a tutorial video of making a paper plane, (B) Find a trip to Italy video to know more about what places to visit, and (C) Find a video about Boston Dynamics robot dog to know more information about it. We selected these tasks because they contain videos with diverse production styles and predicted accessibility (e.g., for task B, videos include a Italy travel Vlog with mostly background music, but also a top 10 places to visit video with extensive narrations). For each search task, we selected the top 10 search results from YouTube by entering relevant keywords using an empty account.

We also built three different interfaces for each search task: (1) the "original interface" is designed to work like the YouTube interface, including typical information like the title, author, length, view counts and description for each result. (2) The "metrics interface" has a similar search results page, but also includes a predicted accessibility score (1-very inaccessible to 7-very accessible) along with an explanation for the score. (3) The "metrics and filter interface" includes the same search result page with the added accessibility information and also a filter that can select videos with predicted accessibility score $\geq 5/7$ (somewhat accessible). Accessibility metrics for all 30 videos are automatically computed using methods described in section 4.

**Prediction Model:** We used the reduced linear regression model described in section 5 to generate accessibility predictions for videos. The model achieved a Mean Absolute Error (5-fold cross-validated) of 1.17 on a 1-7 scale, which means that on average the model predicts the accessibility score of a video within $\pm$ 1.17 of its real value.

**Procedure:** We recruited 8 participants with age ranged from 25-53 years old (4 female, 4 male) who all watched YouTube videos regularly. Participants were recruited from an e-mail list of blind and visually impaired participants who have previous participated in our accessibility research. 7 out of 8 participants have participated in our formative study or the survey. We conducted a 50-minute long remote interview with each participant and each was paid $25 in cash or gift cards. We started by demonstrating accessibility scores and explanations through a 5-minute tutorial. Then, each participant conducted tests for all three interfaces (order counter-balanced). For each interface, one of the three search tasks (paper plane, trip to Italy, robot dog) is randomly selected without repetition. For each test:

1. We asked participants to select among the search results for a video that satisfies the task goal and their accessibility preferences. Participants can view all video information (title, author, # views, length, description), accessibility information (prediction score

and explanations) for metrics and metrics+filter interfaces, and click on the video link to go to YouTube and view the video.

2. After participants have finalized their selection, we first asked them to describe their thought process of selecting videos using the interface, and explain reasons they selected this video. Then, we asked them to rate the accessibility (1-very inaccessible to 7-very accessible) based on the information they currently have (e.g. title, accessibility prediction, the first thirty seconds of the video previewed). We also asked them to rate their confidence of their ratings.

3. We then asked them to watch the entire video and rate the accessibility of that video after watching it.

After completing all three tasks, we asked participants to compare their experience of video searching with all three interfaces. We audio recorded all sessions and screen recorded participants' interactions with the three interfaces. We also timed how long each task took to complete and how many videos participants clicked into and previewed before selecting the final one.

### 2.6.1 Findings

Participants unanimously preferred the augmented interface with accessibility metrics+filter, followed by the augmented interface without filter, and then the original interface. Participants especially liked how the two augmented interfaces showed both the scores and the explanations. All participants described it as a "neat" way to surface accessible videos among search results. They all liked that the two augmented interfaces provide them accessibility information ahead of time, so they can avoid inaccessible videos that they would otherwise be wasting time on:

| | | Task time | | | # Videos clicked | | |
|---|---|---|---|---|---|---|---|
| **P#** | **Tasks** | *Original* | *Metrics* | *Filter* | *Original* | *Metrics* | *Filter* |
| P1 | B2, C3, A1 | 10:05 | 5:14 | 1:25 | 9 | 4 | 1 |
| P2 | A2, C3, B1 | 2:06 | 1:33 | 1:05 | 2 | 1 | 1 |
| P3 | B2, A1, C3 | 5:34 | 1:34 | 3:04 | 5 | 1 | 2 |
| P4 | A1, C2, B3 | 5:58 | 5:51 | 3:20 | 5 | 3 | 3 |
| P5 | C1, B3, A2 | 5:17 | 4:08 | 4:15 | 3 | 3 | 2 |
| P6 | C1, A2, B3 | 10:03 | 8:39 | 4:03 | 2 | 2 | 1 |
| P7 | A3, B1, C2 | 6:26 | 4:18 | 5:07 | 2 | 1 | 1 |
| P8 | A3, C2, B1 | 5:16 | 3:55 | 8:33 | 3 | 2 | 3 |
| **Avg.** | | 6:20 | 4:24 | 3:51 | 3.9 | 2.1 | 1.8 |

*"It shows what's gonna be in the video and how accessible the model thinks it is. So I can choose based on that. I know kind of what I'm getting into before I click. For the YouTube interface you kinda just have to ... hope."* – P1

In video searching tasks, participants generally spent less time and clicked into fewer videos to find an accessible and suitable one to watch using the two augmented interfaces, compared to the original interface (Table 2.6). There were few cases of exception, e.g., P8 found an accessible video quickly using the original interface because the first video he randomly selected happened to be accessible.

All participants expressed enthusiasm about using this augmented interface in the future, and they hoped that this could work on different websites (e.g. Facebook, Twitter) and different platforms (e.g. PC, smartphone).

**Video searching and browsing behavior**. We discuss how participants use different interfaces to find videos that satisfy the task goal and their accessibility preferences.

(1) With the original interface without any accessibility information, all 8 participants

mainly relied on contextual information, such as video title, video description, author, video length, and number of views, to speculate accessibility. All participants also utilized the trial-and-error approach, as we discovered in our formative study. Participants generally took a long time completing the task and often could not accurately estimate accessibility based on such contextual information, causing a lot of 'try-and-exits' (quitting a video after briefly watching it). P4 actually gave up looking for an accessible video after viewing 5 videos and finding all of them to be inaccessible using the original interface (for the "learn to make a paper plane" task), because she felt all the videos in the search results would probably just be the same and did not want to waste time (there were actually 4 videos with predicted accessibility score greater than 5—somewhat accessible—in the search results).

(2) With the augmented interface, all 8 participants prioritized the accessibility of videos. They would first identify accessible videos based on predicted accessibility scores and explanations, and then among these select the ones that are more relevant to the task. All participants liked the structure of presenting a general prediction score followed by detailed metrics information:

> *"I love the details like the metrics and the score and they just work so well together. I love the way it is organized. It's a good combination of enough information, but not too much."* – P3

All participants understood the accessibility score and metrics easily. P2, 7 and 8 mainly relied on the accessibility score, because it conveys key information quickly and succinctly. They found it to be especially helpful when going through a large number of videos. P1, P2, P3, P4 and P6 also found explanations with accessibility metrics to be particularly useful, because it provides transparency and extra explanations. 6 out of 8 participants found the percentage of speech to be the most important metric they would consider, which aligns with our regression analysis. Two (P3, P6) cared about visual changes, taking that as an indication of how hard-to-follow the video would be.

(3) With the filter interface, 6 out of 8 participants turned on the filter to surface the accessible videos, and did not care about other inaccessible results:

> *"I liked that it helps me filter out stuff that otherwise I would be wasting my time on."* – P7

P5 and P6 did not use the filter because they wanted to explore what was available in the search result, and also the amount of videos was limited. However, both participants stated that they would prefer to have the option to filter with a larger number of search results.

**Interpreting automated predictions and trust**. Trust could be an important issue for algorithmic systems [29, 27]. All 8 participants in our study found the prediction scores to be accurate, based on their experience with the system. In our interviews, we asked participants to rate how accessible they thought the selected video was. Our model's predictions achieves a Mean Absolute Error of 0.53 comparing to participants' ratings. P4 mentioned how the scores accurately match with her perception of accessibility:

> *"It's kind of surprising because, there was one video I watch I think she is doing some type of vlog. And it wasn't too accessible, it wasn't too inaccessible, it was exactly like the prediction said it was 'somewhat accessible'. Because she definitely had a lot of speech in there but I definitely noticed a lot of visuals also, that you did really need to see. The score is really really on point."* – P4

4 out of 8 participants reported that they were unsure about the scores initially, and needed to play with it for a while to evaluate how accurate it is. P1 and P3 clicked into several predicted inaccessible videos during the test to check if the scores were accurate. In the augmented interface test, P5 selected a video even though it had a predicted accessibility score of 3, because he felt the title and description of the video indicated that it should an accessible one. He also previewed a small segment of the video and found it to be descriptive.

However, after he watched the entire video he discovered large segments without any speech in latter parts, and agreed with the model prediction. P1, P2 and P3 also described that they felt more confident with model's predictions having explanations available:

> *"The metrics are really good. If it just gives a score of 7 then I might be a little uncertain. But it's got so much information that is really helpful about what to expect."* – P3

**Ideas for improvement:** Participants were generally satisfied with the augmented interface and only suggested small feature changes. P3 and P5 wanted to have more granularity for the filter. Our prototype interface only had one option to filter videos that have a score greater than 5, they would like to have different score thresholds available. P3 and P5 also suggested to make the explanations customizable, since different users may care about different metrics. P2 and P7 suggested to make the interface more structural rather than laying them out one by one. P3 and P7 also would like to know more about the implementation details of how the metrics were computed.

At the end of the user study, we presented participants the two insignificant metrics we excluded from the implementation of the interface due to insignificance in the regression analysis — M4: number of visual objects and M6: number of on-screen texts — and asked them if they would like to know that information. All 8 participants stated that they would be interested in knowing about on-screen texts, but did not care too much about number of visual objects. P3 and P4 mentioned videos they watched that had keynotes or textual/visual jokes and on-screen texts could be very important.

## 2.7 Discussion and Future Work

Our research confirms that considering the inherent accessibility of videos through *video accessibility metrics* is a useful tool for allowing blind and visually impaired people quickly find videos of interest. Our formative study with BVI YouTube users, 8/12 of whom were already using YouTube daily, often selected between multiple comparable videos (*e.g.*, search results for DIY Christmas Ornaments) to watch based on their accessibility in terms of the audio, visuals, and audio-visual factors. Our regression analysis showed that: (1) BVI people agreed on perceived video accessibility ratings for most videos, and that (2) our video accessibility metrics derived from the formative study correlated with BVI's accessibility ratings. Our user study demonstrated that video accessibility metrics and predictions could be immediately applied in the context of video search (for previewing and filtering by accessibility) to improve the experience of BVI people searching for videos.

**Prioritizing Content for Description:** Our work helping people find videos that already have high-quality and built-in descriptions during video search can advance existing work on helping authors add high-quality audio descriptions to individual inaccessible videos [31, 30, 47, 107, 68, 65]. Whereas prior work already surfaced silent video segments for audio description [68, 107, 65], our metrics can be immediately applied to surface non-silent, but still inaccessible video clips for further description (*e.g.*, by calculating the metrics on each 15s segment of a longer video). For instance, our metrics could detect inaccessible moments including undescribed on-screen text such as recipe amounts or corrections, or confusing visual references. Alternatively, our metrics can help video authors identify areas where they could add more descriptive language (as in [70] for slides) to their own videos before publishing.

**Perceived Accessibility vs. True Accessibility:** Accessibility ratings from BVI peo-

ple reflect their perception, but do not capture cases where BVI people are not aware of the inaccessible information they are missing (e.g. on-screen text and visual objects not indicated in the audio). We focused on perceived accessibility because BVI people will be the end users of the tool and we aimed for the final ratings to match their preferences. In the future, we will study the differences between perceived accessibility and true accessibility by (1) performing a summary analysis by comparing BVI users' summaries with summaries generated by sighted people; (2) providing BVI participants original then audio-described versions of the video to watch and rate consecutively.

**Video Samples:** We selected a 60 videos from the YouTube trending page for our regression analysis. While our sample size is small, we obtained expert (from BVI YouTube users) rather than non-expert (*e.g.*, from general population on AMT) accessibility ratings and our sample size falls around the recommended 10 data points per independent variable [35]). In addition, we sampled videos from the YouTube trending page as they represented highly popular videos on YouTube that people may be likely to encounter due to chance or YouTube recommendation. But, our sampling approach revealed videos that were more accessible than inaccessible and this might not be true for videos people usually encounter. In the future, we will collect a larger dataset of ratings with more diverse video content (*e.g.*, more production styles, budgets, and topics) to improve our analysis and predictions.

**Impact of Longer Term Use:** Our user study investigated use by first-time users on three defined search tasks. Despite the learning curve to interpret our accessibility metrics and scores, users experienced efficiency gains and unanimously preferred using our tool. In the future, we will conduct a long-term deployment and analysis to find out when the system is in-the-wild (*e.g.*, task specific search or browsing for entertainment), and if the system impacts browsing behavior (*e.g.*, users more likely to explore new creators or domains, or recommendation algorithm gets better at predicting relevant accessible videos due to less

trial-and-error click-throughs).

**Platform Support and Scalability:** Providing accessible video searching, browsing, and consuming experiences is the responsibility of the platform rather than the user. Video hosting platforms such as YouTube, Vimeo, and TikTok should enable authors to upload audio descriptions, and implement the ability for people to filter and browse videos by their accessibility (*e.g.*, presence of audio descriptions, our metrics of built-in accessibility). Given that YouTube already lets people filter out videos without Closed Captions, a straightforward addition would be to let people filter by the amount of speech in the video (*e.g.*, an option to filter out all videos no speech — videos with only background music were a common complaint). In the meantime, we are building a Chrome Extension for our tool by computing on-the-fly video accessibility metrics for search results. But, our metrics that require video processing (*e.g.*, # on-screen text) are computationally intensive. In the future, we will explore ways to make our metric computation more efficient while retaining accuracy by: sub-sampling videos, storing results, and improving predictions when only a subset of metrics are available.

## 2.8 Conclusion

Surfacing accessible videos on online user-generated video platforms like YouTube is a time-consuming burden for blind and visually impaired users. From heuristics our BVI participants used to describe accessible and inaccessible videos, we instantiated 7 accessibility metrics that can be computed automatically from videos. Through a regression analysis, our metrics correlated with BVI users' perceived accessibility ratings. Participants using our augmented video search interface in a user study unanimously preferred our filtering and browsing support to the traditional interface. The combination of video accessibility heuristics, accessibility metrics and augmented video interface opens up possibilities for sur-

facing accessible videos in a systematic and scalable way, and making video platforms more accessible to all.

# CHAPTER 3

# Understanding How Blind and Visually Impaired People Leverage Accessibility Metrics In Practice

## 3.1 Introduction

In Chapter 2, we identify BVI peoples' challenges of searching and browsing for information when some videos are more accessible than others, and propose metrics to measure and annotate accessibility of online videos. Prior research has also proposed accessibility metrics for websites [95, 92, 93], mobile applications [9] and presentations [70]. Similar to our findings, previous studies have shown that with accessibility metrics, people were able to identify and browse within a subset of the most accessible content and find accessible content more efficiently [95].

However, none of these studies have deployed the proposed metrics in a real world scenario with in-the-wild purposes and content needs. Prior work [16] shows that BVI people also consume content that are not accessible to them for other reasons (*e.g.*browsing a webpage to complete a training, watching movie trailers to stay up-to-date). Accessibility is not the one and only goal people consider when browsing and selecting a content to watch. During the process of content foraging, people have to satisfy different informational needs, and sometimes make trade-offs. There is still a lack of understanding of how people leverage accessibility metrics within the broader context of their everyday content searching and browsing activities.

