What Makes Videos Accessible to Blind and Visually Impaired People?

Xingyu Liu
UCLA
Los Angeles, USA
xingyuliu@ucla.edu

Patrick Carrington
Carnegie Mellon University
Pittsburgh, USA
pcarrington@cmu.edu

Xiang 'Anthony' Chen
UCLA
Los Angeles, USA
xac@ucla.edu

Amy Pavel
Carnegie Mellon University
Pittsburgh, USA
apavel@cs.cmu.edu

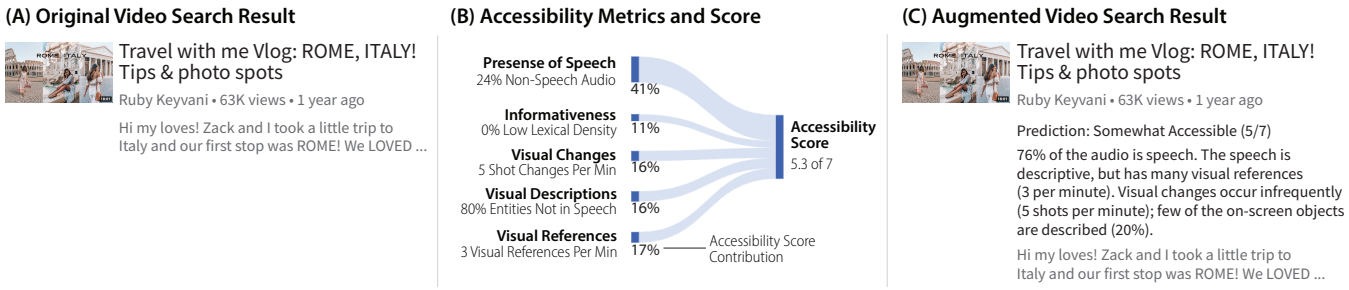


Figure 1: Video search results (A) contain no information about whether or not the video is accessible. People must use trial and error to find an accessible video. Our system calculates *video accessibility metrics* (B, left) informed by BVI formative study participants, and predicts the overall non-visual 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.

ABSTRACT

User-generated videos are an increasingly important source of information online, yet most online videos are inaccessible to blind and visually impaired (BVI) people. To find videos that are accessible, or understandable without additional description of the visual content, BVI people in our formative studies reported that they used a time-consuming trial-and-error approach: clicking on a video, watching a portion, leaving the video, and repeating the process. BVI people also reported video accessibility heuristics that characterize accessible and inaccessible videos. We instantiate 7 of the identified heuristics (2 audio-related, 2 video-related, and 3 audio-visual) as automated metrics to assess video accessibility. We collected a dataset of accessibility ratings of videos by BVI people and found that our automatic video accessibility metrics correlated with the accessibility ratings (Adjusted R^2 = 0.642). We augmented a video search interface with our video accessibility metrics and predictions. BVI people using our augmented video search interface selected an accessible video more efficiently than when using the original search interface. By integrating video accessibility metrics, video hosting platforms could help people surface accessible videos and encourage content creators to author more accessible products, improving video accessibility for all.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05.
https://doi.org/10.1145/3411764.3445233

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in accessibility**; *Accessibility systems and tools.*

KEYWORDS

blind, visual impairments, online videos, accessibility

ACM Reference Format:
Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445233

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 [65], the second most used mobile application [5], and reaches 81% of internet users under 25 [65], 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 [11]), 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 [64]. 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 indicate 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 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).

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 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 summary, 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 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.

