User-Driven Automated Audio Description to Enhance Video Accessibility for Blind and Low Vision Users

by

Maryam Saadat Cheema

A Thesis
Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved October 2024 by the
Graduate Supervisory Committee

Hasti Seifi, Chair
Pooyan Fazli
Sri Kurniawan

ARIZONA STATE UNIVERSITY
December 2024

ABSTRACT

Audio descriptions (AD) make videos accessible for blind and low vision (BLV) users by describing visual elements that cannot be understood from the main audio track. AD created by professionals or novice describers is time-consuming and lacks scalability while offering little control to BLV viewers on description length and content and when they receive it. To address this gap, this work explores user-driven AI-generated descriptions, where the BLV viewer controls when they receive descriptions. In a study, 20 BLV participants activated audio descriptions for seven different video genres with two levels of detail: concise and detailed. Results show differences in AD frequency and level of detail BLV users wanted for different videos, their sense of control with this style of AD delivery, its limitations, and variations among BLV users in their AD needs and perception of AI-generated descriptions. The implications of these findings for future AI-based AD tools are discussed.

ACKNOWLEDGMENTS

I would like to express my gratitude to my professor and committee chair, Dr. Hasti Seifi, whose insightful feedback and guidance have greatly enhanced this work. I am also deeply appreciative of my coauthors, Dr. Hasti Seifi and Dr. Pooyan Fazli, for their invaluable contributions and support throughout this project.

Furthermore, I extend my sincere thanks to my defense committee for their generous sharing of knowledge and expertise.

I am also thankful to my lab members, whose feedback and moral support were a constant source of encouragement. Lastly, I am grateful to the study participants, whose contributions not only enriched this work but also inspired me throughout the process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

The rapid growth and popularity of short videos on online video-sharing platforms have exacerbated the accessibility gap for blind and low vision (BLV) users. With billions of users using platforms such as YouTube, TikTok, and Instagram, videos make up 82% of all internet traffic [63, 67]. As a result, video content has become more diverse, with various types of user-generated content and varying video and audio quality. To make videos accessible, audio descriptions (AD) are recorded by professional describers to narrate the key visual elements of a video, such as actions, characters, scene changes, on-screen text, and other visual content [60]. While community platforms, such as YouDescribe [71], allow sighted describers to volunteer and describe video content, authoring pre-recorded ADs for the ever-growing number of online videos is simply not an option anymore.

Recent advances in artificial intelligence (AI) models, particularly multimodal large language models (MLLMs), offer new automated methods for generating descriptions [64, 49, 66, 9, 61, 11, 15, 48]. However, further research is needed to understand BLV users' preferences and requirements for AI-generated ADs across different types of videos. MLLMs can analyze one or more frames from a video and generate a textual description of the visual content [36, 51, 8, 17]. Despite this, the effectiveness of AI-generated descriptions in directly serving BLV users remains largely unknown. The timing of ADs poses another challenge for automated systems. While these systems can detect scene changes, they struggle with short videos, particularly

those that are single-shot or feature rapid scene transitions [18].

While some guidelines for AD timing exist (e.g., *"Avoid describing over audio that is essential to comprehension."* [19]), professional describers often rely on their tacit knowledge and the video content to time ADs in a way that supports BLV users' comprehension and enjoyment. Little data is available on how frequently ADs are needed for different types of videos. In addition, recent qualitative studies suggest that BLV users have diverse AD needs depending on video content (e.g., entertainment vs. educational videos) and personal preferences [31, 66]. Still, human and automated systems generally produce fixed descriptions at set times, which might not account for a BLV viewer's preference. Therefore, more research is needed to inform the design of AI-based AD platforms.

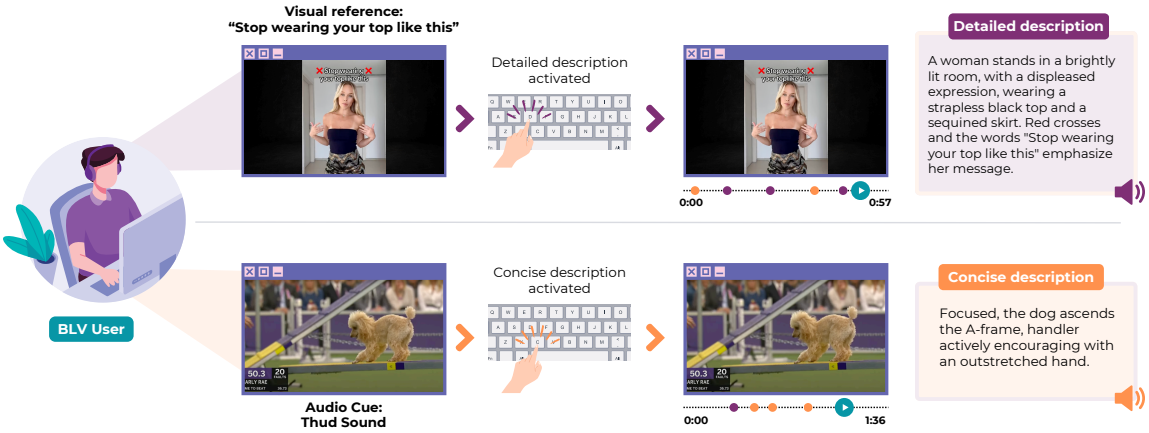


Figure 1.1: User-driven descriptions for different video genres. When watching a video, blind and low vision users can press the C or D keys to activate concise and detailed descriptions generated by a multimodal large language model. In the user study, participants activated descriptions based on different use cases such as visual references or audio cues in the main audio track.

To address these gaps, we investigate *user-driven AI-generated descriptions* to gather BLV users' AD needs for different types of videos and analyze their perceptions of watching videos using this approach. User-driven ADs allow BLV viewers to

activate a description at any point in the video based on their preferences and audio cues. AI-generated descriptions can be verbose and inefficient for video watching [57]. On the other hand, detailed descriptions may improve BLV users' understanding of visual content. Thus, we examine two levels of detail for ADs: a shorter, concise description vs. a more detailed description. Specifically, we aim to address the following questions: (1) How do BLV users' preferences for AD timing and detail differ between different video genres? (2) What are BLV individuals' perceptions and experiences with user-driven AI-generated ADs?