We conducted a 4-week long deployment study to investigate how BVI people leverage

accessibility metrics in practice. To deploy accessibility metrics, we implemented a browser extension that efficiently computes and dynamically displays video accessibility metrics as people search and browse videos on YouTube. During the study, we collected both qualitative and quantitative data through automated data logging, experience sampling surveys, and semi-structured interviews with participants. In total, we collected 4971 instances of video searching and browsing behavior and 275 of responses for the experience sampling surveys from 10 BVI participants. We asked the following research questions:

RQ1 Do people consider accessibility metrics in practice?

RQ2 What aspects of accessibility are most important to people in practice?

RQ3 How does knowledge of accessibility impact people's browsing and searching strategies?

Our analysis reveals that in contrast to prior in-the-lab study results [54, 95], BVI users considered a video's accessibility only after they first determined if a video met their content needs, and their further consideration of accessibility varied based on the specificity of their content-seeking goal (*e.g.*open-ended browsing vs. searching for a specific topic). Participants also reported that they primarily considered audio accessibility metrics (*e.g.*percentage of speech) to estimate how understandable the video is by the audio alone, then additionally consider visual accessibility metrics (*e.g.*visual changes) only if visual information is important to their task such as in following a tutorial. While participants became more efficient at identifying accessible videos during search, participants also reported feeling more confident in their ability to find accessible videos with the accessible metrics across all searching and browsing tasks.

As a result, participants explored new approaches to find videos, watched new types of videos, and used videos in more diverse ways. Based on our observations, we discuss three trade-offs BVI people consider when selecting to watch a video with accessibility metrics: (1) prioritize content vs. accessibility, (2) prioritize perceivable accessibility (*i.e.*the video

is understandable or enjoyable from audio alone) vs. full accessibility (*i.e.*all important visuals are described), and (3) exploit known searching and browsing routines vs. explore new approaches.

In this chapter we contribute:

- A browser extension that efficiently computes and displays video accessibility metrics on YouTube.
- A deployment study revealing how BVI people leverage accessibility metrics within the context of everyday content browsing and searching activities.
- Discussion and future opportunities to improve the design of accessibility metrics on social media platforms.

## 3.2 Related Work

### 3.2.1 Studying Information Seeking In-The-Wild

Long-standing research in information systems and retrieval categorizes two main types of information seeking behavior [37, 101]: (1) *searching*, or querying, which is exploration with a specific target in mind [101], and (2) *browsing*, or navigating, which is casual, undirected exploration over a set of links [37]. As browsing relies on human's "greater ability to recognize what is wanted over being able to describe it" [38], it can be easier than search if appropriate options are available and they include "meaningful cues about the underlying information" [37]. In practice, users often switch between multiple information seeking behaviours [37]. Because information seeking is a complex and context-dependent task, longitudinal field studies reveal behavior change that is difficult to assess within a lab study with formulated tasks [84]. We aim to learn how access to accessibility metrics may change searching and browsing by users with visual impairments. Prior accessibility research has also studied how blind people search and browse in-the-wild in the context of general web

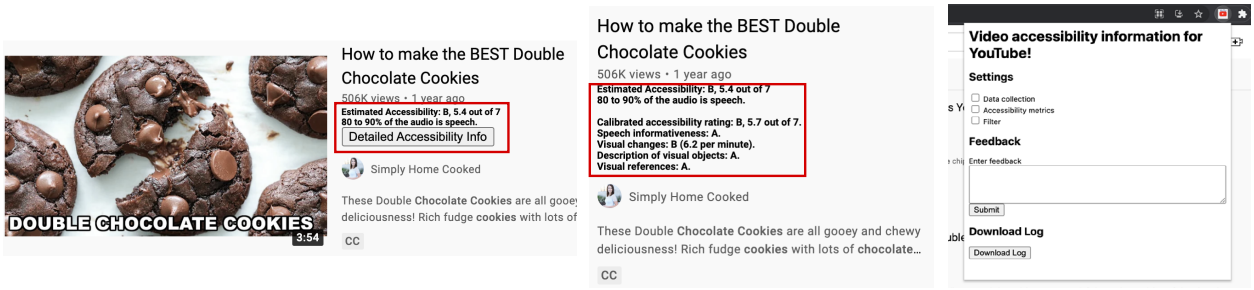
content [16]. Bigham et al. proposed an approach for tracking web exploration and found that while blind people exhibited similar searching and browsing behavior to sighted people overall, blind people were more likely to avoid inaccessible elements [16]. We introduce an approach to track video exploration rather than web exploration, and we study how letting blind users know about the inaccessible content ahead of time could change their searching and browsing approaches.

### 3.2.2 Searching and Browsing on YouTube

While the majority of prior research on information seeking considers text content (*e.g.*, search engines, web pages, library catalogs), research has begun to explore how people search and browse for videos. For the purpose of this paper, we consider YouTube as a platform for exploring searching and browsing. YouTube is widely used for users seeking information, entertainment, and social conectedness [34, 72, 21, 102, 45] — and, like many search engines and social media sites today displayed links are driven by a recommendation algorithm [25, 13]. Prior work studying individual video searching and browsing behavior focuses on how people use YouTube either from a macro scale (*e.g.*, by studying public metrics such as video views [109, 108]) or on an individual scale by asking people about their YouTube use [72, 45, 21, 22, 12, 46]. Beyond public data, Baluja et al. [13] uniquely sample individual-specific YouTube browsing data to study a new recommendation algorithm, but no prior public research has studied the individual's searching and browsing behavior on YouTube. In this work, we create a browser-extension based approach to log YouTube data, and conduct the first public study of an impact to video searching and browsing behavior from an intervention (in this case, Accessibility Metrics).

To provide context for our study on video searching and browsing behavior, we summarize the two primary pages on YouTube that people use to search and browse for videos (for more information see, "Search and Discovery on YouTube" [88]):

**Homepage:** The YouTube Homepage displays a set of video recommendations based




**(A) Accessibility Metrics Displayed on YouTube** **(B) More Accessibility Information** **(C) Extension Popup**

Figure 3.1: (A) Accessibility metrics that our system computed and displayed onto YouTube. It is displayed as a heading level 4 element next to each video title (heading level 3). Screen reader users can quickly navigate to accessibility metrics with headings. (B) More accessibility information that are related to visual (*e.g.*, visual changes, description of visual objects) are computed upon request, and will be automatically added to the interface. (C) Extension's popup page in which people can turn on/off settings, leave feedback, and download their log data for this study. We highlight parts that are generated by our extension on YouTube in red.

on the viewer's watch history and video performance as well as subscriptions, videos watched by similar viewers, and new videos.

**Search:** YouTube Search lets users type in a query and returns a series of search results based on aspects including the match of the title, description, and content to the query, as well as videos that have "driven the most engagement for a query".

## 3.3 Implementation

To investigate how BVI people leverage video accessibility metrics *in-situ*, we designed and developed a system that computes and displays accessibility metrics [54] of YouTube videos and collects information about people's video browsing and watching behavior on YouTube. Our system consists of a browser extension and a backend server. The Chrome extension

serves as the front end interface that shows video accessibility metrics on YouTube. The extension automatically captures YouTube video IDs, and sends them to the backend server[1] that computes accessibility metrics by analyzing the video and captions. The resulting metrics are then sent back to the front end extension and added onto the YouTube interface (Figure 3.1). To protect users' privacy, study participants can turn off data logging of our extension any time. Log data was also saved only locally to participants' computers and were reviewed by participants before sharing. API requests to the backend server were not inspected.

### 3.3.1 Computing Video Accessibility Metrics

To allow people access information about the video's non-visual accessibility as they browse video results, we dynamically compute video accessibility metrics for video results and insert the results into the YouTube page, improving efficiency to make Liu et al.'s metrics practical to compute for real-time interactive use [54].

#### 3.3.1.1 Optimized Metrics

We optimized the efficiency of computing each metrics proposed in the prior work [54]. Our optimized metrics provide comparable accuracy (5-fold MAE = $1.05, R^2 = 0.537$) to prior methods (5-fold MAE = $1.17, R^2 = 0.642$). Computing all metrics and the score takes $0.94s$ ($\sigma = 0.43s$) but downloading the video is slow ($\mu = 23.92s$, $\sigma = 21.25s$). Thus, by default our browser extension first displays metrics that only require captions (M1, M2, and M3), and then computes video-based metrics (M4, M5) only by per-video user request. The extension also initially shows a default accessibility score fitted based on only the caption metrics (5-fold MAE = $1.08, R^2 = 0.496$), and re-calibrates the score if a user requests the additional metrics. All metrics are cached in a central server for future reference.

[1]the backend server is hosted on an Amazon AWS m4.xlarge instance

While metrics in prior work were too computationally intensive for real-world use (around 10 minutes per video total), our design decisions and optimized metrics were necessary to make such metrics work in interactive searching and browsing scenarios. We make our dataset and video accessibility metrics API available for future use.

### 3.3.2 Video Accessibility Metrics Extension

Our browser extension (Figure 3.1) consists of two parts: (1) a YouTube overlay that inserts computed accessibility metrics next to original video metadata, and (2) an extension popup to manage extension settings and provide feedback.

#### 3.3.2.1 Displaying Accessibility Metrics

When Accessibility Metrics are enabled (via the extension pop-up, Figure 3.1A), video accessibility metrics are added as HTML elements adjacent to each video's metadata (Figure 3.1B). The extension by default shows the accessibility score and percentage of speech for each video. If the speech of a video is detected to be not informative, or there is a large number of visual references in the video, the extension displays warnings: *"Speech may not be descriptive"* or *"May contain many visual references"*. Accessibility metrics are shown in letter-grade levels A ($\geq 6$), B ($\geq 4$), C ($\geq 2$), D ($< 2$), where A represents the most accessible and D the least accessible. The levels were discussed and determined through pilot studies with three BVI participants. Users click "Detailed Accessibility Info" to obtain two additional visual features (M4, M5). Our extension works on several types of YouTube

pages: Homepage, Search, Video[2], Explore[3], Subscriptions[4], Channel[5] and Library[6].

#### 3.3.2.2 Extension Popup Interface

The extension popup (Figure 3.1C) allows users to customize settings, report feedback or download log data. Users can open or close the extension popup by clicking on the extension icon in the Chrome tool bar, or use the hot-key: Command (or Control) + Shift + K. Using the pop-up, participants can: toggle accessibility metrics on or off, toggle data logging on or off (*e.g.*to preserve privacy), and toggle filtering by accessibility score on and off. When users choose to enable the filter, videos that with a low accessibility score (below 'B') will be automatically removed from search results.

Users can also use the popup to report their thoughts, problems and suggestions for this extension. When users are on a YouTube video page, two additional things are added to the feedback section: *Fill out an experience sampling survey about watching this video*: By clicking on this link, users will be directed to our experience sampling survey designed for this study. *How accessible do you think the video is*: users can also report how accessible they think the video is, agreeing or disagreeing with the automated prediction.

#### 3.3.2.3 Data Logging

To collect YouTube behavior data, our extension tracked timestamps and URLs of YouTube page events, including *create tab*, *redirect tab* (*i.e.*go to a different URL within the current tab) and *activate tab* (*i.e.*switch to an existing tab). The extension was limited to only log-

[2]youtube.com/watch?

[3]youtube.com/feed/explore

[4]youtube.com/feed/subscriptions

[5]youtube.com/channel/ or youtube.com/c/

[6]youtube.com/feed/library

| UID | Age | Gender | Vision | Screen Reader | YouTube Usage | Platform | Types of Videos Watched | Video Source |
|---|---|---|---|---|---|---|---|---|
| P1 | 27 | M | Blind | NVDA | Daily | Phone | Gaming reviews, comics book | Homepage |
| P3 | 21 | M | Blind | NVDA | Daily | Computer | Music related videos, concerts, interviews, comedy | Search |
| P4 | 27 | F | Light Perception | NVDA | Daily | Phone | Interview of artists, vintage TV shows | Homepage |
| P5 | 36 | M | Blind | NVDA | Daily | Computer | Video games, sports highlights, movie trailer | Search |
| P6 | 54 | M | Light Perception | JAWS | Daily | Computer | Tutorial, recipe, information | Homepage |
| P8 | 21 | F | Blind | VoiceOver | Daily | Phone | Everything | Homepage |
| P9 | 35 | M | Blind | NVDA | Daily | Computer | Video games playthrough, music, news | Homepage |
| P11 | 43 | M | Blind | NVDA | Daily | Phone | News, recipe, tutorial, tech, music related | Search |
| P12 | 25 | M | Blind | NVDA | Daily | Computer | Comedy, sound design, audio games, history, science | Homepage |
| P13 | 48 | M | Low Vision | NVDA | Daily | Phone | Engineering explanation | Homepage |

Table 3.1: Demographic and YouTube experience information of our study participants. 10 participants completed our 4-week long deployment. P2, P7 and P10 did not complete the study and are thus not included in this table.

ging data on the youtube.com domain. We post-processed the recorded URLs and extracted detailed YouTube behavioral data such as *YouTube page type* (*i.e.*homepage, search, subscription), *search queries*, *video id*, *video title*, *video author*, *video category*, *video accessibility metrics*, etc.

#### 3.3.2.4 Pilot Studies

Before deploying our extension, we conducted three one-hour pilot studies with BVI YouTube viewers (also P1, P2 and P3 in our deployment study). Participants provided feedback around adding a loading cue when the accessibility metrics are not yet loaded, and adding a notification sound when finished computing more accessibility metrics. No major usability or accessibility issues were found. Participants also helped us determining thresholds for letter-grade levels of accessibility metrics. Each participant was paid a $20 Amazon gift card.

## 3.4 Deployment Study

To gain a rich understanding on *how do people leverage accessibility metrics within the broader context of their everyday content searching and browsing activities*, we conducted a 4-week long, *in-situ* deployment study with 10 blind and visually impaired participants. We used a mixed-methods study design to gather both quantitative data from YouTube activity logs and qualitative data from participants' experience survey responses and exit interviews. This study was approved by our university's IRB.

### 3.4.1 Participants

We recruited 13 potential participants from mailing lists and our school's Disabilities and Computing Program; 10 participants completed the entire 4-week study (Table 3.1). Participants who completed the full study were 24–54 years old ($\mu = 34.4$, $\sigma = 11.3$) and described their visual impairments as blind (7 participants), light perception (2 participant), and low vision (1 participant). All participants used a screen-reader and watched videos on YouTube daily. 5 participants reported using YouTube mainly on their computer, and the other 5 participants using it on the phone app. We asked all participants to use YouTube on a Chrome browser with our extension installed. From our entry interviews, participants reported using two different approaches to find videos on YouTube: Primary homepage users (P1, P4, P6, P8, P9, P12, P13), and Non-homepage/Search users (P3, P5, P11).

### 3.4.2 Procedure

We onboarded participants with a 45-minute entry session. We first asked participants demographic and background questions (how often do they use YouTube, what videos do they watch on YouTube, how do they find videos on YouTube, *etc.*) in a semi-structured interview. We then walked them through the installation and features of our extension.

The deployment study consists of three stages: (1) In the first week, the accessibility

metrics were turned off, and participants were asked to use YouTube normally as they would such that we could collect baseline data about searching and browsing behavior. (2) In the following two weeks, we enabled the extension such that participants could turn on and see video accessibility metrics. (3) In this last one week, we disabled extension again to let participants to reflect on their experience with and without the extension. Throughout each stage, the extension automatically logged participants' activities on YouTube pages to their local computers. After each stage, we emailed participants to let them review and send us their latest log data (stored locally, and then exported using the browser extension).