2.1 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 [11]), and evaluating website adherence to these guidelines through manual or automated methods [35, 58]. Automatically assessing web accessibility could help non-expert developers identify and fix accessibility problems [49], or help web users surface sites that might be accessible [60]; in practice such automated accessibility evaluation results can be challenging for developers to interpret [49], lack sufficient coverage of accessibility issues [59], and inadequately represent user's perceptions of accessibility [60]. 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 [59] and WCAG guidelines provide only high-level guidance [11]. 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 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" [47] – associated with the video (much as images accessibility is based on the presence of alternative text [11]). For instance, the Section 508 Rehabilitation Act, the Web Content Accessibility Guidelines (WCAG 2.1) [11], 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 [41]. But, unlike traditional media (TV and movies) where professionally produced audio descriptions are becoming increasingly common on streaming services [46], 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 [1, 10], feedback on the content at production-time [43], feedback on audio descriptions [40, 51], and hosting descriptions [24]. 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, 20], but such methods still create inaccurate and unspecific captions in comparison to humans. With few exceptions [48], 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 [63]). Other research instead used computational assistance to help authors write audio descriptions more efficiently by: using computer vision to detect key visual content [18, 19], deep learning to provide a computer-drafted summary [19, 66], synthesized voice to convert text to speech [18, 29, 30, 55], and automatic editing to fit human-authored descriptions into the space provided [42]. 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 [50], and broadly making media players more accessible [39].

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.

2.3 Searching and browsing online videos

Prior work studies how the general population searches, browses, and watches videos online [12, 13, 28, 45, 67]. 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) [22]. Other work also suggests information seeking, entertainment, along with social connectedness, as key reasons for watching videos online [12, 28, 45, 62]. 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 [14, 21], Facebook [61], YouTube creators [52], and Snapchat [9]). 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.

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.

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), tunnel 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

	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 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).

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

3.2 Results

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 [22, 28]). 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.

[1]Transcribed using rev.com

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 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).

3.2.3 What video moments were inaccessible? During the co-watching exercise, participants requested additional information when the video segments lacked speech, failed to describe on-screen subtitles, and referenced visual content in speech, confirming interview

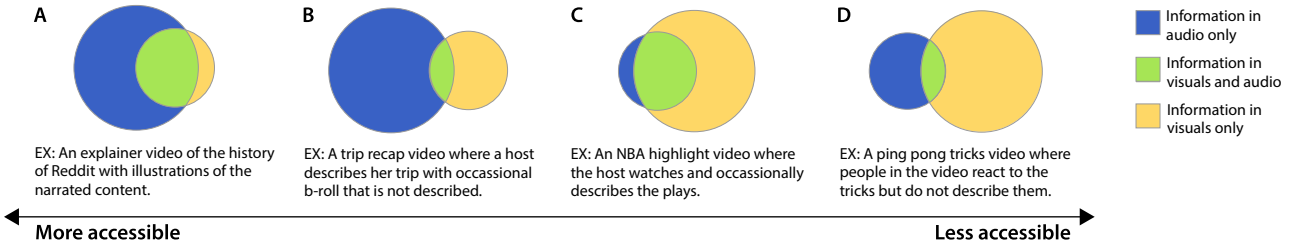


Figure 2: The accessibility of a video depends on what information is in the audio track (blue, left), what information is in the video track (yellow, right), and what information is redundant between the audio and visual channels (green, middle). In (A) more information is contained in the audio than 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.

results (Table 1, #M). Further, the in-situ video responses revealed an additional property of inaccessible videos:

Undescribed sounds. Participants frequently asked for more information due to undescribed sounds including environmental noise (e.g., hearing a crowd, P4 asked the location of the video), reactions (e.g., hearing cheering, P12 asked the reason for cheering), sound changes (P10 inferred a scene change due to the change in the background sound, but didn't know what the scene changed from or to), and potential metaphors (e.g., P8 asked if an explosion noise indicated a literal or metaphorical explosion).

3.2.4 How do participants find accessible videos? Participants currently lack support to help them find accessible videos amidst a large quantity of inaccessible videos. They relied on the following suboptimal strategies:

Trial and error. 9 participants mentioned watching a portion of a video and then clicking off if the video was inaccessible, trying multiple videos to find an accessible video. When previewing inaccessible videos, most times:

> "I can tell right away. I open it and there's no dialogue and I can't really tell what it is. But sometimes I have to watch it for a while before I realize, okay, I can't follow what's happening here." – P7

Topic and length. within videos of a single type or topic. One participant tried using length to predict which crocheting videos would be well-explained:

> "If it's top 10 things to crochet for Christmas or something and the video is only a minute long, then I know that video won't have visual description." – P12

Curated feeds. When not searching for a new topic or domain, participants relied on their curated feeds (e.g., the YouTube homepage) and subscriptions to increase the likelihood that the recommended videos would be accessible.