To answer these questions, we developed a prototype for activating ADs and conducted a study with 20 BLV users across different video genres. We prompted an MLLM, specifically GPT-4 Vision (GPT-4V) [51], with AD guidelines from professional describers to generate descriptions for seven short online videos spanning various categories, such as film and animation, education, and cooking.

BLV users could press a key to activate either a concise or detailed description at any time. They interacted with the videos to activate ADs, rated the effectiveness, efficiency, and enjoyment of the ADs, and shared their experience with user-activated AI descriptions.

Our quantitative results highlight significant differences in the frequency of descriptions needed for different video genres, with shorter AD intervals required for *Film and Animation*, and longer intervals for *Education, Health and Fitness*, and *Beauty* videos. Moreover, BLV individuals differed in their frequency and type of AD requests. Our thematic analysis of the interviews identified three key themes: (1) BLV users' increased sense of control and active watching experience when activating ADs, accompanied by a higher cognitive workload, (2) preferences for pre-recorded vs. user-driven ADs depending on the video content, viewing context, and individual BLV users, and (3) the positive and negative aspects of AI-generated descriptions.

Drawing on these results, we discuss implications for future AI-based AD platforms and the evolving roles of BLV users and sighted describers. Our contributions include:

- Insights into BLV individuals' perceptions of user-driven AI descriptions when watching short online videos
- Empirical data on the frequency and type of AD requests for seven video genres as well as variations among BLV users

Chapter 2

RELATED WORK

We review prior work on video accessibility practices and needs and interactive tools for the creation and use of ADs.

Audio Description Practices and Needs

Although AD has been around for over three decades and produced for high-budget films and movies [60, 33], the field has gained momentum in the past decade due to the exponential amount of online videos [40]. The WCAG 2.0 Level AA compliance mandates that AD be provided for all prerecorded video content in synchronized media [69, 6]. This push for video accessibility has resulted in AD authoring guidelines [1, 19, 46, 50, 7], such as Netflix, which focuses on entertainment content, and DCMP, focusing on educational/instructional content. These guidelines, initially designed for professional describers, have been used to train novice describers on community-driven platforms [71].

Yet, these guidelines and practices primarily focus on pre-recorded ADs. Furthermore, UK Ofcom guidelines state that "some programmes are too fast-moving, or offer little opportunity to insert AD" [56], indicating a gap in how to describe other forms of video content.

Recent research suggested that BLV people wish to interact with video content in new ways besides pre-recorded ADs. Specifically, Bodi et al. [9] investigated the viability of providing video accessibility via interactive visual question answering and showed that BLV users requested descriptions more frequently than asking questions. Chang et al. investigated AD for 360° videos and found BLV participants preferred

immersive ADs over standard ADs for this format [13]. Similarly, SPICA system [49] enabled BLV users to interactively explore video content. Participants found the object exploration feature and object-specific sound effects enhanced overall video consumption. These works underscore the importance of moving beyond traditional ADs to offer BLV users a more customised experience. Similarly, our work investigates user-driven descriptions as a way for BLV users to access video content.

Others studied variations in BLV users' AD needs depending on their visual impairment and the video content. Chmiel and Mazur [14] examined AD preferences between congenitally blind, non-congenitally blind and low vision participants and found some specific preferences related to character naming and the use of metaphors in AD. Another study highlighted how the amount of information in AD can impact the experience of BLV users [21], where segmented step-wise descriptions were less cognitively demanding and led to better recall for BLV users. A recent interview study by Jiang et al. suggested that BLV users have different goals and preferences when watching different video genres [31]. For example, BLV users wanted detailed descriptions of people and their appearance in short-form videos of family and friends but not in educational videos. Relatedly, Natalie et al. found that customized ADs can improve BLV users' video understanding, immersion, and information navigation efficiency [43]. We build on these studies to collect quantitative data on the frequency and amount of detail needed for seven video genres across users with various visual impairments.

Interactive Systems for Audio Description

To assist sighted users in creating descriptions, several tools have been built to streamline different parts of the process [10, 54, 39, 34, 72, 70, 27, 37, 73, 32, 45, 20]. LiveDescribe was one of the first tools that investigated the potential of using vol-

unteers to create audio descriptions by automating gaps to fit the audio description [10]. Rescribe helped describers optimize description placement by dynamic programming [54], whereas CrossA11y assisted authors in detecting visual and auditory accessibility issues in a video [39].

Yuksel et al. developed a system that generated baseline descriptions which could then be revised by sighted individuals to produce high-quality descriptions [72]. Other tools have included BLV users in AD and video creation process. Viscene studied the efficacy of collaboration between novice describers and BLV reviewers to create high-quality descriptions [44]. This work informed an automatic feedback tool to support novice users in authoring ADs [45]. Similarly, AccessibleAD expanded access to AD writing to make BLV users become an active part of AD creation [32]. These approaches still require manual authorship, which even with technology support, remains a challenging task for describers.

Recent advances in artificial intelligence (AI) models have fueled research in automated AD authorship. Wang et al. built a system that analyzed the audio and visual content of the video to create automated descriptions using deep learning [66]. Strangl et al. developed a hybrid approach that generated preset descriptions and provided additional information based on user queries [61]. Automated AD approaches have been dominated by deep learning, which focused on content extraction (encoding) and text generation (decoding) [2]. Due to the lack of context awareness, the automated descriptions can be rather verbose and or lack detail appropriate to the content of the video. To cater to this, recent systems have used large language models (LLMs) to create descriptions [64, 49]. ShortScribe used GPT-4 to create long descriptions, short descriptions, and shot-by-shot descriptions for short-form videos. This is the first of its works that has provided different types of AI descriptions for videos. Similarly, SPICA used GPT-4 to create natural language object descriptions

in an interactive video content exploration system [49]. These systems used vision models to extract information from individual frames and then created descriptions using LLMs. This two-step approach can disconnect the semantic meaning between the frames and the generated descriptions [35]. Our work builds on this prior research by using an MLLM to directly generate descriptions from a sequence of frames and preserve the spatiotemporal context of the videos.