Participants were also asked to fill out a daily experience sampling survey, consisting of multiple-choice and short-answer questions regarding one specific video watching experience of the day (*e.g.*How did you find this video? What made you click on this video?). Whenever a participant visited a new YouTube video page, the extension displayed a link to the experience sampling survey in the extension pop-up and on the video page, and participants were asked complete at least one survey for a video they watched each day. Participants were also given the extension three days prior to the start of the study to accommodate to this study routine and to eliminate novelty effects.

We instructed participants to use YouTube for at least 15 minutes per day (all participants already reported using YouTube daily), and fill out at least one experience sampling survey. Every morning, we sent an email to participants with a reminder of their study progress. We did not provide any specific task or prompt to participants during the study as we wanted to observe their day-to-day use. If a participants missed a day, they were allowed to make it up on a subsequent day and continue the study. Participants completed the study in 34 - 53 days. At the end of the deployment, we conducted a 45 minute exit interview with participants regarding usage patterns and behavioral changes we observed from their log data. All completed participants were paid a $100 Amazon gift card as compensation.

### 3.4.3 Data

We recorded and transcribed participants' entry and exit interviews, and collected experience sampling surveys. Following Affinity Diagram approach [39], two researchers organized notes of participants' comments from experience sampling surveys and exit interviews to iteratively develop meaningful and coherent themes. We also recorded participants' YouTube behavioral data from our extension. We post-processed the log data and conducted exploratory data analysis and statistical tests on themes we identified. In total, we collected 4971 instances of video searching and browsing behavior (with 1246 selected videos, and 20,030 videos links displayed in the course of searching and browsing) and 275 of responses for the experience sampling surveys from 10 participants.

### 3.4.4 Limitations

Our extension extracts video information from YouTube and relies on transcript and the video file to compute accessibility metrics. Sometimes accessibility metrics cannot be computed due to copyright issues or video author disallowing captions. In addition, participants may not exhibit their usual behavior as our study required people to use Chrome's desktop browser rather than their mobile device app to search and browse YouTube (5 participants were mainly using YouTube on their phone prior to the study). To account for using Chrome to browse YouTube for the first time, we provided users a 3-day adjustment window before recording the baseline data. The accessibility metrics have limitations in terms of the time required to compute. If metrics were computed on the back-end (by YouTube) rather than using our application, users would have access to the metrics faster and might behave differently (*e.g.*, querying for more metrics more often).

## 3.5 Findings

In this section, we first describe participants' overall usage of our extension. We then report on five themes across the three research questions from qualitative and quantitative analysis of participants' content consumption behavior with accessibility metrics: (1) Do people consider accessibility metrics when selecting to watch a video in practice? (2) What aspects of accessibility are most important to people in practice? (3) How does knowledge of accessibility metrics impact people's browsing and searching behavior?

### 3.5.1 Overall Usage

We first examine how users made use of our extension during Stage 2 (when accessibility metrics were enabled).

#### 3.5.1.1 Overall Usage and Accuracy

All participants kept our extension on for the entire duration of Stage 2. The logged data indicates that the extension provided accessibility metrics for 65.29% of videos participants encountered; the remaining videos did not have captions (due to copyright issues or video author disallowing caption access) available for calculating the metrics In exit interviews, participants reported that they found the overall accessibility score and individual metrics to be accurate. In 80.1% of the responses, participants agreed with the statement *"I found the overall accessibility rating of this video to be accurate."* with a *"6-Agree"* or *"7-Strongly Agree"*, and in 70.7% of the responses, participants agreed with the statement *"I found the other accessibility information (e.g., percentage of speech, visual changes) to be accurate."* with a *"6-Agree"* or *"7-Strongly Agree"*.

All participants found the accessibility metrics to be helpful because the metrics provided them extra information before watching the video, which made them more confident about the more confident about the usefulness of the content. As P12 described, *"I have the scores*

*so I can actually know if it's going to be something meaningful, instead of no idea what's going on."*

#### 3.5.1.2 Usage of "More Accessibility Info"

Participants only clicked on "detailed accessibility info" (Figure 3.1B) to request more accessibility metrics 8 times on average ($\sigma = 7.78$) over the course of the deployment. Most participants felt that the initial overall score and metrics were already helpful enough, and that the loading time was long. P4, P8 and P11 requested detailed information once every four page visits (9, 14 and 27 times in total respectively), which is more frequent than other participants. They used it for content that were especially interesting or important to them.

#### 3.5.1.3 Usage of Filter

Our extension can filter videos based on predicted accessibility for YouTube search pages. Participants turned the filter on for 29.41% ($\sigma = 21.06\%$) of the searches during stage 2. Participants reported that they did not use the filter often because they want to look at all results they might be interested in:

> *"I think it's a cool feature and I could see people wanting to use it. I just didn't have enough faith, just in case I don't want you eating some results that I may actually want."* – P5

Other participants liked the filter because they felt already had an excessive amount of content to choose from, as P11 described: *"I just picked something above a five, cause there's too many to pick from."*

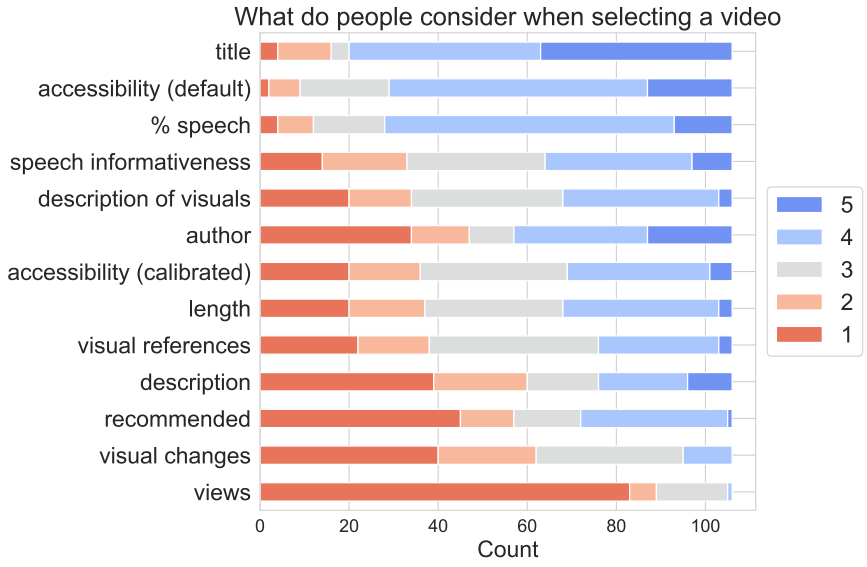



Figure 3.2: Participants' Likert-scale ratings (From 1 - not important at all to 5 - very important) to the question, *"Rate the importance of factors that made you decide to click on and watch the video."* Participants rated title, accessibility (default) and percentage of speech, to be the top-3 most important reasons.

### 3.5.2 RQ1: Do People Consider Accessibility Metrics in Practice?

The first research question we want to investigate is when and how much do people consider accessibility metrics given other interests and informational needs.

From experience sampling survey results in the second stage (where accessibility metrics were displayed), participants considered the *Title*, *Accessibility Score (default)*, *Percentage of Speech* to be the top three most important factors that affected their video selection process (Figure 3.2). Post-hoc Mann-Whitney U test shows that even when accessibility metrics were turned on, people still considered video title ($\mu = 4.03, \sigma = 1.11$), which participants used as a proxy of the expected content, as more important than video accessibility rating ($\mu = 3.80, \sigma = 0.87$) with statistical significance ($U = 7007, p < 0.005$). Participants further described how they would prioritize to select an interesting content first and then considered accessibility in different scenarios.

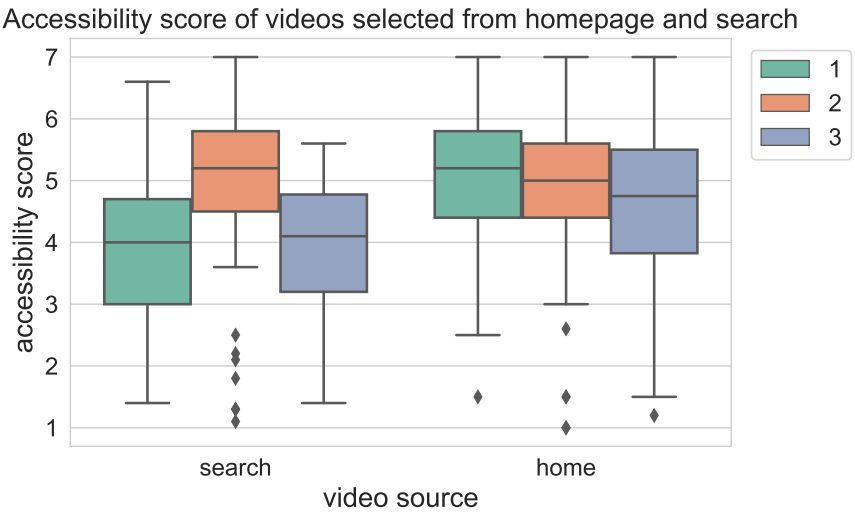



Figure 3.3: Boxplots of accessibility scores of videos participants click into from different source (search, homepage) in different stage 1, stage 2 and stage 3. We observe that participants clicked into more accessible videos from search (semi-specific) with accessibility metrics (stage 2). However when selecting videos from homepage (open-ended), accessibility of videos stays around the same.

**Theme 1: People first consider their video content needs, then consider accessibility when deciding between videos that meet their video content needs.** On content sharing platforms like YouTube, a content need might be met by many possible topics (open ended, *e.g.*casually browsing on homepage), a single topic (semi-specific, *e.g.*searching for a recipe), or a single video (specific, *e.g.*watching the latest update of a favorite sports team). In exit interviews, participants reported that different levels of specificity of what they are looking for affect how much they pay attention to the accessibility metrics alongside the video.

**(1) Open-ended: accessibility is considered.** When participants were casually browsing without a specific topic in mind, they would first choose to watch a video based on its title, author, or description. For example, P5 described that he primarily would choose not to watch a video based on aspects of the content (*e.g.* the presenter's microphone qual-

ity), rather than accessibility. But, accessibility *"is definitely a reason if it's just background music"* (P5). Similarly, P3 mentioned that he would still try inaccessible videos when the content looks interesting enough, and would watch the video even if the accessibility rating is low:

> *"...but if there was a really good title and there was no accessibility rating, or if it was low, I would still try it. I had experiences where they would get low ratings. But I would still, you know, subjectively feel an enjoyable experience."* – P3

This observation is further confirmed by our log data analysis. A Kruskal-Wallis H-test shows that accessibility ratings of videos that participants selected from homepage (*i.e.*open-ended browsing) do not significantly change with the introduction of accessibility metrics (stage 1 accessibility score: $\mu = 5.08, \sigma = 1.08$, stage 2: $\mu = 4.86, \sigma = 1.10$, stage 3: $\mu = 4.61, \sigma = 1.29$). Overall, most videos that the participants watched before using accessibility metrics were around a B rating (Figure 3.4), such that the level of accessibility before metrics (5.08 of 7 for homepage) may be considered adequate for video consumption while open-ended browsing. As P4 expressed: *"[On homepage] I don't feel like I'm not watching videos that are not accessible."* Thus, for browsing, people considered accessibility to the extent that they would "subjectively feel an enjoyable experience" (P3), but did not use the metrics to further watch videos with an even higher accessibility score.

**(2) Semi-specific: accessibility is a priority.** 8 out of 10 participants found accessibility metrics to be the most useful when they were looking for something semi-specific. Semi-specific queries usually happen when people are searching on YouTube, where the search results is comprised of videos about the same topic. Participants reported that they aimed to select the most accessible video in such cases:

> *"If I was searching for something a little more vague. You know, Brian Adams interview and a bunch of came up, I would pick the most accessible one because they're all the same. I want to watch the most accessible one. "* – P3

Similarly, P6 mentioned:

> *"If you're looking for like backpack reviews, you're going to, obviously have a dearth of selections to choose from. So it's not like if you're looking for like 'Jump' by Van Halen. A semi-specific piece of content, which could have multiple results and done differently. That's where I see it really being useful."* – P6

This result is also confirmed by participants' log data. With accessibility metrics, people selected more accessible videos from search results (stage 2 accessibility score: $\mu = 4.91, \sigma = 1.36$), compared to without seeing the accessibility metrics (stage 1: $\mu = 3.92, \sigma = 1.30$; stage 3: $\mu = 4.00, \sigma = 1.09$). Kruskal-Wallis H-test result shows a significant difference in videos' accessibility ratings between stage one, two and three (statistic= $22.7, p < 0.001$). With further Mann-Whitney U tests, we observed a significant increase from stage one to stage two (statistic = $816.5, p < 0.001$), for the accessibility scores of videos that participants selected from search results; and a significant decrease from stage two to stage three (statistic = $1401.5, p < 0.001$). The difference between stage 1 and stage 3 was not significant (statistic = $623.5, p = 0.35$) Participants selected more accessible videos with accessibility metrics than without when they are looking for something semi-specific (Figure 3.3).

**(3) Specific: accessibility is not considered.** When looking for a specific video (e.g., an assigned video course, video update from one's favorite sports team, etc.), participants did not consider accessibility metrics and chose to watch the video anyway. P5 described how he would always watch a video from his favourite baseball team regardless of its accessibility:

> *"I don't actually care about the accessibility because it's the Cubs. So I just want to see what the Cubs are doing. I actually, you know, the metrics, I didn't even use them. It wasn't even something that I needed because I knew what I wanted. I was going to watch it regardless."* – P5

P6 reported a similar experience in which he would not care about accessible and watch any video from his favorite rock band:

*"But if it's a specific thing, so if I'm looking for some interviews or documentaries of Van Halen, I don't care. I just want the song. I don't care how accessible or inaccessible it is. I just know. I know what I want."* – P6

**Theme 2: People prioritize accessibility more for informational content goals than for entertainment content goals.**

Participants' accessibility needs for some specific types of content may be more important than others:

*"If I'm using it for information, if I'm looking for a tutorial video or I'm looking for a recipe video or something, um, for my college, the accessibility is really important. "* – P9

4 participants mentioned a contrast between consuming the content for entertainment and consuming the content for information. When just using the video for entertainment, participants considered accessibility to be less important:

*"It depends on the context of what I'm trying to do with it. If it's just entertainment then most times I do not [care about accessibility metrics]. "* – P4.

This theme reflects Theme 1 as participants reported using open-ended video browsing primarily for entertainment (and considered selecting for accessibility less in both cases), and search primarily for finding information (and considered selecting for accessibility more in both cases).

### 3.5.3 RQ2: What aspects of accessibility are most important to people in practice?

Once a user decides that accessibility matters for this video they then proceed to consider which of metrics is more important. Specifically, we observed that participants divided the

metrics into two categories: (1) audio metrics that indicate how understandable the video is by listening to the audio track alone, and (2) visual metrics that estimate how inaccessible the unseen visuals are. In this research question we investigate what factors of accessibility are most important to BVI people in practice, under different scenarios.

**Theme 3: People primarily consider audio accessibility metrics.** Results from experience sampling surveys (Figure 3.2) show that participants considered *Percentage of Speech* ($\mu = 3.71, \sigma = 0.91$) and *Speech Informativeness* ($\mu = 3.04, \sigma = 1.16$) to be the top two important metrics our extension provides (besides the overall accessibility score). Visual features were rated as less important (*e.g.*participants only rated visual changes with $\mu = 2.14, \sigma = 1.04$). 8 out of 10 participants reported that an audio track with frequent speech was typically sufficient. As P13 reported: *"I would kind of look and see like what the metrics said. If there was a lot of speech, I was probably more likely to click on it."* (P13).