Finding extra information. Participants occasionally tried to find more information about specific videos (assigned, recommended, or of particular interest) via external description services like AIRA and YouDescribe (9 participants), time-stamped comments or secondary online sources (10 participants), or by asking friends or family (8 participants).

3.2.5 What are the consequences of inaccessible videos? Although participants cited leaving the video as the primary consequence of finding inaccessible videos, 5 participants mentioned a loss of trust in the visual content as a consequence (e.g., P3 found trusting visual content important when selecting videos for their children to watch), and 4 mentioned social disconnection as a consequence (e.g., P4 missed out on topical graphs, and P5 missed joke references to online content).

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 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 2B vs. 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 2A vs. 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).

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

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'

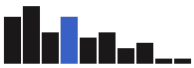
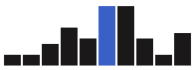
Heuristics (H) and Metrics (M)	Distribution	Median	Examples
Audio			
H1: Presence of speech **M1:** % Non-speech		16%	*Accessible*: 0%, explainer video on the history of Reddit [16]. *Inaccessible*: 73.6%, demonstrations of Ping Pong trick shots [44].
H2: Informative language **M2:** % Low lexical density speech		0%	*Accessible segment*: 0.80, a narrator reads a scripted description [53]. "...and **taro root**. **Next, boiled cassava**, **known locally** as **Yuca**." *Inaccessible segment*: 0.27, two people talk about their food as they eat [53]. "Oh this is **awesome**. Oh that is **pretty crispy** it's like..."
Visual			
H3: Infrequent visual changes **M3:** Rate of shot changes		17.43	*Accessible*: 3.8 shots/min, interview video after a mixed martial arts match [57]. *Inaccessible*: 51.6 shots/min, video game trailer [33].
H4: Simple visual content **M4:** # Visual entities / min		10.87	*Accessible*: 2.6 entities/min, late night talk show [32]. [audience, crowd, people, performance, reality television] *Inaccessible*: 63.9 entities/min, car advertisement [25]. [mountain, car, climbing, dust, dirt road,...] and 45 more.
Audio-visual			
H5: Description of visual objects **M5:** % Visual entities not in speech		79%	*Accessible*: 60% visual entities not in speech, TikTok food hack reaction video [56]. [black, cake, **chocolate**, kitchen]; "...putting **chocolate** on a saltine cracker..." *Inaccessible*: 94% visual entities not in speech, car advertisement [25]. [mountain, car, dust, dirt road,...]+45 more; "...to see things from a new perspective..."
H6: Description of on-screen text **M6:** # Undescribed on-screen text / min		5.16	*Accessible*: 0 instances/min, customizing fingerboards [36]. No on-screen text. *Inaccessible*: 4.18 instances/min, video game trailer [33]. Release date, producer, platforms, etc. *APEX LEGENDS SEASON 04 COMING FEBRUARY 4!*
H7: Few visual references **M7:** # Visual references / min		2.46	*Accessible*: 0 instances/min, story of El Chapo with animated illustrations [54]. "...Francisco 'El Chito' Camberos Rivera opened the electronic door..." *Inaccessible*: 13.8 instances/min, TikTok food hack reaction video [56]. "...we're gonna pour **this** all on **here** to melt..."

Table 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.

relationship to accessibility consistently negative). For example, a Reddit explainer video [16] which explains how Reddit works in details and keeps on talking (0% non-speech) would be more accessible than ping pong trick shots video [44] 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). We use lexical density [27], 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).

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 [26]) of the video (shifted by 0.5s offsets). We calculate the lexical density using NLTK Part of Speech Tagger [34] 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 [27]) over the over the total speech time to find the percentage of low lexical density speech.

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 [57], 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 [33], 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 [32] with a simple setup would have less visual content that needs to be described, and a car advertisement [25] may 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.