Chapter 3

STUDY MATERIALS AND INTERFACE

To study user-driven AI descriptions, we selected short videos from different genres, created concise and detailed descriptions for them using GPT-4V, and developed a user interface for watching videos with user-driven ADs.

Video Selection

We chose short videos (50-110 seconds) from seven different genres to evaluate user-driven AI descriptions across a wide range of video content. The chosen genres include: Food and Cooking, Beauty, Pets and Animals, People and Blogs, Health and Fitness, Film and Animation, and Education (Figure 3.1). We selected these genres for several reasons. First, prior literature suggested that BLV users' needs for ADs can vary depending on the video genre [31]. Second, the genres covered a range of visual elements, helping test the MLLMs' descriptive capabilities for different content. Third, these genres varied in audio and speech content, and hence, how frequently ADs are required may vary. Lastly, videos from these genres were frequently accessed and requested on YouDescribe, making them relevant contenders for audio descriptions [55].

In these genres, we chose videos that contained speech and visual references and thus could benefit from ADs. We did not include videos that primarily relied on speech or had a music soundtrack. Our initial pilot study with a blind user on 14 different videos suggested that the former category was already accessible and did not require ADs, and the latter category was difficult to use with user-driven ADs due to the limited auditory cues. Thus, we selected videos with a speech track where visual

elements were important (e.g., a workout video) or were referenced without explanation [38]. For example, in the *Beauty* video, the speaker said, "stop wearing your top like this," while pointing to her top without describing its appearance (Figure 1.1).

Audio Description Generation

Our goal was to evaluate the effectiveness of user-driven AI descriptions as a tool for BLV users to watch and interact with videos. However, generating descriptions and speech from visual frames can introduce delays and negatively impact the user experience. Moreover, the output of MLLMs is not deterministic and can change on each API call. Thus, we pre-generated ADs for the videos to run a controlled experiment and minimize the influence of these factors on the results. To generate the descriptions, we collected AD guidelines from online sources and prompted GPT-4V with the guidelines and video frames.

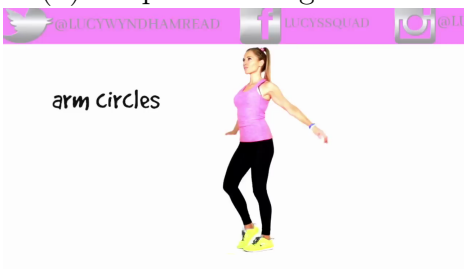
Collecting AD guidelines. We collected AD guidelines from four online sources. These guidelines focused on how to describe educational content [19], entertainment content [41, 46], and general guidelines for audio describers [50]. An example guideline was *"Description should convey facial expressions, body language and reactions, especially when in opposition to the dialogue. These elements can be omitted if they completely mimic the dialogue they are accompanying."* [46]. We removed guidelines that focused on timing, context, and audio content of the video, as these aspects cannot currently be used to prompt GPT-4V. This process resulted in 42 guidelines for prompting MLLMs.

Generating descriptions. For each video, we generated two types of ADs, a shorter version (*concise*) and a longer version (*detailed*), to let BLV participants choose the level of detail based on their preferences and video content. We created two versions by specifying the maximum length of descriptions in the input prompt:

(a) Test Video

(b) People and Blogs Video

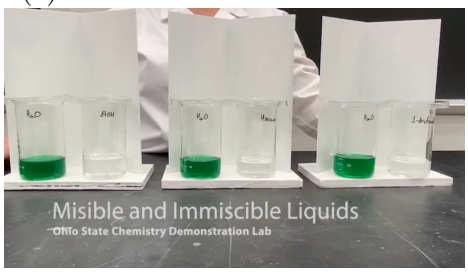
(c) Pets & Animals Video



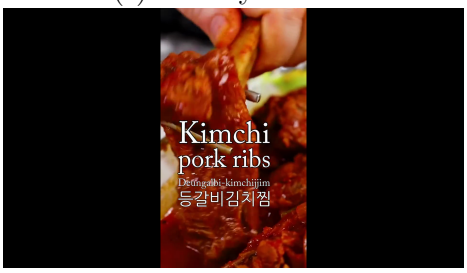
(d) Health & Fitness Video

(e) Film and Animation Video

(f) Beauty Video



(g) Education Video



(h) Food & Cooking Video

Figure 3.1: Figure (a) shows the test video used to familiarize participants with activating descriptions. Figures (b) through (h) illustrate the videos included in the study. The viewing order of the seven videos was randomized for each participant in the user study.

100 words for detailed descriptions and 25 words for concise descriptions. Based on our iterative prompt-engineering and testing, these limits were the most effective in creating the two types of descriptions without altering the prompt. On average, the detailed ADs were 13 words longer than the concise ADs. For a user-driven experience, we generated an AD for every second of the video using GPT-4V API. Each API call consisted of the prompt and ten frames from the video, with one frame extracted from each second. Thus, each API call described ten seconds of the video, requiring six API calls to describe a one-minute video. We repeated the process twice for the seven videos to get detailed and concise descriptions. During the AD generation process, there were a few instances when the generated descriptions referenced previous descriptions. For instance, in the *Film and Animation* video, the description for a frame started as: *"Same as description 5"*. Because these errors were few and required manual correction for removal, we decided to keep them in the descriptions and assess their impact on BLV users' experience.