Although accessibility scores were fitted and tested with videos labelled by BVI people [54], participants in our study reported that they felt some videos had an overall accessibility score that was lower than their expectations. Participants used the fine-grained accessibility metrics to identify such cases:

> *"there was one that had a lot of speech. But it also had a lot of visual changes. So [the overall accessibility score] was quite a little bit lower. That's where the percentage speech was helpful"* – P1

Given the knowledge of the metrics used to calculate the overall accessibility score, others would predict that surprisingly low scores were due to undescribed visuals that they are not aware of:

> *"There are some things that they might've been even more accurate that I realized. Cause there could be a lot of visual stuff happening. I just have no idea what it is."* – P3

When encountering such situations, participants would watch the video if they believed that the audio itself would be coherent and understandable. In particular, participants used the type of content as a signal when considering whether videos would be useful from audio alone:

> *"Interviews and documentaries may have a lot of visual changes happening. But they're narrated very well. [...] So I'll still watch it."* – P12

> *"I watch a lot of music-related videos on YouTube, and the vast majority of what I watch is live shows or interviews or things like that. And so the visual accessibility information doesn't matter. "* – P3

Occasionally, participants would still be willing to watch videos with a low audio accessibility score. For example, P1 frequently watched video gaming streams even though the audio track lacked high-quality narration:

> *"Like video games, they're not going to have very a good scripture anyway, I just try to enjoy the sound effects."* – P1

**Theme 4: People also consider visual accessibility metrics when pursuing informational goals** Visual metrics can be important when the visuals contain essential information for users' needs or tasks. For example P5 described a case where he found a video with a high percentage of speech left out visual information that was important for an informational goal:

> *"So you went into the video, but then the maybe 2% that wasn't audio described, might've been something I would've liked to see. For instance, there was one video it was like how to tie your shoes in a way that they wouldn't come on down. And it sounded really interesting. So the video was a hundred percent described and at the very end, the guy said, 'here's how to'."* – P5

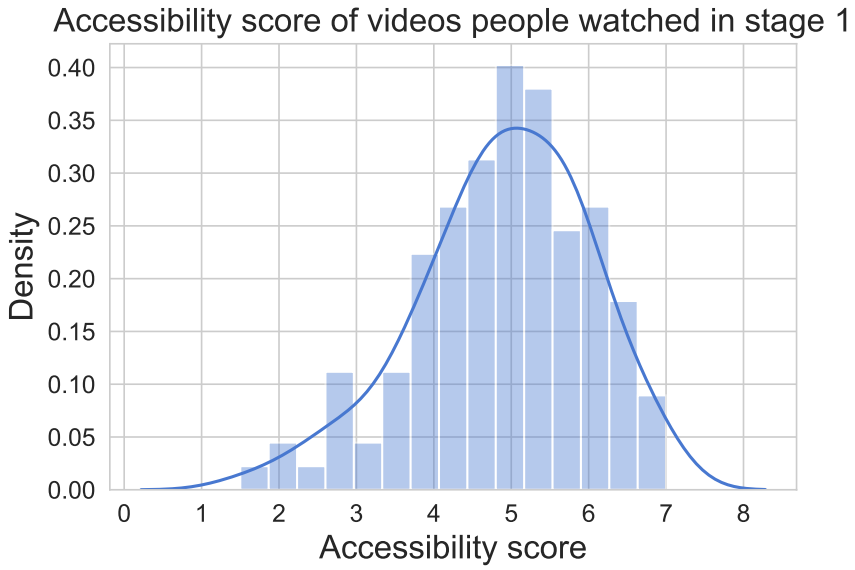



Figure 3.4: Histogram of accessibility scores (from 1 - not accessible at all to 7 - very accessible) for all selected videos when not having accessibility metrics available. People were already picking more accessible videos even without our extension.

P9, similarly, noted that he would leverage visual metrics for informational videos that can be especially visual:

> *"Especially with something as my hobby is rock tumbling, and some of it could be kind of visual. So it [visual-related metrics] was useful to figure out what videos would be helpful for me to watch and what worked. "* – P9

### 3.5.4 RQ3: How does knowledge of accessibility metrics impact people's browsing and searching behavior?

YouTube is a content sharing platform that provides multiple ways for people to find videos. We observed that participants were already watching accessible videos (Figure 3.4, $\mu = 4.89, \sigma = 1.11$) on YouTube without accessibility metrics via different ways: (1) Primary homepage users (P1, P4, P6, P8, P9, P12, P13). Participants had their homepage curated

with familiar authors and with consistently accessible content. (2) Non-homepage users (P3, P5, P11). Three participants reported rarely using homepage before, and mainly relied on search to find accessible video types (*e.g.*, interview, comedy). We report on how accessibility metrics impact people's existing video foraging strategies.

**Theme 5: Accessibility metrics enable people to more actively explore and venture out.** A theme emerged from our analysis is that accessibility metrics enabled people to venture out and diversify their (1) ways to find videos, (2) video content they watch and (3) what do they use YouTube videos for.

**(1) Accessibility metrics enable people to use new approaches to find videos.** Participants who relied on homepage before reported that although they did not find the metrics to be especially helpful on their homepage, they started to explore beyond their curated lists. From our log data, we observed a trend that primary homepage users used YouTube search more often with accessibility metrics available (frequency of search in stage 2: $\mu = 0.233, \sigma = 0.210$) than they did without the accessibility metrics (stage 1: $\mu = 0.202, \sigma = 0.178$; stage 3: $\mu = 0.188, \sigma = 0.201$) (Figure 3.5). 5 out of 7 primary homepage participants increased their search frequencies and reduced their homepage visits with accessibility metrics:

> *"I did use it for some searches and that was where it's the most useful. I can now look up YouTube videos more confidently."* – P8

And after turning off the extension stage three, they started to feel less confident in using search and used search less:

> *"[In stage 3] I didn't really want to search very much. I felt less confident in doing all that. I searched less but the homepage was still all right. "* – P1

In addition, 3/3 participants who only relied on search started to explore their homepage recommendations that was not previously curated (Figure 3.5). P3, P5 and P11 used home-

page more often with accessibility metrics (stage 2: $\mu = 0.403, \sigma = 0.041$), than without (stage 1: $\mu = 0.256, \sigma = 0.084$; stage 3: $\mu = 0.349, \sigma = 0.157$). P3 mentioned how he only used YouTube for searching specific types of videos he had previously found to be accessible (*e.g.*musician interviews), and accessibility metrics allowed him to explore interesting content on homepage that he never tried before.

> *"The homepage is a huge example of something that I'm able to browse more efficiently and view a lot more diverse content. I never really used YouTube that way before. I typically only get on YouTube to search for something if I'm in the mood for something specific."* – P3

P5 and P11 used less homepage after turning accessibility metrics off again. They reported starting running into videos with just background music:

> *"When I had it turned off in stage three, there were times that I found videos on homepage that I clicked on and played and it was all music and somebody showing something that was like, oh, well the plugin would have worked."* – P11

**(2) Accessibility metrics enable people to watch more diverse video content.** As a result, participants found that they were able to watch videos with new topics. With accessibility metrics, participants who used homepage discovered new videos in search that was beyond their typical recommended content:

> *"The search page does give you a different, you know, a lot more diversity in terms of the content, as opposed to your typical home page or your recommended content from your channels. I'm a little more adventurous in terms of what videos I watch."* – P4

As for participants who only searched for videos, they felt that the homepage recommendations provided more possibilities:

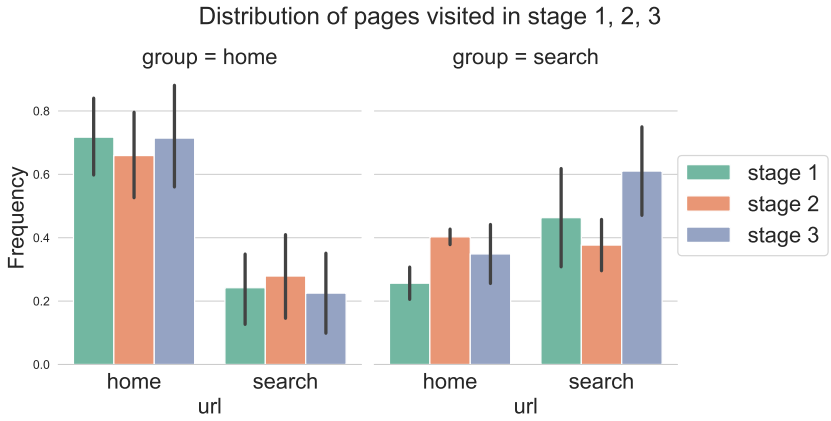



Figure 3.5: Percentage of video foraging approach (page visits) of primary homepage and search users in stage 1, 2, 3 of the study. With accessibility metrics people tend to explore new approaches to find videos.

> *"The homepage is a huge example of something that I'm able to browse more efficiently, um, and, and view a lot more diverse content."* – P5

**(3) Accessibility metrics enable people to use videos in more diverse ways.** The knowledge of accessibility metrics empowers people with more choice to use YouTube. Participants reported being able to use YouTube for new tasks that they would not before look on YouTube:

> *"Oftentimes I'll look up articles as tutorials before I look up YouTube videos. But now I can also look up YouTube videos more confidently because the metrics that are telling me how accessible that YouTube tutorial is going to be, which is really nice."* – P13

Similarly, P12 who only followed his homepage recommendations and used YouTube for entertainment, started to use YouTube videos for researching a new hobby:

> *"I did start to use it for some searches. I was researching a hobby that I was thinking of getting into and I use that for that. And that was interesting and*

*useful."* – P12

## 3.6 Discussion

Our work is the first in-situ study to understand how people leverage accessibility metrics within the broader context of their everyday searching and browsing activities. Through the four week deployment of our extension, we observed whether and how much BVI people consider accessibility when selecting a video, what aspects of video accessibility metrics do they value, and how does knowledge of accessibility metrics impact their video foraging strategies. We further elaborate on our findings and highlight three trade-offs people consider when selecting to watch a video with accessibility metrics: (1) when to prioritize content vs. accessibility, (2) when to prioritize perceivable accessibility vs. full accessibility, and (3) when to exploit known searching and browsing techniques vs. exploring new approaches.

### 3.6.1 Content vs. Accessibility

Our work demonstrated that BVI users considered a video's accessibility only after they first determined if a video met their content goals. In addition, their further consideration of accessibility depended on the specificity of their content goal (*e.g.*, open-ended, semi-specific, specific video) (Theme 1).

In open-ended browsing tasks, that considered videos with diverse content, participants would select a video primarily for its content (*e.g.*, title and author), and then would consider accessibility as a secondary criteria (*e.g.*, is this video accessible enough to be enjoyable?). This finding that when browsing videos in-the-wild users first prioritize content then consider accessibility contrasts a prior in-the-lab study finding that in a webpage browsing task users would first prioritize links annotated as accessible, then search among only those links for content of interest [95, 92, 94]. In practice, once users choose a video with relevant content, if the video is accessible enough to be enjoyable, they typically watched the video without

conducting an additional search for a more optimal alternative (*i.e.*"satisfing" [103]). Information foraging theory suggests that people satisfice when it is expensive, *e.g.*in terms of time and efforts, to perform an additional search [73]. On the other hand, we found that for semi-specific (search) tasks that are often informational, participants used search to find many videos that similarly met their content goal (*e.g.*, how to tie shoes), so they considered accessibility as a primary factor to pick between the options — echoing prior work that finds users prioritize accessibility during search tasks [95, 92, 94, 54]. Thus, while filtering out inaccessible videos may support users when searching [54], future work should consider alternative ways to make use of accessibility metrics for open-ended browsing. For example, to reduce the time required to find a more accessible content, future systems may allow users to select a content of interest then quickly preview accessible alternatives with similar content.

### 3.6.2 Perceivable Accessibility vs. Full Accessibility

Participants in our study reported that they primarily considered audio accessibility metrics (Theme 3), and only considered visual accessibility metrics when missed visuals could impact their informational goals (Theme 4). Thus, participants prioritize an experience of the video that they perceive to be accessible (*i.e.*the video is seemingly understandable or enjoyable from the audio alone), and consider full accessibility (*i.e.*all important visuals are described including unknown unknowns) primarily for informational tasks.

Prior work had previously observed a discrepancy between what users perceived to be inaccessible vs. what was not possible on a website [17], and between predicted accessibility metrics and user perceptions of accessibility [54, 80, 93, 95]. Our work demonstrates that by providing the lower level information about what contributed to the accessibility score, users were able to use components of the audio accessibility (*e.g.* percentage of speech) to make decisions about what to watch, and infer that a lower than expected score may mean that there is visual content that is inaccessible. Future systems that help users surface accessible content via accessibility metrics or scores could enable users to toggle parts of

the accessibility scoring on or off so that the overall score reflects their preferences. In addition, future systems may consider providing two different aggregate scores: a "perceived accessibility", or quality of experience score (similar to audio accessibility metrics in our case), and a "full accessibility", or completeness of important information score (similar to visual accessibility metrics in our case).

### 3.6.3 Exploitation vs. Exploration

Participants in our study were already exploiting their known strategies to watch videos that were accessible to them before using accessibility metrics. For example, participants would stick to tightly curated homepage recommendations that included creators they knew to be accessible, or use known keywords to find an accessible video during search [54]. This is similar to prior work that demonstrated blind users searched and browsed the web similar to sighted users but visited pages with less inaccessible content [16]. Prior work in accessibility metrics show in lab studies that the knowledge of accessibility of search results helps people more efficiently browse and search for accessible content [15, 54, 95]. From interviews and log data we find that for BVI users, a primary benefit of having accessibility metrics is reducing potential accessibility barriers for content that they were not familiar with. Participants reported that accessibility metrics improved their confidence in being able to find an accessible video efficiently, and thus their willingness to use new strategies to find videos, and use YouTube videos for new purposes (Theme 5). Future work should consider how to expand the presence of accessibility metrics to other forms of content to promote ease of finding content of interest. For example, adding accessibility metrics to applications may help users efficiently find the most accessible version of a common application type (*e.g.*, a to-do list application). Future work could also explore augmenting search with metrics that support people with other accessibility needs such as a specific reading level to enable people to make informed decisions when selecting between multiple content options. Content platforms could obtain such accessibility metrics via manual annotation, user feedback, or

developing automated approaches based on feedback or ratings from people with disabilities to make searching and browsing more useful to all.

## 3.7 Conclusion

In this chapter, we investigate people who are blind or have visual impairments use accessibility metrics within the broader context of their everyday content searching and browsing activities. We present a deployment study of a browser extension that adds video accessibility metrics onto YouTube. We find that when selecting to watch a video, BVI people first considered their content needs, and then considered accessibility need when deciding between videos that meet their content needs. Participants also reported they primarily considered audio accessibility metrics and only considered visual accessibility metrics for informational tasks. Additionally, with accessibility metrics, participants diversified their video selection approaches, content types and use cases of videos. Based on these findings, we highlight three trade-offs BVI people consider when incorporating accessibility metrics in their content foraging process, and identify future design opportunities.

# CHAPTER 4

# Identifying Video Accessibility Issues via Cross-modal Grounding

## 4.1 Introduction

In previous chapters, we have explored systems to help blind and visually impaired video audience surface accessible videos. In this chapter, we aim to enable content creators to make more accessible videos at the authoring stage. To make videos accessible, video authors can add audio descriptions (AD) that describe important visual content, and closed captions (CC) that transcribe the speech and non-speech sounds. However, to identify parts of the video that require further description, authors must manually watch the video all the way through, playing and pausing to check if: (1) the important visuals have not been described in the audio (*e.g.* a travel montage in a vlog), and (2) the important audio is not present in the visuals and captions (*e.g.* a door slams off-screen). This process of identifying inaccessible video segments is challenging and time-consuming, especially for video accessibility novices.