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 [56] would be more accessible than a car advertisement in which the speech is just motivating quotes that are completely irrelevant to the visual [25]. 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 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 [37, 38] 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 [33], 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 ±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

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

[2]https://github.com/Breakthrough/PySceneDetect
[3]https://cloud.google.com/video-intelligence
[4]https://cloud.google.com/video-intelligence/
[5]https://demo.allennlp.org/coreference-resolution

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.

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). 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 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.

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' 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 3, 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.

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.	

Table 3: Questions used to collect BVI users' perceived video accessibility ratings. Median values are colored blue. For Q1 the accessibility scale was: 1-very inaccessible, 2-inaccessible, 3-somewhat inaccessible, 4-neutral, 5-somewhat accessible, 6-accessible, and 7-very inaccessible. For Q3-Q6 the Likert-scale questions were framed as statements with agreement from 1-strongly disagree to 7-strongly agree.

Per-video accessibility ratings: To obtain the per-video accessibility ratings, we averaged participant ratings (Table 3). Overall, participants achieved moderate to substantial agreement [31] 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) ≥ 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

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

	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:		$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$	

Table 4: Regression models of how 7 accessibility metrics contribute to BVI users' perceived accessibility scores. The initial model shows the result including all metrics, and the reduced model shows result after removing insignificant metrics. Both models have statistically significant correlations to users' accessibility ratings, and 5 out of 7 our proposed metrics are significant independent variables.

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).

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), 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.

Model Assumptions: First, we want to confirm that our data and model satisfy the assumptions of linear regression: (1) No Multicollinearity — we computed the correlation matrix of our independent variables and all magnitudes of correlation coefficients are less than 0.4 (✓); (2) Homoscedasticity — we plotted a scatter plot of residuals versus predicted values of our model, residuals were centered around 0, and no clear pattern was found. The variance of error terms are similar across different values (✓); (3) Multivariate Normality — we plotted the normal Q-Q plot of residuals and the points show high linearity. The residuals are normally distributed (✓) (All Figures attached in Appendix). Thus, all assumptions are satisfied and our linear regression model is suitable.

Results: For the first research question **R1**, we are confident that there is a statistically 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 4). 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 4). 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%.

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 ± 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 counterbalanced). 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.

		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

Table 5: Summary of task time to select a video and number of videos clicked on and previewed (trial-and-error) during the 8 user study sessions for original, metrics, and metrics+filter video search interfaces. Video search tasks are: (A) paper plane, (B) trip to Italy, and (C) robot dog. Video search interfaces are: (1) original, (2) metrics, and (3) metrics+filter.

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:

> "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 5). 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 [15, 17]. 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.

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 [18, 19, 29, 40, 42, 66]. Whereas prior work already surfaced silent video segments for audio description [40, 42, 66], 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 [43] for slides) to their own videos before publishing.

Perceived Accessibility vs. True Accessibility: Accessibility ratings from BVI people 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 [23]). 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 straight-forward 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.

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

REFERENCES

[1] [n.d.]. 3PlayMedia. https://www.3playmedia.com/
[2] [n.d.]. American Council of the Blind, Audio Description Project, Guidelines for Audio Describers. https://www.acb.org/adp/guidelines.html.
[3] [n.d.]. Average Speaking Rate and Words per Minute. http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html
[4] [n.d.]. Gentle Forced-aligner. https://lowerquality.com/gentle/
[5] [n.d.]. These are the 10 most used smartphone apps. https://www.businessinsider.com/most-used-smartphone-apps-2017-8
[6] N. Reviers A. Remael and G. Vercauteren. [n.d.]. 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. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–37.
[8] Tania Acosta, Patricia Acosta-Vargas, Jose Zambrano-Miranda, and Sergio Lujan-Mora. 2020. Web Accessibility Evaluation of Videos Published on YouTube by Worldwide Top-Ranking Universities. *IEEE Access* 8 (2020), 110994–111011.
[9] Cynthia L Bennett, Jane E, Martez E Mott, Edward Cutrell, and Meredith Ringel Morris. 2018. How teens with visual impairments take, edit, and share photos on social media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
[10] Carmen J Branje and Deborah I Fels. 2012. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness* 106, 3 (2012), 154–165.
[11] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* (2008).
[12] Ronald Chenail. 2011. Youtube as a qualitative research asset: Reviewing user generated videos as learning resources. *Qualitative Report* 16 (01 2011), 229–235.