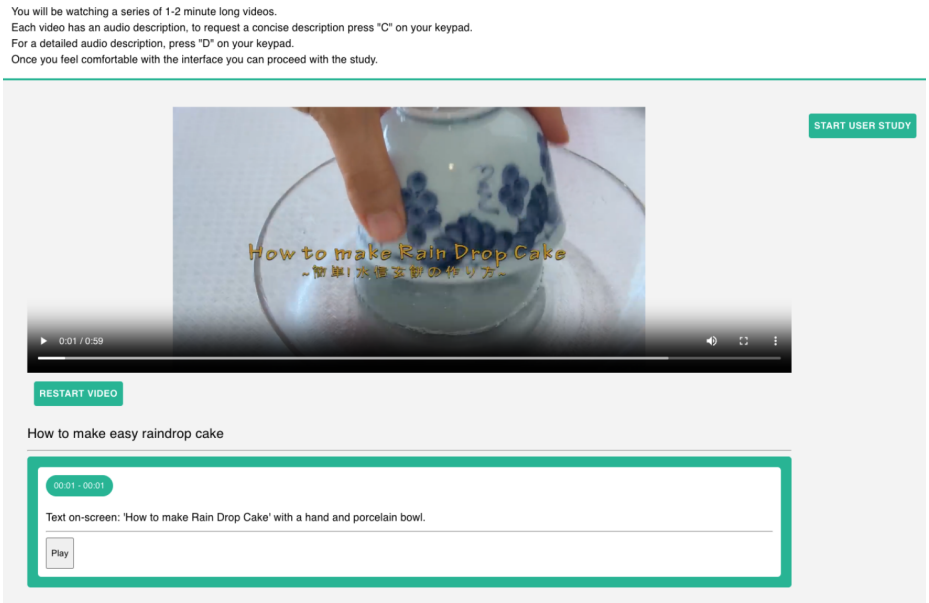


Figure 3.2: User interface for the prototype to test user-driven descriptions.

User Interface

We prototyped a simple interface to test the user-driven AI descriptions (Figure 3.2). We used ReactJs for creating the interface and Firebase for storing all the generated ADs and videos. Participants could press "C" on their keyboard to get concise descriptions and "D" for detailed descriptions. We used an extended audio description approach, where the video paused when the participant pressed a key, the text-to-speech model read out the description, and then the video playback continued. We used OpenAI's `alloy` text-to-speech model to read out the descriptions with a natural humanlike tone. The interface logged the key presses (C or D) and their timings. After watching a video, participants could rate their experience and type questions on the interface.

Chapter 4

USER STUDY

We conducted a user study to obtain data on the frequency and type of AD requests across different videos and BLV individuals and elicit user perception of this interactive approach to watching videos. The study was approved by our institution's IRB. The study was conducted over Zoom to facilitate participation from different locations. Each study session took about 75 minutes, and participants received $50 for their time.

Participants

We recruited 20 participants (7 female, 13 male) through BLV organizations, Facebook groups, and snowball sampling. The participants were between 27 to 79 years old and had various vision impairments ranging from legal blindness, blindness with some light/color perception, and total blindness (Table 4.1). Most participants (n=18) used screen readers to navigate the interface, while two participants used magnification.

Procedure.

The study session included four phases: completing a demographics survey, a practice segment, video viewing segment with user ratings, and a post-interview segment. After completing a demographic survey, participants interacted with a testing page to get familiar with the interface and used C or D on the keypad to activate descriptions. Then, participants watched the seven videos in random order with user-driven descriptions. They watched each video once and pressed the C and D keys

whenever needed to receive an AD for the previous second in the video. The UI logged the time and type of key presses. After each video, participants rated the efficiency, enjoyability, and effectiveness of user-driven ADs on a 5-point Likert-type scale from "strongly disagree" to "strongly agree". The participants could also type any questions regarding the video. Finally, we conducted a semi-structured interview with participants to gather in-depth feedback on their experience with user-driven ADs and their thoughts on the AI descriptions. The interviews were audio-recorded.

Data Analysis.

We analyzed the interview transcripts using thematic analysis inspired by Braun and Clarke's approach [16]. Specifically, two authors independently applied open coding to each transcript using MAXQDA qualitative analysis software. After coding every five interviews, the two authors met and discussed the similarities and differences between their codes before moving on to the rest. The authors also wrote memos to capture interesting patterns and relationships between codes. After open-coding all the transcripts, the two authors separately identified recurring patterns or themes from the codes and discussed the themes together. One author wrote a draft of the themes, and both authors discussed and revised it by referencing the codes. This process resulted in three themes, each including 2-4 sub-themes (Section 6). We also applied statistical testing on the user ratings, timing, and type of AD requests (Section 5).

P#	Age	Race	Gender	Visual Impairment	Screen Readers	AD Use
$P1_B$	63	White	Female	Total vision loss due to car accident	VoiceOver, JAWS	Frequently
$P2_B$	57	Unknown	Male	Blind due to retina detachment and glaucoma	JAWS	Frequently
$P3_{LV}$	68	Black	Female	Low vision since birth	VoiceOver, Zoom, Magnifying glass	Occasionally
$P4_B$	32	Unknown	Female	Totally blind due to retinoblastoma as an infant	JAWS	Occasionally
$P5_B$	64	White	Male	Born with glaucoma, had some sight until the age of 15. Lost sight at 15 and have no light perception, totally blind	VoiceOver, JAWS	Frequently
$P6_B$	44	White	Male	Totally blind	VoiceOver, JAWS	Frequently
$P7_B$	28	Asian	Male	Total congenital blindness	JAWS	Occasionally
$P8_B$	39	White	Male	Totally blind from retinal arterial occlusion	VoiceOver, JAWS	Frequently
$P9_B$	46	White	Male	Congenital blindness due to glaucoma and optic nerve damage	VoiceOver, JAWS	-
$P10_{LV}$	79	White	Male	Low Vision	VoiceOver, Mechanical magnifier	Occasionally
$P11_{LV}$	27	Hispanic/Latino	Male	Low vision since birth	VoiceOver, JAWS, Talk Back	-
$P12_{LB}$	40	American Indian	Female	Slowly losing vision due to stargardt disease	JAWS	-
$P13_B$	30	White	Female	Totally blind	JAWS, NVDA	Rarely
$P14_B$	59	Black	Male	Retinitis Pigmentosa, birth defect that led to progressive visual loss with age	JAWS, NVDA	Frequently
$P15_{LB}$	27	White	Male	Legally blind since birth, no vision in left eye	VoiceOver, Screen Magnification	Frequently
$P16_B$	38	White	Female	Total congenital blindness	JAWS, NVDA	Frequently
$P17_{LB}$	44	Hispanic/Latino	Male	Legally Blind, Light perception only Retinitis Pigmentosa	JAWS, NVDA	Frequently
$P18_B$	30	Asian	Male	Totally blind, almost since birth	VoiceOver, JAWS	Frequently
$P19_B$	42	Asian	Male	Totally blind caused by scarlet fever	VoiceOver, NVDA	Frequently
$P20_B$	32	Hispanic/Latino	Female	Total congenital blindness	VoiceOver, NVDA, JAWS	Frequently

Table 4.1: Description of the 20 BLV participants in our user study. The participant numbers are sub-indexed to indicate their vision status: Blind (B), Legally Blind (LB), and Low Vision (LV)

Chapter 5

QUANTITATIVE RESULTS

User Ratings

Figure 5.1 shows the distribution of user ratings for the efficiency in watching, effectiveness in comprehending the video, and enjoyability of videos with the user-driven ADs. The medians for efficiency and effectiveness are 4 (agree), while the median for enjoyability is 3 (neutral). The lower median for enjoyability compared to efficiency and effectiveness suggests that while the descriptions were useful, the overall enjoyment varied across the participants. Participants noted that the extended (rather than inline) presentation of user-driven ADs reduced their enjoyment of the videos.