To guide authors to describe inaccessible video segments, existing audio description authoring tools surface "gaps in speech" as a proxy for moments where the visuals are unlikely to be verbally described [99, 68, 107]. However, many video genres including tutorials, vlogs, and lectures may not feature significant gaps in speech [54, 69], and audio description guidelines as well as prior research [99, 54, 20] indicate that visuals can be inaccessible to blind and visually impaired (BVI) people even when there is accompanying speech. For example, a speaker may make an ambiguous verbal reference to visual content (*e.g.*"make sure to

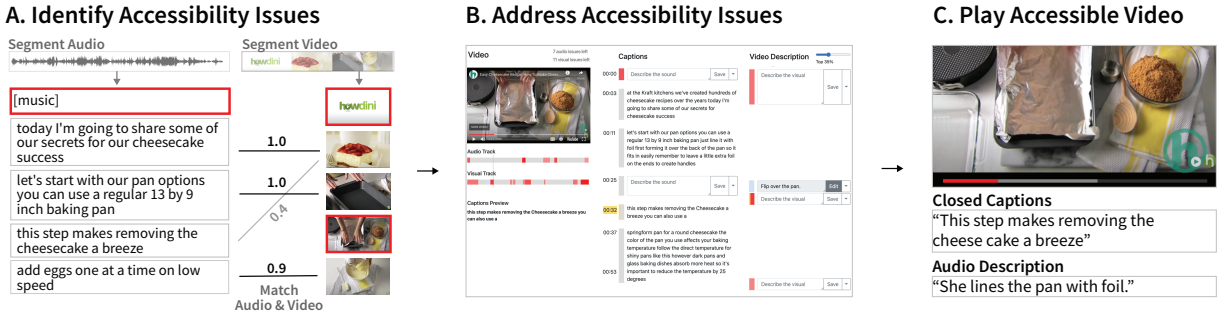



Figure 4.1: Our system (A) identifies accessibility issues by locating *modality asymmetries* (in red) between audio segments and video segments using cross-modal grounding. (B) lets authors address accessibility issues using the CrossA11y interface to write captions and video descriptions, and (C) creates a more accessible video from the authored descriptions.

have *these* before you get started") or share a personal story while demonstrating a tutorial step. In addition, visuals without speech can be accessible if they are understandable from non-speech sounds alone. Thus, using gaps in speech alone, authors will miss important inaccessible moments or be prompted to describe already-accessible moments. Similarly, caption authoring tools [50, 77, 26] let authors correct errors from automatic speech recognition (ASR), but they fail to surface moments when important audio does not also appear on-screen (*e.g.*someone leaves and we hear a door slam).

To help authors efficiently identify and address audio and visual accessibility problems, we present CrossA11y. CrossA11y surfaces asymmetries between the visual track and the audio track, or *modality asymmetries*. By identifying moments in the visuals that are not available in the audio, CrossA11y surfaces moments that are not accessible to blind and visually impaired (BVI) audience members. Similarly, by identifying moments in the audio that are not available in the visuals, CrossA11y surfaces moments that are not accessible to d/Deaf and Hard of Hearing (DHOH) audience. To automatically identify modality asymmetries, CrossA11y's computational pipeline segments the audio and visual tracks and uses *cross-modal grounding* to identify mismatches between the two tracks (Figure 4.1A).

CrossA11y then displays the results in an interface where authors can jointly author closed captions and audio descriptions by easily navigating to inaccessible moments (Figure 4.1B). Authors can then preview and export their resulting audio descriptions and closed captions (Figure 4.1C).

We evaluated CrossA11y in a user study with 11 video authors creating captions and audio descriptions for four videos. Authors more efficiently authored audio descriptions and captions with better precision and recall in addressing accessibility issues when using CrossA11y's modality asymmetry predictions than without these predictions. We also invited two video authors who frequently posted videos to YouTube to use CrossA11y to make two of their own videos accessible, and they reported that they would use CrossA11y in their workflow to produce more accessible videos.

In this chapter, we contribute:

- A pipeline to compute accessibility scores of the visual and audio segments of a video by checking for *modality asymmetries* via cross-modal grounding.
- A unified tool that helps video authors to locate and address visual and auditory accessibility problems of a video.
- A user study demonstrating that CrossA11y improves people's efficiency and reduces their mental demand in identifying accessibility issues.

## 4.2 Related Work

### 4.2.1 Authoring AD and CC

Prior work aims to help people manually author audio descriptions with task-specific authoring tools [41, 18, 1], feedback on the content at production-time [70], with feedback on audio descriptions [82], and with hosted descriptions [41]. Since authoring descriptions is a

time-consuming process, other prior work seeks to provide computational support for this task including using: computer vision to detect visual content [31, 30], using deep learning to provide a computer-drafted description [31, 107, 99], synthesized voice to convert text to speech [30, 47, 48, 89], and automatic editing to fit human-authored descriptions into the space provided [68]. While focusing on methods to help people write better descriptions, such tools only find inaccessible moments for description by surfacing silent portions of the video [18, 68, 107, 30], or by helping people find film-specific visual content that may need descriptions (*e.g.*, scene changes, characters [30]). Rather than assessing video in a single modality, we explore to find accessibility problems by assessing asymmetries between the auditory and visual content.

Current caption authoring tools [50, 77, 26] transcribe speech and allow creators to correct the transcript. In this work, we also surface non-speech sounds to facilitate caption authoring and add modality matching score to help people prioritize points where additional description could be needed (*e.g.*a sound that happens off screen may be highly important to describe, while a silent section would not be).

### 4.2.2 Assessing Audio and Visual Similarity

Recent work in unsupervised cross-modal machine learning explores learning a joint embedding space for information in different modalities, including text and images [76, 100, 87], text and video [60, 10, 24, 98], and audio and video [10, 64, 11]. Such models enable comparison between any visual, text, or audio segment. While these models can be used for retrieval across modalities (*e.g.*, text-image retrieval [76, 100], and text-video retrieval [24, 98]), we use a cross-modal approach to inform authors of accessibility issues due to low correspondence between the modalities, or modality asymmetry.

Prior video work in video accessibility has also considered the similarity between video and audio tracks. Wang et al. filter possible accessibility problems first by gaps in speech then use video and audio similarity to prioritize what non-speech segments to describe [99].

Liu et al. checks if detected objects is mentioned in the transcript, along with other metrics, then assigns an accessibility score to a video to help blind viewers find accessible videos [54]. We instead compute the fine-grained similarity between audio and visual segments to help authors find accessibility issues outside of gaps in speech that have not yet been addressed by prior work.

## 4.3 CrossA11y Interface

CrossA11y enables authors to efficiently identify and address visual and auditory accessibility problems in videos. The interface consists of three main components: 1) a *video pane* that lets authors navigate via an audio segment timeline or video segment timeline to identify inaccessible video segments (Figure 4.2A), 2) the *video description pane* that lets authors identify and address visual accessibility problems (Figure 4.2F), and 3) a *caption pane* that lets authors address auditory accessibility problems and navigate the video with a time-aligned caption transcript (Figure 4.2E).

### 4.3.1 Video Pane

The *video pane* (Figure 4.2A) displays the video and lets authors play/pause the video and seek within the video using two timelines: (1) the audio timeline that lets authors navigate to auditorily inaccessible segments, and (2) the visual timeline that lets authors navigate to visually inaccessible segments. The audio timeline displays audio segments that each represent a segment with continuous speech, or non-speech sound. The visual timeline displays visual segments that each represent a segment of continuous footage (i.e. a shot). Each segment is colored with its estimated accessibility[1] from grey (accessible) to red (inaccessible) using sRGB inverse gamma mixing. Using either timeline, authors can gain an overview of accessibility issues, or quickly navigate to an inaccessible segment by clicking on

[1]Computed using the cross-modal grounding pipeline as described in Section 4.4

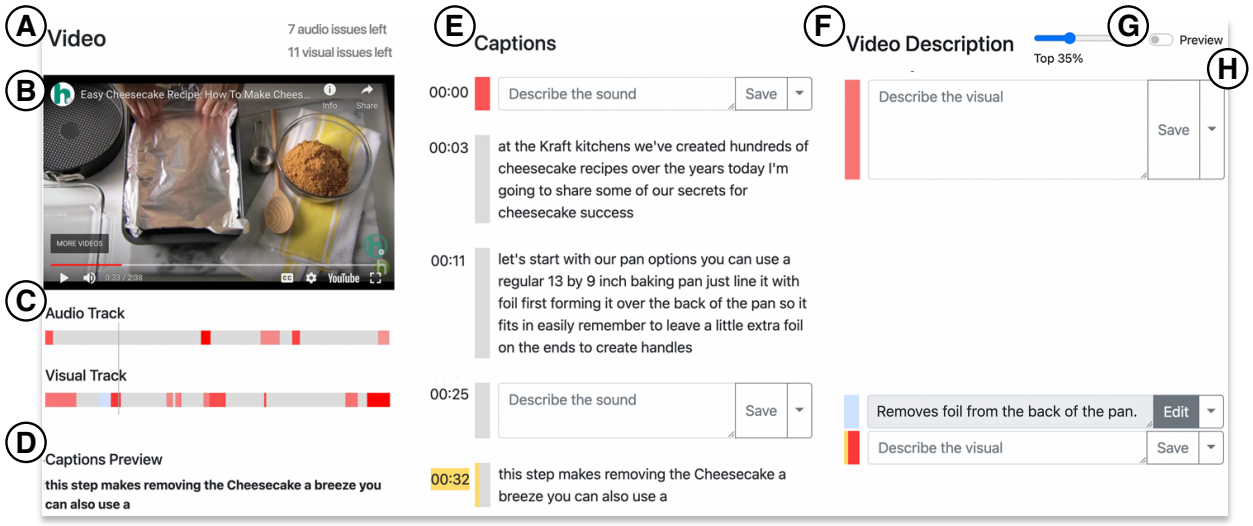



Figure 4.2: In CrossA11y's interface, the *video pane* (A) displays audio and visual timelines with accessibility visualization that allows authors to quickly identify and navigate to accessibility issues. The *video description pane* (F) surfaces inaccessible visual segments and lets authors to add descriptions. The *captions pane* (E) provides time-aligned captions and detected non-speech sound segments for authors to seek within the video and add captions.

the segment to play the corresponding point in the video. For example, by clicking on the first red segment in the audio track (Figure 4.2C) an author will hear an inaccessible audio segment — music plays that is not available in the captions preview (Figure 4.2D). Authors may inspect inaccessibility prediction results displayed in the timeline by hovering over an audio or visual segment to see the audio segments that are predicted to match that segment (displayed with higher opacity). As the author navigates and plays the video with the video pane, the corresponding segments are highlighted in the linked video description pane and the caption pane.

### 4.3.2 Video Description Pane

The *video description pane* (Figure 4.2F) lets address inaccessible visual segments by writing text descriptions of the visual content. Each video description segment consists of a vertical side bar that is colored according to its predicted accessibility, an editable text field where an author may add descriptions, and "save"/"edit" and "dismiss" buttons to address or ignore surfaced visual accessibility issues. Video description segments are relatively aligned with the caption segments in the *captions pane* such that authors can preview the nearby narration. The height of each video description segment represents its relative length such that authors can estimate the approximate length of description required.

When an author locates a visual accessibility issue (*e.g.*, the last displayed segment in Figure 4.2H), the author can click the segment to play the clip and to check if the visual content is described in the audio or existing descriptions. For example, in this case, the author may notice that the host placing the foil inside of the pan is not yet described in the captions, and add a description by typing "Flips the and over and put the foil inside" and clicking "Save". The vertical side bar, and the corresponding segment in the video pane's visual timeline, then change to blue to indicate the issue has been addressed. If an author decides that a suggested visual accessibility issue does not need description, they can dismiss the problem. The vertical side bar for that segment and corresponding horizontal bar in the visual track timeline will turn dark grey to indicate the issue has been dismissed. Authors can manually add a description to a point in the video where an accessibility issue was not detected by double clicking the visual segment in the video pane's visual timeline to create a corresponding video description segment in the video description pane.

By default, the video description pane displays visual segments with an estimated accessibility scores lower than 0.35 (range 0-1). Authors can use the slider (Figure 4.2G) to surface more visual accessibility problems when making sure they covered everything, or fewer accessibility problems when prioritizing for a time constraint.

### 4.3.3 Captions Pane

The *captions pane* lets authors navigate the video with a time-aligned transcript and write captions to address inaccessible audio segments. CrossA11y automatically provides captions for the speech, so authors can focus on making non-speech sounds accessible. Each caption pane segment has a similar structure with video description segments. Authors can quickly locate, review, script captions in-place, or dismiss a suggested problem.

The caption pane displays predicted audio accessibility by coloring the vertical bars (similar to the video description pane) to help authors understand and prioritize audio accessibility issues. Unlike the video description pane, we do not use predicted audio accessibility to filter audio accessibility issues as Closed Caption guidelines state that all important sounds should be synchronously described whether or not they can be inferred from the visual content alone [2, 6]. Using the captions pane, authors can click on a caption segment to hear the segment, then script a caption if the sound is important (music at 0:00 in Figure 4.2E) or dismiss the segment if the sound is not important (silence at 00:25 in Figure 4.2E).

### 4.3.4 Accessible Video Preview

After authors create captions and video descriptions for inaccessible segments, they can then preview their results as the video plays. The video's original captions and author-created captions are displayed under "Captions Preview" (Figure 4.2D). Audio descriptions are synthesized via a text-to-speech engine (Web API's SpeechSynthesisUtterance Interface[2]). Our system renders audio description in the format of *extended description* [97], which pauses the video, plays the synthesized speech descriptions, and continues the video.

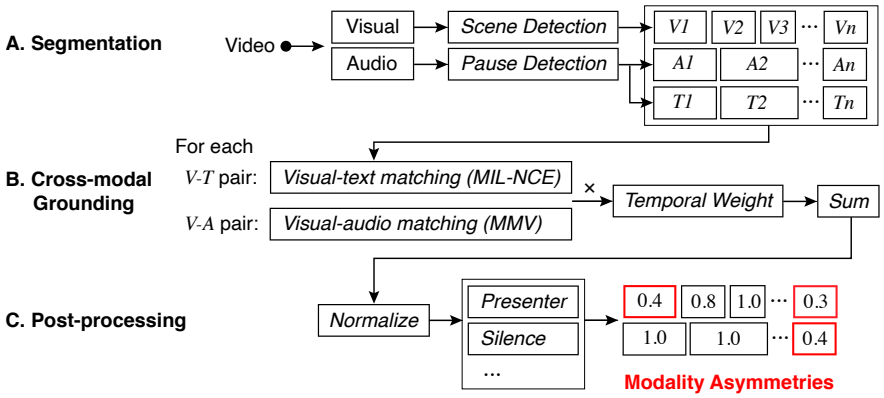



Figure 4.3: Our computational pipeline includes: (A) Segmentation that segments the video into audio and visual segments, (B) Cross-modal Grounding that finds correspondences between visual and audio segments, and (C) post-processing that filters identified correspondences.

## 4.4 Cross-modal Grounding Pipeline

We present a computational pipeline that segments the auditory and visual track of the video (Figure 4.3A) and identifies asymmetries between auditory and visual tracks using cross-modal grounding analysis (Figure 4.3B).