[13] Hsiu-Sen Chiang and Kuo-Lun Hsiao. 2015. YouTube stickiness: The needs, personal, and environmental perspective. *Internet Research* 25 (02 2015), 85–106. https://doi.org/10.1108/IntR-11-2013-0236

[14] Amy Pavel Cole Gleason, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. [n.d.]. Twitter A11y: A Browser Extension to Make Twitter Images Accessible. ([n. d.]).

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

[16] Casually Explained. [n.d.]. Casually Explained: Reddit & Casually Explained. https://youtu.be/Uy9V_v-XV8Q

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

[18] L Gagnon, C Chapdelaine, D Byrns, S Foucher, M Héritier, and V Gupta. 2010. Computer-Assisted System for Videodescription Scripting. In *Proceedings of Computer Vision Application for Visually-Impaired (CVAVI), a satellite workshop of CVPR*.

[19] Langis Gagnon, Samuel Foucher, Maguelonne Heritier, Marc Lalonde, David Byrns, Claude Chapdelaine, James Turner, Suzanne Mathieu, Denis Laurendeau, Nath Tan Nguyen, et al. 2009. 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 (2009), 199–218.

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

[21] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*. 549–559.

[22] Google/Insight Strategy Group. [n.d.]. "What the world watched in a day" from Premium is Personal studies. https://www.thinkwithgoogle.com/feature/youtube-video-data-watching-habits/

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

[24] The Smith-Kettlewell Eye Research Institute. [n.d.]. YouDescribe. https://youdescribe.org/

[25] Jeep. [n.d.]. 2020 Jeep Grand Cherokee. https://youtu.be/oXzvyoFkWwY

[26] Jingyang Jiang and Haitao Liu. 2015. The effects of sentence length on dependency distance, dependency direction and the implications–based on a parallel English–Chinese dependency treebank. *Language Sciences* 50 (2015), 93–104.

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

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

[29] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. 2009. Providing synthesized audio description for online videos. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 249–250.

[30] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. 2010. Are synthesized video descriptions acceptable?. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility*. 163–170.

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

[32] The late show. [n.d.]. Jon Stewart Climbs Out From Under Colbert's Desk To Debut "Irresistible" Movie Trailer. https://youtu.be/0WSxCOTEyQA

[33] Apex Legends. [n.d.]. Apex Legends Season 4: Assimilation Gameplay Trailer. https://youtu.be/DFY_scgPl80

[34] Edward Loper and Steven Bird. 2002. 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*.

[35] Jennifer Mankoff, Holly Fait, and Tu Tran. 2005. 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*. 41–50.

[36] MARKO. [n.d.]. CUSTOM FINGERBOARDS!! (GiVEAWAY). https://youtu.be/Hv7hqBFnLmA

[37] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

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

[39] Lourdes Moreno, María González-García, Paloma Martínez, and Yolanda González. 2017. Checklist for Accessible Media Player Evaluation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 367–368.

[40] Rosiana Natalie, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

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

[42] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 747–759.

[43] Yi-Hao Peng, JiWoon Jang, Jeffrey P. Bigham, and Amy Pavel. [n.d.]. Say It All: Feedback for Non-visual Presentation Accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (To Appear)*.

[44] Dude Perfect. [n.d.]. Impossible Ping Pong Trick Shots. https://youtu.be/0ADQauuOJto

[45] Paul Haridakis Ph.D and Gary Hanson M.A. 2009. Social Interaction and Co-Viewing With YouTube: Blending Mass Communication Reception and Social Connection. *Journal of Broadcasting & Electronic Media* 53, 2 (2009), 317–335. https://doi.org/10.1080/08838150902908270

[46] Audio Description Project. [n.d.]. Master AD List. https://acb.org/adp/masterad.html

[47] Audio Description Project. [n.d.]. What is Audio Description? https://acb.org/adp/ad.html

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

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

[50] Andreas Sackl, Franziska Graf, Raimund Schatz, and Manfred Tscheligi. 2020. Ensuring Accessibility: Individual Video Playback Enhancements for Low Vision Users. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4.