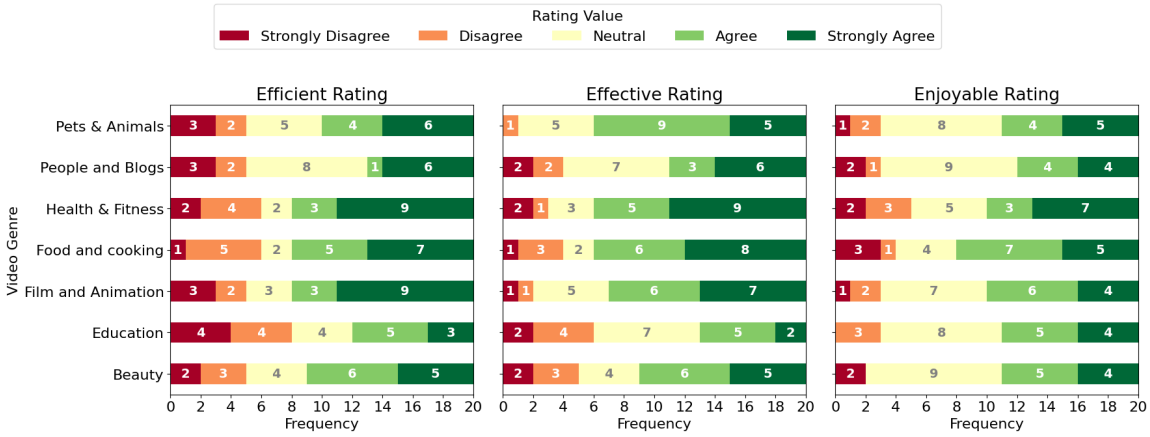


Figure 5.1: Frequency Distribution of Participant Ratings Across Video Genres

To assess the effect of video genre on user ratings, we ran a Friedman test on each of the three ratings with one factor of video genre (seven levels). There was a statistical difference in perceived efficiency ($\chi^2(2) = 15.702$, $p = 0.015$, W=0.131) and

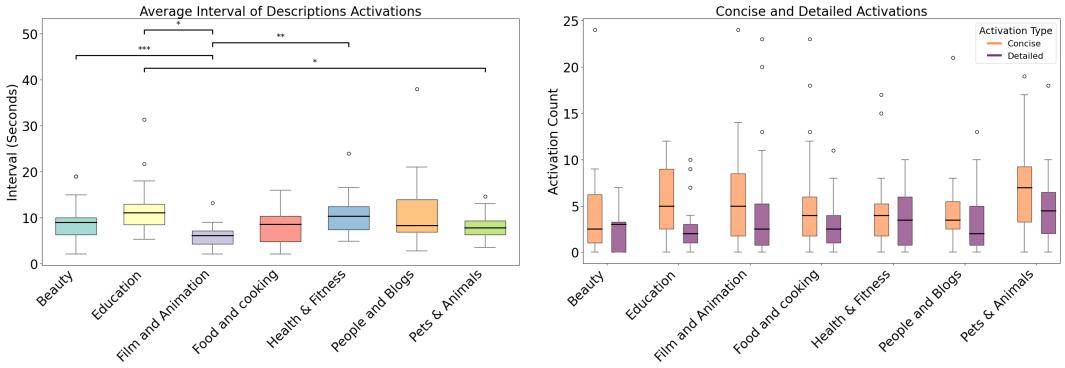
effectiveness ($\chi^2(2) = 15.582$, $p = 0.016$, W=0.130) but no statistical difference for the ratings of enjoyability, ($\chi^2(2) = 1.716$, $p = 0.944$, W=.014). The post hoc analysis on efficiency and effectiveness ratings using Wilcoxon signed-rank tests and Holm-Bonferroni correction showed no statistical differences between ratings of different genres. The effect sizes, calculated using Kendall's W, were small for all three ratings which suggests that the video genre had little practical impact on the user ratings. Spearman's rank correlation analysis showed strong correlations between efficiency and effectiveness ($\rho = 0.83$) ratings and between the efficiency and enjoyment ratings ($\rho = 0.67$) and a moderate relationship between effectiveness and enjoyment ratings ($\rho = 0.54$).

Frequency and Type of AD Requests

We conducted a repeated measures ANOVA to determine statistical differences in the time intervals between subsequent AD activations (Figure 5). The mean scores of the AD intervals were statistically significantly different, $F(6, 114) = 5.460$, $p < .001$, η^2=0.223. Post hoc pairwise comparison using the Bonferroni correction showed a significant difference in the AD intervals for *Education* vs. *Film and Animation* ($p = 0.012$), *Film and Animation* vs. *Health and Fitness* ($p = 0.009$), *Beauty* vs. *Film and Animation* ($p < 0.001$), and *Education* vs. *Pets and Animals* ($p = 0.046$). The results highlight that the frequency of descriptions required can significantly vary for different genres of video content.

For the number of AD activations, the assumption of normality was violated. Thus, we conducted Aligned Rank Transform (ART [68]), a non-parametric alternative to two-way repeated measures ANOVA, to investigate the effects of request type (concise, detailed) and video genre on the counts of AD activations (Figure 5). The description type had a significant main effect $F(1, 247) = 12.52$, $p = 0.00048$, indicat-

ing concise descriptions were activated more frequently ($Mean = 5.42$, $SD = 5.26$) than detailed ones ($Mean = 3.58$, $SD = 3.95$) for all the videos. There was no significant effect of video genre $F(6, 247) = 2.11$, $p = 0.053$, and no interaction effect between video genre and the type of requests $F(6, 247) = 0.44$, $p = 0.85$.