### 4.4.1 Segmentation

To create visual segments, we detect shots, or segments with continuous footage. To segment shots, we used scenedetect[3]'s content-aware scene segmentation algorithm that compares the HSV color space in adjacent frames against a threshold to determine if the two segments belong to the same shot. To create audio segments, we follow prior work [54, 68] by aligning the transcript and audio using Gentle forced-aligner [66] to get word-level timings, and then consider any gap between words longer than 2 second to be a non-speech audio segment, and

[2]https://developer.mozilla.org/en-US/docs/Web/API/SpeechSynthesisUtterance

[3]http://scenedetect.com/

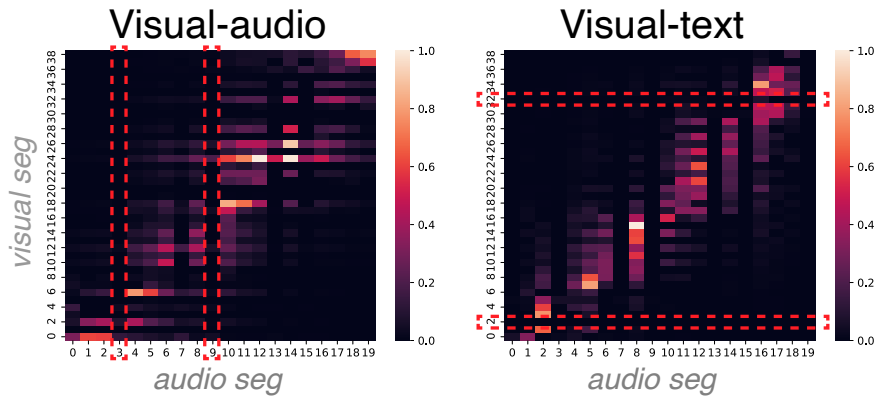



Figure 4.4: Example of visual to audio similarity matrix, and visual to text (transcript) similarity matrix. Red dotted lines highlight examples of asymmetric and potentially inaccessible segments.

any gap longer than 0.5 second but shorter than 2 seconds as a pause in speech. We segment the audio into speech and non-speech audio clips based on the gaps. In addition, we also generate a list of segmented transcripts according to the start/end of audio segments.

### 4.4.2 Cross-modal Grounding

To assess if each visual segment and audio segment is described in the other modality, we compute visual-text and visual-audio matching scores for all video and audio segment pairs using multimodal machine learning algorithms.

#### 4.4.2.1 Visual-text/audio Matching

CrossA11y uses multimodal machine learning algorithms which learn a symmetric joint embedding space for visual and auditory or textual data. With the embeddings, we can measure if a visual segment and a audio segment is semantically similar by computing the dot product of visual embeddings and audio/text embeddings. Specifically, for visual-text matching, we use the MIL-NCE model [60] which was trained on HowTo100M, a dataset of 100 million

clips-narrations from YouTube. For visual-audio matching, we use the state-of-the-art MultiModal Versatile (MMV) networks [10] which was trained on AudioSet, a dataset consists of 10 seconds clips coming from 2 million different internet videos. AudioSet contains a variety of audio types including musical instruments, animal, mechanical sounds, etc.

By matching all $n_v$ visual segments to $n_a$ audio segments in a video, MIL-NCE and MMV each produces a $n_v \times n_a$ matrix, where cell $(i, j)$ is the matching score for visual segment $v_i$ and audio segment $a_j$. Each matrix is normalized to range $0 - 1$. Figure 4.4 displays examples of such cross-modal grounding results.

To estimate the accessibility of a visual segment, $v_i$, we compute its matching scores to all audio segments in the same video. When matching to audio segments that contain speech, we use the MIL-NCE score (since match different visuals to human speech sound does not make sense). We remove stop words of all transcripts before computing its correspondence to visuals. When matching to audio segments without speech, we use the MMV score. We then compute a weighted sum of all scores based on each audio segment's temporal position to $v_i$ (as explained in Section 4.4.2.2). Thus, for a video with $n_v$ many visual segments and $n_a$ audio segments:

$$\text{score}(v_i) = \sum_{j}^{n_a} w_{i,j} * \text{matching}(v_i, a_j) \tag{4.1}$$

where,

$$\text{matching}(v_i, a_j) = \begin{cases} \text{MIL-NCE}\left(v_i, t_j\right), & \text{if speech.} \\ \text{MMV}\left(v_i, a_j\right), & \text{if non-speech.} \end{cases} \tag{4.2}$$

Similarly, to estimate the accessibility score of an audio segment, $a_j$, we compute its degrees of matching to all visual segments. If the current audio segment is non-speech, we only use the MMV score. However, when checking whether the content of a speech audio is presented in the visual, it is prevalent that the speech is transcribed and added as subtitles

using automatic speech recognition technology. Even if subtitles of speech are not added, systems can quickly apply ASR to incorporate them into the visual modality. Thus, if an audio segment is detected as speech, we will assign it a constant value $c$ and consider it as accessible (since the speech information is displayed as subtitles in visual). For an audio segment, $a_j$:

$$\text{score}(a_j) = \begin{cases} c, & \text{if speech} \\ \sum_i^{n_v} w_{j,i} * \text{MMV}\left(a_j, v_i\right), & \text{if non-speech.} \end{cases} \tag{4.3}$$

Note that because both the MIL-NCE and MMV scores are symmetric (e.g., $\text{MMV}\left(a_j, v_i\right) = \text{MMV}\left(v_i, a_j\right)$), we only need to conduct matrix multiplications one time to compute scores for both visual and audio segments.

#### 4.4.2.2 Temporal Weighting

A visual segment can be matched to an audio segment that is far away from each other in time. In such cases, although the information is grounded in the other modality, it would be hard for people to connect and make sense of such cross-reference. Thus, as shown in the above equations, we apply a temporal weighting to the output visual-text and visual-audio matching scores. Specifically, the matching between a visual segment $v_i$ and an audio segment $a_j$ diminishes exponentially by a factor of $w$ ($0 \leq w \leq 1$) for every 5 seconds' distance in time:

$$w_{i,j} = w_{j,i} = w^{\frac{\left|TS_i - TS_j\right|}{5}} \tag{4.4}$$

Where $TS_i$ is the timestamp in seconds of segment $i$, and $w$ is the weighting factor. We empirically found that $w = 0.45$ works well.

### 4.4.3 Post-processing

After we compute segment visual accessibility scores, score($v_i$), and audio accessibility scores, score($a_j$), for the video, we normalize the scores into $0-1$ ranges. We then remove commonly detected accessibility issues that do not need further description including: the presenter speaking to the camera, and silences.

#### 4.4.3.1 Presenter Speaking

In initial tests, our approach detected moments where the host was speaking to the camera to be inaccessible, leading to low precision for how-to and recipe videos like [40] (precision=0.356, recall=0.875) and [57] (precision=0.435, recall=0.929). While the visual content and speech did not match, these segments were accessible as they could be implied from the audio alone (the presenter's voice). To address this, we detect faces using OpenCV[4], and compute the area of the detected face bounding box per second for each visual segment. We consider a visual segment to be "presenter speaking" and thus not inaccessible if the area per second metric is greater than a threshold $TH_{presenter}$. We empirically determined a threshold $TH_{presenter} = 58000$. With presenter detection, the precision improved to 0.636 on [40] and 0.867 on [57].

#### 4.4.3.2 Silences

Similarly, our initial approach predicted segments with silent or insignificant audio to be inaccessible as the quiet noises were not detected to match the visuals. For instance, the algorithm considered a scene in a recipe video [40] where the host flips the pan over making minor noises, and a scene in a food review video [78] where the host was eating and making a chewing sound, as inaccessible. To address the issues, we detect silences by computing the

[4]https://docs.opencv.org/3.4/db/d28/tutorial_cascade_classifier.html

| | *Visual* | | | *Audio* | | |
|---|---|---|---|---|---|---|
| | Random | Gaps | **CrossA11y** | Random | Gaps | **CrossA11y** |
| Precision | 0.275 | 0.833 | **0.694** | 0.125 | 0.909 | **0.983** |
| Recall | 0.390 | 0.385 | **0.984** | 0.381 | 0.843 | **0.843** |
| F1 | 0.323 | 0.526 | **0.814** | 0.188 | 0.874 | **0.908** |

Table 4.1: CrossA11y's experimental test results on a sample of 20 manually labeled videos. CrossA11y generally performs better than random guess and using "gaps in speech" as heuristics, for both detecting visual and auditory accessibility issues.

average intensity of audio segments using librosa[5] and compare it to a threshold $TH_{silence}$, which we empirically set to 0.007. If the average intensity score is lower than the threshold we label this audio segment as insignificant and thus not inaccessible.

#### 4.4.3.3 Threshold

As discussed in 4.3.2, CrossA11y displays visual segments with grounding score larger than a threshold as the visual accessibility issues. We selected 0.35 empirically as it worked consistently well for diverse videos, and favored recall over precision such that the system showed more potential accessibility issues. Authors may easily dismiss false accessibility issues using the "Dismiss" button.

### 4.4.4 Technical Evaluation

We evaluated CrossA11y's cross-modal grounding pipeline using 20 randomly selected videos from YouDescribe[6], a platform where people can request audio descriptions for YouTube videos. In particular, we limited our random selection to videos that were less than 5 minutes

[5]librosa.org

[6]https://youdescribe.org/

with captions available (implies some narration). All 20 videos were not tested when we built the system, i.e. out-of-bag, and they covered diverse topics: how-to (5 videos), recipe (4 videos), vlog (3 videos), campus tour (3 videos), documentary (2 videos), educational (2 videos) and review (1 video).

Two researchers independently identified visual and auditory accessibility issues in the videos based on guidelines [2, 20]. 70.78% of initially identified issues were the same. Researchers discussed the remaining labels until agreement was reached. In total, the sample included 182 visual accessibility issues and 79 auditory accessibility issues. We then used CrossA11y to predict visual and auditory accessibility issues in these videos. We also predicted accessibility issues using two baselines for comparison: (1) mark each segment as inaccessible with 50% chance (Random) , and (2) mark each segment as inaccessible if it did not include speech (Gaps) following prior work [68, 107, 99]. To assess if a labeled visual or auditory accessibility issue was accurately detected, we compared the start and end times of all segments that were predicted to be inaccessible with the start and end times of manually labeled inaccessible segments. We defined a prediction as accurate if there was a $> 50\%$ overlapping manually labeled accessibility problem.

Using the thresholds we selected in Section 4.4, CrossA11y achieved a higher F1 score compared to the baseline methods (Table 4.1). For visual accessibility issues, the recall score increased significantly from 0.385 with gaps in speech to 0.984 with CrossA11y, meaning that CrossA11y identified visuals with accompanying speech that are still inaccessible to BVI audience members. The precision decreased from 0.833 with gaps in speech to 0.694 with CrossA11y. For CrossA11y, we prefer high recall (the ability to show all issues) over high precision (the ability to show few incorrect issues), as authors may easily review and dismiss inaccurate issues (false positives), but they may struggle to find issues that we do not surface (false negatives). CrossA11y identified auditory accessibility problems with higher precision (0.983) compared to gaps in speech (0.909), as CrossA11y removes false positive issues when the gaps in speech correspond to silence. The recall remains the same as all

accessibility issues in our sample occurred during gaps in speech.

### 4.4.5 Limitations

From our technical evaluation, we discuss some limitations of the current implementation of CrossA11y.

#### 4.4.5.1 Segmentation Limitation

We noticed that algorithms in some cases failed to segment visual and audio tracks into semantically coherent segments. For visual, the shot detection algorithm would sometimes segment the same visual with different filming angles into different shots. This leads to lots of repetitive segments that would be annoying for authors to dismiss. In addition, the algorithm sometimes consider a long shot with different pieces of information as one large segment. This is especially common for tutorial videos that only has one shot angle (*e.g.,* an origami tutorial where the camera is always facing the table).

Similarly, the pause detection algorithm also in some cases produce an disproportionally long (*e.g.*, a host speak very fast with no pause) or short (*e.g.*, a host talking slowly very demonstrating a step in a how-to video) segments. Our algorithm also does not address overlapping sounds like a sound effect that is covered by speech, since audio source separation still cannot produce desirable results and often require training on specific examples.

Moreover, visual and audio segment is only a proxy of "information piece" that we truly want to extract. In future work will explore methods to address these issues and extract more fundamental units of information pieces.

#### 4.4.5.2 Grounding Limitation

From our observations, current cross-modal grounding algorithms do not work well on visual details that are specific to current video's context (*e.g.,* in a recipe video the host instructs

to mix the batter until it looks like "this", the model will label this as matched and cannot detect that the specific state of the batter is not described), and smaller or rarer visuals (*e.g.,* sprinkling salt). Cross-modal machine learning algorithms can also sometimes generate in consistent results due to its unexplainability (giving divergent matching scores to similar visuals that are close to each other).

#### 4.4.5.3 Video Production Style

CrossA11y works better on videos with relatively dense audio and visual information, and are partially inaccessible. For videos with a monotonous visual (*e.g.,* podcast video stream) or audio track (*e.g.,* only background music), our system will still correctly match visuals and audio, only to show that the entire track is not described. In such scenarios CrossA11y provides minimal information to authors.

## 4.5 Evaluations: Can CrossA11y Users Efficiently Identify Video Accessibility Issues?

We evaluated CrossA11y with 12 participants who have video creation experience to compare creating AD&CC with and without modality asymmetry visualizations. We want to investigate: *How does CrossA11y enable authors to efficiently identify and address visual and auditory accessibility issues in videos?*

### 4.5.1 Materials

We selected four videos on YouDescribe.com from different genres (Cheesecake recipe [40], handicraft tutorial [57], restaurant review [78] and day-in-the-life vlog [53]). All videos are under 5 minutes. We used CrossA11y to automatically identify inaccessible visual and audio segments and rendered the four videos on CrossA11y's interface. In addition, we

created *Interface 1* where we removed all accessibility visualizations to compare CrossA11y (Interface 2) with. In Interface 1, we provide a transcript-based timeline and display gaps-in-speech, following prior work [68, 107, 99]. With Interface 1, users create AD&CC and click "Add" button to add it to the current video timestamp.

### 4.5.2 Participants

We recruited 12 participants (7 female and 5 male) who all have previous video creation and sharing experience. Participants have created various types of videos including vlog, how-to, music, travel, presentation, product demo, etc. Participants were recruited through our university's internal communication channel and mailing lists. P8 did not complete the study due to technical issues. P5, P6, and P11 have their own YouTube/TikTok channels and have created around 40, 80, and 100 videos respectively. P3, P7, P10 and P12 created 10-20 videos. P1, P2 P4 and P9 less than 10 videos.

### 4.5.3 Procedure

We conducted a 90-minute study with each participant remotely. Each participant was paid $50 in gift card. In each session, we started by asking participants about their experience with videos and experience with accessibility. We asked if participants have ever added closed captions or audio descriptions to their videos, and their reasons for (not) creating AD&CC. We explained in details what AD&CC are for, and what they should describe or not describe based on AD&CC guidelines. We also went through two video examples with AD and CC to provide a more concrete understanding. Then, we demonstrated Interface 1 and 2 with example videos. Each participants tried out all features in both interfaces before continuing to the main study.