[51] José Francisco Saray Villamizar, Benoît Encelle, Yannick Prié, and Pierre-Antoine Champin. 2011. 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*. 1–4.

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

[53] Best Ever Food Review Show. [n.d.]. TWISTED Cuban LECHON in Cuba!!! Pork Hammock!! https://www.youtube.com/watch?v=SxrjdTfR6cI

[54] The Infographics Show. [n.d.]. How Insane is El Chapo's Prison Cell Security? https://youtu.be/XvVCc1Ts0MA

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

[56] Brennen Taylor. [n.d.]. We TASTED Viral TikTok Cooking Life Hacks. https://youtu.be/Sq021CY8_Mg

[57] UFC. [n.d.]. UFC 246: Conor McGregor Octagon Interview. https://youtu.be/6V0UxqD57WI

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

[59] Markel Vigo, Justin Brown, and Vivienne Conway. 2013. 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*. 1–10.

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

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

[62] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. 2012. The YouTube Social Network. (01 2012).

[63] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1180–1192.

[64] YouTube. 2017. You know what's cool? A billion hours. https://youtube.googleblog.com/2017/02/you-know-whats-cool-billion-hours.html

[65] YouTube. 2018. The latest YouTube stats on when, where, and what people watch. https://www.thinkwithgoogle.com/data-collections/youtube-stats-video-consumption-trends/

[66] 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.

2020. 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*. 47–60.

[67] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. 2010. The Impact of YouTube Recommendation System on Video Views. In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement* (Melbourne, Australia) *(IMC '10)*. Association for Computing Machinery, New York, NY, USA, 404–410. https://doi.org/10.1145/1879141.1879193

9 APPENDIX A

9.1 Normality Checks for Linear Regression

9.2 Questions used in survey

Q1: Rate the accessibility of this video when considering the video as is (1 - very inaccessible to 7 - very accessible).

Q2: Provide reasons for your accessibility rating of this video.

Q3: I feel that much of the important information in this video was conveyed via the audio track (1 - strongly disagree to 7 - strongly agree).

Q4: I feel that much of the important information in this video was conveyed via the visual content (1 - strongly disagree to 7 - strongly agree).

Q5: I feel that the audio track described much of the important visual content (1 - strongly disagree to 7 - strongly agree).

Q6: I feel that much of the audio track was confusing or hard to understand without seeing the visual content (1 - strongly disagree to 7 - strongly agree).

Q7: Provide a 3 to 5 sentence summary of this video.

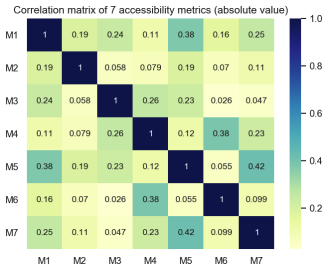


Figure 3: Correlation matrix of independent variables (accessibility metrics). All magnitudes of correlation coefficients are less than 0.4, satisfying the No Multicollinearity assumption of linear regression.

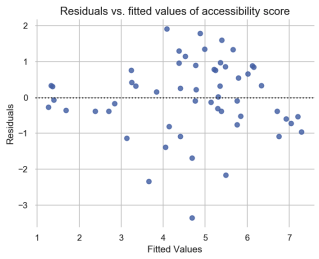


Figure 4: Residual vs. fitted plot of our linear regression model. Residuals were centered around 0, and no clear pattern was found. The variance of error terms are similar across different values. The model satisfies the Homoscedasticity assumption of linear regression.

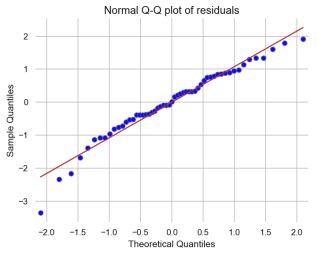


Figure 5: Q-Q plot of our linear regression model. Points show high linearity along the line. The residuals are normally distributed, satisfying the Multivariate Normality assumption of linear regression.