(a) The time intervals between AD activations

(b) Number of concise and detailed AD activations

Figure 5.2: Results for the frequency and type of AD activations in the seven videos. Asterisks (*) mark significant differences, with more asterisks indicating higher levels of significance.

The patterns of AD activations showed variations among BLV individuals (Figure 5.3). Overall, the frequency of AD activations depended on the video type and the participants' interests. For example, $P4_B$ had fewer AD activations for the *Beauty* and *Education* videos than other videos, $P10_{LV}$ activated only one AD for *Health and Fitness*, and $P11_{LV}$ had few requests for *Food and Cooking*. The frequency and type of AD requests also varied drastically depending on the BLV individuals. Some participants activated ADs frequently (e.g., $P4_B$ and $P19_B$), whereas others activated AD less often (e.g., $P5_B$, $P3_B$, $P10_{LV}$, $P11_{LV}$). Some participants primarily used either concise or detailed ADs (e.g., $P3_{LV}$, $P4_B$, $P19_B$), whereas some others used both types of ADs depending on the video content (e.g., $P5_B$, $P11_{LV}$). Furthermore, the three low-vision participants in our study activated ADs less than the blind participants, perhaps due to relying on their functional vision. These variations further

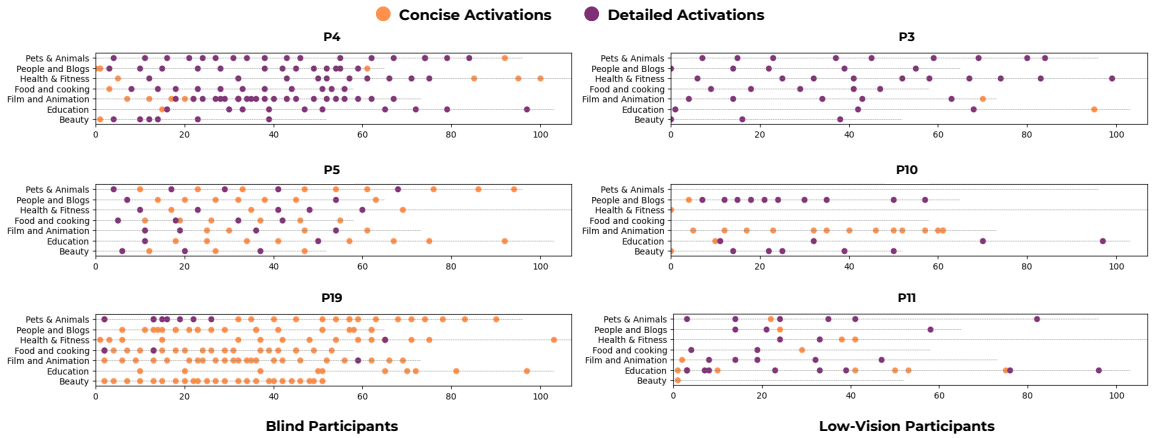


Figure 5.3: Patterns of concise and detailed AD activations from example blind and low vision participants. On each plot, the rows show the seven videos, and the horizontal axis shows the video timeline in seconds. The duration of the videos was between 52 and 107 seconds.

suggest the differing AD needs and preferences of BLV individuals.

Chapter 6

QUALITATIVE FINDINGS

We identified three themes based on the interviews and observation of participants' experiences with user-driven AI descriptions (Figure 6.1).

User-driven Description: Redefining Control for Accessibility

The concept of control was mentioned repeatedly in participant descriptions of AD activations. The options to say when and how the descriptions were delivered created a sense of control that was missing from pre-recorded ADs. As $P15_{LB}$ highlighted: "*We're unlocking this absolutely mind-blowing layer that we can now add, which is an interactive audio description, the idea of having not only a description of what's going on, but a way to have the description describe to us what we want it to describe to us, how we want it to describe it to us.*" Another participant also echoed this sentiment, noting the limitations of pre-recorded ADs: "*When you have audio description built in, what is described is already predetermined... you can't really change that.*" ($P4_B$).

Participants enjoyed the flexibility of how and when they received ADs, giving them the ability to "seek information" ($P4_B$) exactly when they wanted it. For example, $P6_B$ activated descriptions based on whether they wanted to hear the person in the video or more detailed information about the visual content. Participants likened this user-driven description style to having a live "human describer". The ability to control ADs also helped cater to the needs of BLV individuals with varying visual impairments. For example, $P15_{LB}$ expressed their frustration with videos containing on-screen text, highlighting how he has to magnify or use OCR to read text; thus, he could get on-screen text in a more accessible manner using this interaction. In

contrast to pre-recorded ADs, the user-driven descriptions made the participants feel more active and engaged when watching videos. With this sense of control, $P3_{LV}$, $P7_B$, and $P9_B$ imagined creating and saving ADs that they or others could use to watch, summarize, or preview video content later.

Challenges of user-driven ADs. The increased user control came with its challenges. One of the primary challenges was the disruption of video viewing experience and flow: "*It's easy to forget what was happening in the video if you use the on-demand descriptions too much*" ($P17_B$). This disruption was more pronounced for content with greater visual details or little to no silent gaps.

The increased control associated with user-activated AD also led to a higher cognitive load. Participants had to actively engage with the video to know when to receive descriptions. This was mentally taxing, especially for content that was information-dense or had fewer silent gaps to activate descriptions. As $P16_B$ noted, "*I think the harder ones were things like the dog running through the course (Pets & Animals). That's really hard because things are changing really fast... so it's just really challenging to keep up with that.*" This cognitive load reduced utility for some participants, as they had to balance their attention between the content and activating AD.