Each participant was asked to run four tests in total for Interface 1 and 2 with randomized order. For each interface, two videos are randomly selected without repetition. We provided

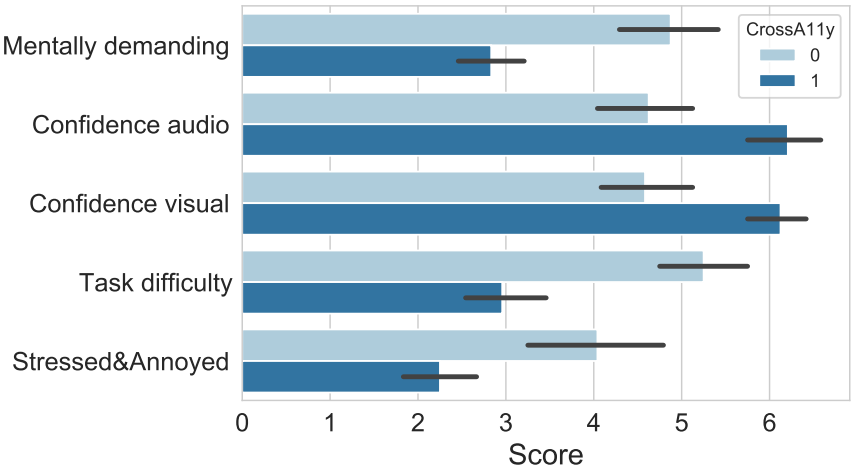



Figure 4.5: Participants' ratings to task load index questions (on a scale of 1-low to 7-high) for their experience adding AD/CC to videos with (1) or without (0) CrossA11y.

an open-ended prompt to participants, asking them to *"use this tool to make this video accessible to BVI and DHOH people"*. Participants are not subject to any time limit. At the end of each test, we asked participants to rate a set of questions on task load index [36]. After completing all four tasks, we asked participants to compare their experience of identifying and address AD and CC with and without CrossA11ythrough semi-structured interviews.

We recorded the audio track for the entire interview and the screen portion of trying out the interfaces. Both interfaces also automatically logged participants' use of different features, including their video navigation, clicks on vertical and horizontal bars, timestamps they chose to add AD/CC, and the content of AD/CC they wrote, etc. In total, we collected 4042 log instances of interaction data. We also recorded the completion time for each task for each participant.

### 4.5.4 Findings

Participants addressed 362 accessibility issues in total with Interface 1 (271 visual and 91 audio), and addressed 390 issues with Interface 2 (CrossA11y, 272 visual and 118 audio).

Participants on average spent 10 minutes and 12 seconds to complete a task. Participants unanimously preferred using Interface 2 (CrossA11y) over Interface 1. On a scale from 1-strongly disagree to 7-strongly agree, participants rated that CrossA11y is useful ($\mu = 6.0, \sigma = 0.58$), easy to use ($\mu = 6.33, \sigma = 0.47$) and would like to use it to make their videos accessible in future ($\mu = 6.50, \sigma = 0.65$).

#### 4.5.4.1 Interface Usage

With CrossA11y, most participants navigate the video using the horizontal timelines in the video pane (Figure 4.2C) or vertical side bars in the captions and video description pane (Figure 4.2E, F). P9 only used the original YouTube player to pause/play the video. She reported that she wanted to be really careful and make sure that she covered all problems. P1, P4, P6, P7 and P11 more frequently used the horizontal timelines to navigate. P4 reported that with horizontal timelines she can more easily understand her current position, and it provided a good overview. Other participants preferred to navigate in the captions and video description pane. P5 especially liked this "in-place" design where he can navigate, identify and edit all in one place. Participants in general navigated more using visual timeline and side bars than audio. And the frequencies of audio/visual navigation are approximately proportional to the number of audio/visual accessibility issues they addressed. While some participants mostly followed CrossA11y's guidance, 9 out of 11 participants used "Dismiss", "Add", and "Filter" to correct CrossA11y's predictions (further discussed in Section 4.5.4.4).

P1, P5, P8 and P10 particularly liked the use of color in CrossA11y. Colors allow them to get an overview of approximately how inaccessible this video is, quickly locate most critical accessibility issues, and monitor their work progress. P8 described:

> *"I like that when you have something undone, it will mark as red. This makes it really easy for me to locate on the tracks and navigate. You can take a glance at how much work is left. "* – P8

P10 also considered the segmentation of visual and audio to be especially useful. It allows him to add AD/CC to where the scene or sound occurs, and adds it for a coherent piece of information, comparing to Interface 1 that he needed to adjust himself.

#### 4.5.4.2 Efficiency and Performance

Participants on average spent 10 minutes 12 seconds to complete a task. There is no significant difference between task completion times using Interface 1 and 2. This could be caused by identifying more accessibility issues and thus spending more time overall to address the issues. For example P5 and P10 reported that Interface 1 required too much efforts and discouraged them from carefully inspecting the videos.

Thus, instead of comparing the overall time spent, which could be affected by video length, number of accessibility issues identified and efforts to write descriptions, we measure time per fix, i.e. total time divided by number of total AD/CC added. This metric represents how efficiently a participant can locate accessibility problems. A Wilcoxon signed-rank test shows that participants were able to create AD&CC more efficiently with CrossA11y ($\mu = 38.5, \sigma = 19.5$) than with Interface 1 ($\mu = 45.0, \sigma = 18.1$), with statistical significance ($W = 182.0, p = 0.037$). Participants found a variety of features CrossA11y provided to be helpful in making their workflow faster. This will be discussed in details in Section 4.5.4.3

Another important measurement for efficiency is how well authors were able to identify visual and auditory accessibility issues, specifically, how many of the captions and descriptions they added address an actual accessibility issue (precision), and how many of the total accessibility issues were addressed (recall). We collected participants' log data and computed the precision and recall scores for locating visual and auditory problems using both interfaces. We label a participant's created AD/CC to be correct if its added timestamp lies within the start and end time of a manually labeled accessibility issue.

Participants reached higher precision and recall for identifying inaccessible audio (pre-

cision: $\mu = 0.766, \sigma = 0.192$, recall: $\mu = 0.693, \sigma = 0.196$) and visual segments (precision: $\mu = 0.921, \sigma = 0.091$, recall: $\mu = 0.895, \sigma = 0.131$) with CrossA11y than with Interface 1. Wilcoxon signed-rank tests indicate statistical significance for the increase in participants' performance for both auditory (precision: $W = 123.0, p = 0.004$, recall: $W = 120.0, p = 0.004$) and visual (precision: $W = 0.0, p < 0.001$, recall: $W = 24.0, p < 0.001$) accessibility problems. This result aligns with participants responses to task load index questions. Participants felt significantly more confident in locating auditory ($W = 35.0, p = 0.003$) and visual ($W = 4.5, p < 0.001$) accessibility issues with CrossA11y than without. 5 participants reported that CrossA11y not only provides them with some guidance but also serves as a confirmation that improves their confidence.

#### 4.5.4.3 CrossA11y Workflows

Participants were able to identify and address accessibility problems with CrossA11y more efficiently. Results from task load index questions (Figure 4.5) also show that participants found using CrossA11y to be significantly less mentally demanding ($W = 9.5, p < 0.001$), less difficult ($W = 10.5, p < 0.001$), and less stressed or annoyed ($W = 11.5, p < 0.001$) than using Interface 1.

All participants reported that in Interface 1 they had to watch the entire video through and have to constantly check if there is an accessibility issue. P6 complained that she had to *"stop at every sentence"*. P9 had to *"pay attention all the time, every moment"*. Moreover, 8 out of 11 participants stated that it was a huge cognitive load for them to surface for visual and audio accessibility problems at the same time, because they have to repeatedly switch their minds between imagining "I cannot see the content" vs. "I cannot hear the content".

As a result, participants either had to play the video two times and only focus on addressing one type of accessibility issue at a time (P3, P6, P7, P9, P11), repeatedly inspect the same segment (P4), or rely on heuristics to make the process easier (P1, P2, P5):

*[With Interface 1] I first look for non-speech, or visual changes to some obvious object or some close-up shots. Then I would imagine that I can only access the video through one of my senses, for example my vision or my hearing, to determine that, ok, here might need a CC or AD.* – P2

CrossA11y enabled participants to more efficiently locate and address accessibility problems with lighter mental demand. With visualization of modality asymmetries, participants can get an overview of, immediately identify, and seamlessly navigate to surfaced visual and auditory accessibility issues. 8 out of 11 participants would directly jump to the highlighted red visual and audio segments in a video and address those problems, especially after they felt that the algorithm is accurate enough:

*"After the first one I felt like the algorithm is pretty accurate and sufficient. So in the second one, I would just click on the undone red marks. It's a much better experience."* – P3

P2, P6 and P10 employed a "dynamic workflow" in which they would skim through the gray segments and pay more attention to the red parts:

*"My workflow isn't linear anymore and I don't have to check for every second. For example, in this video I can click on a gray segment to instantly navigate to the position, and then realize that most of it is just the person speaking, then I can just skip the entire segment and go to the next one. For the first one I'll have to be continuously watching."* – P6

P2, P7 and P10 also explained that CrossA11y's highlight of inaccessible segments reduces the work from searching for all potential accessibility issues to judging if one modality of this video segment is inaccessible, which is much less mentally demanding:

*"Seeing the problem, I can understand it in hindsight. It is so much easier than I have to go over everything, paying attention to visual and audio, thinking if there is potentially an accessibility issue while the video is still playing."* – P7

9 out of 11 participants reported that with CrossA11y they were able to address both visual and auditory problems in parallel. P4 described that with the timelines she realized that most audio and visual issues are not at the same location. So when she was at a segment she can focus on either audio or visual. And even if they are around the same location, P4 explained, *"Since you know that there might be a problem, your attention will be on what potential problem does this segment have instead of which part has a problem. I don't have to distinguish. I feel like that was the hardest part."*

#### 4.5.4.4 Interpreting AI Predictions

9 out of 11 participants used "Dismiss", "Add", and "Filter" to correct CrossA11y's predictions. As discussed in previous sections, participants can easily judge and dismiss false positive predictions by our system. We observed that participants were able to identify visual accessibility issues with significantly higher precision (0.921), compared to the precision of CrossA11y's predictions (0.718). This indicates that participants were not overly relying on the system and able to determine whether the surfaced problem is actually inaccessible. P2, P5, P6, P9 and P10 also checked for false negative errors of CrossA11y, by skimming through gray segments or adjusting the slider (Figure 4.2H) to retrieve more accessibility problems:

*"After I have address all the issues, usually I just slide it to a bigger value and check if there's any red segments. If there is I'll click on those segments and see if they are actually accessible. If all the new problems are ok I'll stop there."* – P10

P1, P2 and P5 reported that they would prioritize workload over *complete* accessibility.

P2 stated that she will first look at the top-left corner to see how many issues remaining and just go with the default if not too many or not too few. P5 told us that having some description to cover some important visual stuff in his video is more important than completeness:

> *"Using this tool, I'm not trying to achieve a 100% accuracy. For any suggestions it provides it's already better. it increases my willingness to address them and at least try to make my videos more accessible. If I'm using the first one [Interface 1] I'll probably just choose to skip."* – P5

#### 4.5.4.5 Feedback & Improvement

During the study, participants suggested new features that could improve our interface. P1 and P4 thought that sometimes the segmented visual and audio content are repetitive, and they have to enter the same descriptions again and again. They suggested that we could cluster similar visuals/audio segments and allow authors to apply a description to all similar segments.

Although our system focuses on the identification of accessibility issues, as a number of prior work [82, 70, 68] have explored ways to author higher quality descriptions, 7 out of 11 participants hope that our system could automatically generate descriptions and captions, or at least some simple words to start with. They felt like this is the "last missing piece" of our system and would consider use it on all of their video creations.

P10 felt that our current design of the slider is a bit hard to understand, and we could potentially replace it with more well-defined levels. For instance, "You should fix this", "Recommend to fix", "Make your video completely accessible", etc.

## 4.6 Case Study: Usage of CrossA11y by Content Creators

We recruited 2 YouTubers who did not participate in the first study and conducted a 60-minute long study with each participant remotely. Each author was paid $50 in gift card. We demonstrated CrossA11y and asked the authors to make two of their own videos accessible using our system. We then conducted a semi-structured interview to discuss their experience, concerns and expectations. Specifically, we wanted to understand: *How could CrossA11y fit within content creators' video creation workflow, and help them make their own videos more accessible in the real-world settings?*

### 4.6.1 Participants

Author 1 mainly created life vlogs and music videos, with around 70 videos published in YouTube. Author 2 mainly created life vlogs and talking videos, with around 200 videos published in YouTube and TikTok. Only the second author has added closed captions to her videos before (not often). Neither of them has added audio descriptions to their videos.

### 4.6.2 Findings

#### 4.6.2.1 Integrating CrossA11y into Video Production

Both authors gave a 7 when rated on the usefulness of CrossA11y (from 1-strongly disagree to 7-strongly agree). The authors agreed that CrossA11y helps identifying inaccessible parts while doing it manually, even on their own videos, is hard for them. Both authors expressed enthusiasm and showed strong expectations towards integrating CrossA11y into existing video uploading process. Author 1 mentioned that she might consider how she designs the production style (*i.e.,* more description of the visual) of her video content based on accessibility feedback:

*"If it's integrated within my editor, I would try to edit the video in a way that's*

*more accessible. It would help me keep an eye out for parts are particularly inaccessible, or if I noticed that a lot of scenes are inaccessible, I might rethink how to structure the video better."* – Author 1

However, author 2 would still prioritize what she wants to express first when producing the videos, and would only consider accessibility until the editing is complete. She preferred to have a completely separate tool or website that she could upload the video and check for issues before she publish it onto YouTube.

#### 4.6.2.2 Balancing Efforts and Accessibility

Both authors gave a 6 when rated on " I would like to use CrossA11y to check and address for accessibility issues in future." (from 1-strongly disagree to 7-strongly agree) Their major concern that stops them from using this tool is balancing between time invested in fixing accessibility issues and how many audiences indeed benefit from it. Author 1 reported that she did no know how many BVI or DHOH people were watching her videos and required AD/CC, so she had not thought about making her videos more accessible. However if she knew that even a small part of her demographic needed it, she would definitely invest some time and use this tool to make her content accessible in future. Author 2 explained that she would try to make her video to be as accessible as possible with the time she had:

*"I always check my video through a color accessibility test, because it's something that takes a short amount of time, but allows my videos to be a little bit more accessible. So if your tool were to come out I will definitely do it because it's gonna take me less than 30 minutes for two videos while 45 minutes for one video manually [adding closed captions]. It saves me so much time."* – Author 2

## 4.7 Discussion and Future Work

Our work explores using cross-modal grounding to detect modality asymmetries in the visual and auditory tracks in a video, and instantiates the scores as a unified interface that allows users to efficiently identify and address video accessibility issues. Next, we describe the limitations of our system, discuss the implications, and envision future opportunities:

**Improved Segmentation of Information.** In CrossA11y , visual and audio tracks of media are first divided into semantically coherent pieces of information. In our implementation, we chose to use pauses in speech to segment the audio track and shot changes to segment the visual track. However, this proxy can be inaccurate sometimes. In the future, we hope to explore more semantically meaningful approaches to segment visual and audio information. For example, we will segment visuals into object-level segments such that each segment corresponds to one important visual object.

**Leveraging Information Importance.** In our current implementation, CrossA11y only detects unmatched segments and does not discern if an unmatched visual/audio segment contains important information. This results in the presenter (*e.g.*a host talking to the camera does not contain much useful information) and silence issues (*e.g.*silence or background noise does not to be explicitly described) that we have to address. Future system could estimate how important an unmatched visual/audio is (*e.g.*, based on the topic's uniqueness or consistency with respect to the rest of the video [68]), then surface and prioritize segments for authors based on importance in addition to accessibility.