User-driven ADs also instilled a fear of missing out (FOMO) visual information in some participants who expressed concern over missing critical information if they activated descriptions infrequently. This concern was particularly felt in fast-paced videos, where they could discern scene changes from audio but could not activate to keep pace with the scene changes: "*I was requesting descriptions too often, because I couldn't make those judgment calls.* ($P7_B$)." Some participants felt they needed to press the keys at the right time or that they didn't press at the right moment to request the AD. This sense of FOMO could be a direct result of needing more time with user-driven ADs since it is a new way of receiving descriptions, as highlighted

by five participants.

Strategies for activating ADs. Participants had different strategies for activating descriptions. While some felt they activated descriptions randomly without any particular pattern ($n = 3$), most noted that they relied on the audio track to decide when to activate an AD ($n = 8$). They often avoided activating AD during conversations or speech and waited for a pause or silent gap in the video. Others noted activating descriptions when there were visual references (e.g., *"do it like this*" for the *Health & Fitness* video – $P18_B$), when a change in the ambient sounds indicated a scene change ($P19_B$), or when they wanted to know the source of sounds in the audio track (e.g., *"Chopping up something, okay what?*" – $P13_B$). User-driven ADs were particularly helpful for videos with difficult speech. For the *Food and Cooking* video, even though the presenter explained all the steps, the speech was difficult to follow for some participants due to the presenter's accent, and the AD activations helped them understand the visual content better.

Participants also mentioned other strategies. Some noted that they requested descriptions early in the videos "*to get a context*" ($P9_B$, $P6_B$, $P20_B$). Their interest in the content also affected when and how frequently they requested descriptions, with fewer and concise requests for videos that did not interest them. Finally, low-vision participants also leveraged their functional vision to request ADs, especially to read on-screen text. "*Where it's really helpful is the text, because reading text is something I just don't do when it comes to videos.*" ($P10_{LV}$).

Preference for ADs: Dependent on the Video Content, Context, and BLV Individuals.

Participants noted that various forms of ADs (pre-recorded vs. user-driven, concise vs. detailed) supported different uses and that *video content* and *viewing context*

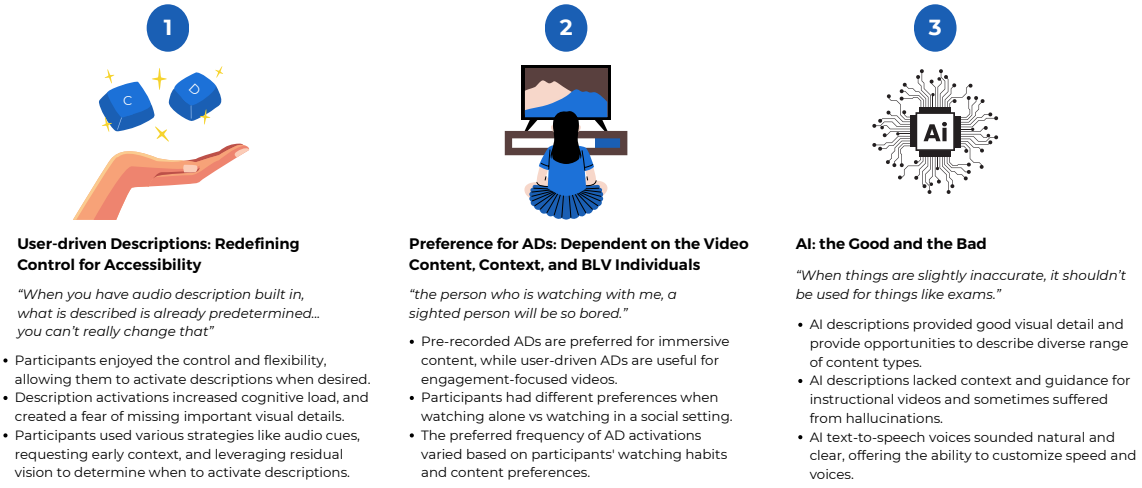


Figure 6.1: Overview of the three themes and sub-themes from interviews about BLV users' experience with user-driven AI descriptions.

played a crucial role in their preference for AD type.

Content dependency. Conventional or pre-recorded AD format was preferred for entertainment content, where an immersive experience was essential ($n = 6$). Conversely, user-driven ADs were useful for content requiring active engagement over immersion, such as educational, workout, and cooking videos, where the ability to request ADs precisely when the action was happening enhanced their understanding of the content ($n = 9$). The preference for concise vs. detailed ADs also depended on the video. For fast-paced videos or action sequences, concise descriptions were preferred to get sufficient information while minimally disrupting the video flow. Conversely, participants preferred detailed descriptions for heavily visual content (e.g., *Film and Animation*), as $P9_B$ highlighted: "*The detailed ones to me like they create a visual image in my mind more than you know short ones.*" For instructional videos, participants preferred to have detailed descriptions but noted that the detailed information was unrelated to the context of the video, which reduced the utility of longer descrip-

tions for instructional content.

Context dependency. Participants' preference for the form of AD also depended on the viewing context, particularly whether they were watching or re-watching a video and whether they were viewing it alone or in a social setting (i.e., with sighted people). Participants found user-driven descriptions especially useful when re-watching videos. Sometimes, the purpose of re-watching was to get a new perspective or experience on the content. "*some of these Miyazaki films... there's all these different nuances to it...if you watch it again and you click, describe at different times, you're gonna understand all the foreshadowing pictures and things like that. So that part's a huge benefit.*" ($P12_{LB}$). In other cases, they wanted to re-watch the content for specific information. For instance, $P19_B$ noted watching do-it-yourself videos with concise ADs first and activating detailed descriptions when re-watching a part of the video later to find specific information. Participants also described different AD preferences when watching alone vs. with sighted people. $P6_B$ highlighted that in social contexts, "*the person who is watching with me, a sighted person will be so bored.*" Hence, user-driven would be preferred more in a private setting rather than a social one. Similarly, $P12_{LB}$ noted that they would watch a video with frequent AD activation alone, then "*...watch it with other people and only click, describe occasionally when you're like, oh, I forgot what happened.*"