**Modality Asymmetry for Accessibility Beyond Video.** We use cross-modal grounding to check for modality asymmetries in visual and audio. Future research can explore generalizing this pipeline to identify modality asymmetries in other modalities for other groups of people with disabilities. For example, identify asymmetries in textual and visual modalities for mobile app interfaces to locate accessibility problems for people with Dyslexia. Since most media is consists of information from three main modalities, visual, auditory and

textual, there might be opportunity to design a unified system that is able to provide an accessibility diagnosis for all major media contents with a consistent standard.

**Empathize and Motivate Accessibility.** The majority of participants from both evaluations claimed that the small number of BVI/DHOH audience, the lack of system assistance and video platform support are the reasons why they did not provide AD/CC for their created videos. Without assistance and support, they had to invest much effort when they were not sure if someone in their audience could benefit from the effort. In fact, on YouDescribe[7], a crowdsourcing audio description platform, blind and low vision people submit a large amount of requests everyday to make YouTube videos accessible. Providing more specific instructions or reminders on the video platform can help incentivize authors to add CC/AD. As author 2 commented, *"Your [video platforms or system] should educate authors about what is accessibility and why it's important."*

**Automated Generation as a Double-Edged Sword.** 7 of 11 participants in our lab study and both of the content creators mentioned they wanted automated AD/CC as a starting point. Although this research is focused on helping users identify accessibility problems rather than authoring AD/CC, we see an exciting opportunity to combine CrossA11y with existing systems like [82, 70, 68] to create an end-to-end experience for authors. Authors can efficiently identify accessibility problems first with our tool, and then generate automated results, view feedback, or optimize their description to fit within the video.

However, researchers should be cautious when providing AI-generated info as replacement of content authoring since it may decrease the content quality. Prior research [56] on alt text authoring showed that authors crafted significantly lower quality alt text when starting from the automatic alt text compared to starting from a blank box. We see one major opportunity for researchers to design the representation of AI-generated info to assist content authoring without priming authors with low quality automatic generations.

[7]https://youdescribe.org/

## 4.8 Conclusion

We present CrossA11y, a system that enables authors to efficiently identify and address visual and auditory accessibility issues in videos. Our system automatically estimates accessibility of visual and audio segments by checking for modality asymmetries using cross-modal grounding algorithms. It allows authors to quickly locate, review, script and preview AD&CC in a unified interface. Participants using CrossA11y in our user studies were able to author AD&CC more efficiently with lower mental demand. Content creators envisioned integrating our system into their video creation workflow, and expressed enthusiasm in using it to make their videos more accessible in future.

# CHAPTER 5

# Conclusion

The goal of this thesis is to make videos and video platforms more accessible with human-AI systems. To achieve this goal, I designed, developed and evaluated three human-AI systems that help different users (content creators, BVI audience) to improve video accessibility at different stages of video production (authoring, consuming).

The first project, presented in Chapter 2, is an ML-driven interface that leverages visual and auditory features of videos to automatically surface inherently accessible videos for BVI users. The second project, presented in Chapter 3, is a 4-week field deployment of a browser extension that displays automated video accessibility metrics for BVI people on YouTube. The third project, presented in Chapter 4, is a system that helps video authors efficiently detect and address accessibility issues in videos.

There are many opportunities for future work in this area. For example, future work could focus on developing systems that help content creators produce more accessible videos from the start (*e.g.*when recording videos), rather than systems that help content creators fix accessibility issues after the fact. Additionally, future work could focus on developing systems that provide more personalized recommendations to BVI users, *e.g.*by taking into account users' interests and preferences, as well as systems that provide more comprehensive accessibility evaluations of videos, *e.g.*by evaluating the accessibility of the comments section. Future systems could also utilize the large population of video audience without disabilities, to crowd source annotations and descriptions of videos online.

In conclusion, this thesis has presented three projects that aim to make videos and video

platforms more accessible with human-AI systems. By automatically analyzing visual and auditory characteristics of videos, and further assist people in identifying and addressing accessibility issues more efficiently, we can make a large population of videos online accessible to blind and visually impaired people. These projects have the potential to make a significant impact on the accessibility of video content online, and improve the experience of blind and visually impaired people when watching videos.

# REFERENCES


[1] 3playmedia.

[2] American council of the blind, audio description project, guidelines for audio describers. https://www.acb.org/adp/guidelines.html.

[3] Average speaking rate and words per minute.

[4] Gentle forced-aligner.

[5] These are the 10 most used smartphone apps.

[6] N. Reviers A. Remael and G. Vercauteren. Pictures painted in words: Adlab audio description guidelines. https://dcmp.org/learn/captioningkey/624.

[7] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)*, 52(6):1–37, 2019.

[8] Tania Acosta, Patricia Acosta-Vargas, Jose Zambrano-Miranda, and Sergio Lujan-Mora. Web accessibility evaluation of videos published on youtube by worldwide top-ranking universities. *IEEE Access*, 8:110994–111011, 2020.

[9] Patricia Acosta-Vargas, Luis Salvador-Ullauri, Janio Jadán-Guerrero, César Guevara, Sandra Sanchez-Gordon, Tania Calle-Jimenez, Patricio Lara-Alvarez, Ana Medina, and Isabel L Nunes. Accessibility assessment in mobile applications for android. In *International Conference on Applied Human Factors and Ergonomics*, pages 279–288. Springer, 2019.

[10] 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.

[11] Humam Alwassel, Dhruv Kumar Mahajan, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *ArXiv*, abs/1911.12667, 2020.

[12] Janarthanan Balakrishnan and Mark D Griffiths. Social media addiction: What is the role of content in youtube? *Journal of behavioral addictions*, 6(3):364–377, 2017.

[13] Shumeet Baluja, Rohan Seth, Dharshi Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, pages 895–904, 2008.

[14] Cynthia L Bennett, Jane E, Martez E Mott, Edward Cutrell, and Meredith Ringel Morris. How teens with visual impairments take, edit, and share photos on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[15] Jeffrey P Bigham. Increasing web accessibility by automatically judging alternative text quality. In *Proceedings of the 12th international conference on Intelligent user interfaces*, pages 349–352, 2007.

[16] Jeffrey P Bigham, Anna C Cavender, Jeremy T Brudvik, Jacob O Wobbrock, and Richard E Ladner. Webinsitu: a comparative analysis of blind and sighted browsing behavior. In *Proceedings of the 9th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 51–58, 2007.

[17] Jeffrey P Bigham, Irene Lin, and Saiph Savage. The effects of" not knowing what you don't know" on web accessibility for blind web users. In *Proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility*, pages 101–109, 2017.

[18] 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.

[19] 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)*, 2008.

[20] 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.

[21] Ronald Chenail. Youtube as a qualitative research asset: Reviewing user generated videos as learning resources. *Qualitative Report*, 16:229–235, 01 2011.

[22] Hsiu-Sen Chiang and Kuo-Lun Hsiao. Youtube stickiness: The needs, personal, and environmental perspective. *Internet Research*, 25:85–106, 02 2015.

[23] Amy Pavel Cole Gleason, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. Twitter a11y: A browser extension to make twitter images accessible.

[24] Ioana Croitoru, Simion-Vlad Bogolin, Yang Liu, Samuel Albanie, Marius Leordeanu, Hailin Jin, and Andrew Zisserman. Teachtext: Crossmodal generalized distillation for text-video retrieval. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11563–11573, 2021.

[25] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[26] Descript. Descript.com.

[27] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[28] Casually Explained. Casually explained: Reddit & casually explained.

[29] Robert Fildes and Paul Goodwin. Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, 37(6):570–576, 2007.

[30] L Gagnon, C Chapdelaine, D Byrns, S Foucher, M Héritier, and V Gupta. Computer-assisted system for videodescription scripting. In *Proceedings of Computer Vision Application for Visually-Impaired (CVAVI), a satellite workshop of CVPR*, 2010.

[31] 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.

[32] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based lstm and semantic consistency. *IEEE Transactions on Multimedia*, 19(9):2045–2055, 2017.

[33] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. "it's almost like they're trying to hide it": How user-provided image descriptions have failed to make twitter accessible. In *The World Wide Web Conference*, pages 549–559, 2019.

[34] Google/Insight Strategy Group. "what the world watched in a day" from premium is personal studies.

[35] Frank E Harrell Jr, Kerry L Lee, Robert M Califf, David B Pryor, and Robert A Rosati. Regression modelling strategies for improved prognostic prediction. *Statistics in medicine*, 3(2):143–152, 1984.

[36] Sandra G Hart. Nasa task load index (tlx). 1986.

[37] Marti Hearst. *Search user interfaces*. Cambridge university press, 2009.

[38] Morten Hertzum and Erik Frøkjær. Browsing and querying in online documentation: A study of user interfaces and the interaction process. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(2):136–161, 1996.

[39] Karen Holtzblatt and Hugh Beyer. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[40] Howdini. Easy cheesecake recipe: How to make cheesecake.

[41] The Smith-Kettlewell Eye Research Institute. Youdescribe.

[42] Jeep. 2020 jeep grand cherokee.

[43] Jingyang Jiang and Haitao Liu. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel english–chinese dependency treebank. *Language Sciences*, 50:93–104, 2015.

[44] Victoria Johansson. Lexical diversity and lexical density in speech and writing: a developmental perspective. *Lund Working Papers in Linguistics*, 53:61–79, 2009.

[45] M Laeeq Khan. Social media engagement: What motivates user participation and consumption on youtube? *Computers in Human Behavior*, 66:236–247, 2017.

[46] Jane E Klobas, Tanya J McGill, Sedigheh Moghavvemi, and Tanousha Paramanathan. Compulsive youtube usage: A comparison of use motivation and personality effects. *Computers in Human Behavior*, 87:129–139, 2018.

[47] 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*, pages 249–250, 2009.

[48] 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*, pages 163–170, 2010.

[49] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[50] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. Real-time captioning by groups of non-experts. In *UIST*, pages 23–34. ACM, 2012.

[51] The late show. Jon stewart climbs out from under colbert's desk to debut "irresistible" movie trailer.

[52] Apex Legends. Apex legends season 4: Assimilation gameplay trailer.

[53] Annie Liu. my first day of college (ucla) vlog!

[54] 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, 2021.

[55] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

[56] 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.

[57] The Shopping Mama. Easy halloween lantern craft for kids.

[58] Jennifer Mankoff, Holly Fait, and Tu Tran. Is your web page accessible? a comparative study of methods for assessing web page accessibility for the blind. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 41–50, 2005.

[59] MARKO. Custom fingerboards!! (giveaway).

[60] 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, 2020.

[61] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[62] George A Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[63] Lourdes Moreno, María González-García, Paloma Martínez, and Yolanda González. Checklist for accessible media player evaluation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 367–368, 2017.

[64] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12470–12481, 2021.

[65] Rosiana Natalie, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. Viscene: A collaborative authoring tool for scene descriptions in videos. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–4, 2020.

[66] RM Ochshorn and Max Hawkins. Gentle forced aligner. *github. com/lowerquality/gentle*, 2017.

[67] 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.

[68] 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, 2020.

[69] Yi-Hao Peng, Jeffrey P. Bigham, and Amy Pavel. Slidecho: Flexible non-visual exploration of presentation videos. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 2021.

[70] Yi-Hao Peng, JiWoon Jang, Jeffrey P. Bigham, and Amy Pavel. Say it all: Feedback for non-visual presentation accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (To Appear)*, 2021.

[71] Dude Perfect. Impossible ping pong trick shots.

[72] Paul Haridakis Ph.D and Gary Hanson M.A. Social interaction and co-viewing with youtube: Blending mass communication reception and social connection. *Journal of Broadcasting & Electronic Media*, 53(2):317–335, 2009.

[73] Peter Pirolli and Stuart Card. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–58, 1995.

[74] Audio Description Project. Master ad list.

[75] Audio Description Project. What is audio description?

[76] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[77] Rev. Convert audio & video to text.

[78] One Bite Pizza Reviews. Barstool pizza review - pelham pizza (pelham, ny).

[79] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[80] Murray Rowan, Peter Gregor, David Sloan, and Paul Booth. Evaluating web resources for disability access. In *Proceedings of the fourth international ACM conference on Assistive technologies*, pages 80–84, 2000.

[81] Andreas Sackl, Franziska Graf, Raimund Schatz, and Manfred Tscheligi. Ensuring accessibility: Individual video playback enhancements for low vision users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, pages 1–4, 2020.

[82] José Francisco Saray Villamizar, Benoît Encelle, Yannick Prié, and Pierre-Antoine Champin. An adaptive videos enrichment system based on decision trees for people with sensory disabilities. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, pages 1–4, 2011.

[83] Woosuk Seo and Hyunggu Jung. Understanding the community of blind or visually impaired vloggers on youtube. *Universal Access in the Information Society*, pages 1–14, 2020.

[84] Ben Shneiderman and Catherine Plaisant. Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies. In *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pages 1–7, 2006.

[85] Best Ever Food Review Show. Twisted cuban lechon in cuba!!! pork hammock!!

[86] The Infographics Show. How insane is el chapo's prison cell security?

[87] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, Rahul Sukthankar, and Cordelia Schmid. Learning video representations from textual web supervision. *ArXiv*, abs/2007.14937, 2020.

[88] YouTube Creator Studio. Search and discovery on youtube. https://creatoracademy.youtube.com/page/lesson/discovery#strategies-zippy-link-2. Accessed: 2021-04-02.

[89] Agnieszka Szarkowska. Text-to-speech audio description: towards wider availability of ad. *The Journal of Specialised Translation*, 15:142–162, 2011.

[90] Brennen Taylor. We tasted viral tiktok cooking life hacks.

[91] UFC. Ufc 246: Conor mcgregor octagon interview.

[92] Markel Vigo and Giorgio Brajnik. Automatic web accessibility metrics: Where we are and where we can go. *Interacting with computers*, 23(2):137–155, 2011.

[93] Markel Vigo, Justin Brown, and Vivienne Conway. Benchmarking web accessibility evaluation tools: measuring the harm of sole reliance on automated tests. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10, 2013.

[94] Markel Vigo and Simon Harper. Coping tactics employed by visually disabled users on the web. *International Journal of Human-Computer Studies*, 71(11):1013–1025, 2013.

[95] Markel Vigo, Barbara Leporini, and Fabio Paternò. Enriching web information scent for blind users. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*, pages 123–130, 2009.

[96] 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 & Social Computing*, CSCW '16, page 1584–1595, New York, NY, USA, 2016. Association for Computing Machinery.

[97] W3C. Audio description (prerecorded): Understanding sc 1.2.5.

[98] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5075–5084, 2021.

[99] 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, 2021.

[100] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. Camp: Cross-modal adaptive message passing for text-image retrieval. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5763–5772, 2019.

[101] JOHN A WATERWORTH and MARK H CIDGNELL. A model for information exploration.

[102] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. The youtube social network. 01 2012.

[103] Sidney G Winter. The satisficing principle in capability learning. *Strategic management journal*, 21(10-11):981–996, 2000.

[104] 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, 2017.

[105] YouTube. You know what's cool? a billion hours, 2017.

[106] YouTube. The latest youtube stats on when, where, and what people watch, 2018.

[107] 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, 2020.

[108] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. The impact of youtube recommendation system on video views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, page 404–410, New York, NY, USA, 2010. Association for Computing Machinery.

[109] Renjie Zhou, Samamon Khemmarat, Lixin Gao, Jian Wan, and Jilin Zhang. How youtube videos are discovered and its impact on video views. *Multimedia Tools and Applications*, 75(10):6035–6058, 2016.