Individual differences in AD use and preference. The individual experiences of participants varied widely, which further underscores the importance of adaptable and customized ADs. Participants who appreciated having user-driven AD often cited it as a pleasant, enjoyable experience, and a promising alternative to conventional AD ($P1_B$, $P6_B$, $P14_B$, $P17_{LB}$, $P18_B$). $P6_B$ noted: *"I was trying to, you know, get as much as possible."*. However, not everyone shared the same narrative, and some found activating descriptions constantly to be overwhelming and unpleas-

ant ($P8_B$, $P10_{LV}$, $P20_B$). The varied and diverse opinions also translated into how participants activated descriptions, as shown in Figure 5.3. When asked if they would use user-activated ADs in daily video viewing, the preferred frequency of use varied widely among participants depending on their watching habits. Those who watched instructional and educational videos stated a higher frequency of use: "*it would be nice to kind of get the option...*"($P12_{LB}$). Others, however, felt the disruption that comes with user-driven AD would not be suitable for the content they watch, particularly entertainment content. Hence why, participants felt that user-driven ADs would be more helpful for YouTube videos rather than movies or TV shows ($P6_B$, $P15_{LB}$, $P19_B$). Also, participants who mentioned lower frequency of use felt that the content they watched was audio-based and did not require AD ($P7_B$, $P8_B$, $P16_B$). These participants also highlighted how they would use it for occasional descriptions: "*Seth Meyers in his show... a lot of his stuff has like silly graphics or things that they show. So when I hear him say something, a pause, and the audience like cracks up, I can run it and say, okay, there's a funny picture being shown of something, so that would be useful. ($P8_B$)*"

AI: the Good and the Bad

Participants had varied opinions about the effectiveness of AI descriptions in the study and in general based on their prior experience and the types of content they watched. Several saw it as a promising future, while some felt skeptical about how well AI can describe different forms of content ($P11_{LB}$, $P15_{LB}$).

AI descriptions: the good. Ten participants highlighted the quality of the AI-generated descriptions, particularly appreciating the visual detail. Unlike human ADs, which incur substantial costs in production, AI provided an *"opportunity to describe more content"* . The need for more content to be described was evident, as

participants reported watching 11 unique genres. At the same time, they mostly only received descriptions for Film and TV shows. P19$_B$ found that AI descriptions will increase the range of content they can watch: *"You know, like the [Legend of] Zelda, I would love to be able to sit there through a full description narration of game play-through."*. Additionally, participants acknowledged the benefit of AI descriptions to allow customization: *"the power of the technology will allow us to give people these choices to give them the concise or more flowery detail... because up to this point, you know, blind people have had very, very little described, and when we've had something described, we've had very little say in how it's described to us. (P9$_B$)"*.

Some participants noted the benefits of AI-generated descriptions over human descriptions. P15$_{LB}$, who enjoyed the user-driven ADs, highlighted how paid sighted workers are often untrained to provide AD and found these AI descriptions were "fine-tuned" and had an "audio description vibe to them". However, others felt the AI descriptions were similar to novice describers but failed to compete with professional descriptions (P7$_B$, P8$_B$). Participants also noted that AI may help remove the subjective censorship by human describers. P1$_B$ mentioned how narrators are often too careful in describing people and characters to the point where they *"lose some of the experience"*. A similar concern was echoed by P11$_{LV}$, who emphasized that descriptions should be on point and *"if you're embarrassed to narrate... I would hope you give your job to someone else."* These comments illustrate a potential for how AI descriptions can help bridge the gap between unbiased and accurate descriptions, potentially complementary to human AD.

AI descriptions: the bad. AI descriptions could lack the context that human describers would provide, which is important for instructional content. Though the AI descriptions provided visual detail, the descriptions lacked detailed guidance to follow along for instructional and educational content ($n = 6$). Participants also

noted that some descriptions did not adequately relate to the main audio track and were confusing as they provided more "static" rather than dynamic information, exemplified in the *Health and Fitness* video, as highlighted by P16$_B$: *"The motion was important to the content of the video... you have to know first you raise your right hand and your left leg, and that stuff is really really critical, and you can't capture that."* On the other hand, AI descriptions sometimes assumed a certain context that was not provided. Even with a structured prompt, some descriptions referenced previous descriptions, which led to confusion as P7$_B$ stated that they *"don't know what description 5 was"*. Although the prompt explicitly mentioned reading on-screen text central to understanding, the concise descriptions often failed to contain the on-screen text and only mentioned "on-screen text." This further increased FOMO as expressed by P19$_B$: *"What if one of the videos have a secret formula written on the screen, and only you know that the viewer... the sighted viewer could see that. But to the blind person, this just says there's text on the screen."*

AI descriptions were also prone to hallucinations in some cases. The hallucinations were of particular concern for educational settings: *"when things are slightly inaccurate, it shouldn't be used for things like exams (P13$_B$)"*. Some participants explicitly noted concern with hallucinations in AI (P8$_B$). This was true even when AI descriptions were accurate. For example, in the comment for *Pets & Animal* video, P18$_B$ wrote: *"The dog part was mentioned towards the end of the video. However, the AD mentioned about the dog much before. This was helpful in getting the context into the video. Till the video spoke of dog, I was skeptical that the AI is calling something else a dog."* These comments indicate lower trust in AI-generated vs. human descriptions.

Finally, the descriptions did not adapt to the participants' backgrounds. While some participants enjoyed the visual *"flowery detail (P6$_B$)"* provided by the descrip-

tions, others highlighted how the descriptions implored too much background information. For instance, the mention of colors in the description made no sense to $P16_B$ as a congenitally blind person. These aspects highlight further expectations for flexibility from AI descriptions.

AI Text to Speech (TTS) voices: the good and bad. An appreciated aspect of TTS voices was how natural and clear they sounded compared to technologies used by the participants in screen readers. In fact, $P1_B$ thought a live person was speaking out the descriptions upon their key presses. Participants enjoyed the tone of the AI voice, highlighting the voice was clear and easy to understand ($n = 4$). They appreciated having a more distinct voice for the descriptions from all videos, so *"you get to be comfortable with that voice"* ($P2_B$). Several participants also highlighted the customization advantages of AI TTS, particularly the ability to increase speed ($P9_B$, $P13_B$) and to change between male or female voices based on content ($P11_B$, $P15_{LB}$). The most common feedback from participants was wanting to speed up the TTS ($n = 9$), noting that a faster TTS would enhance the efficiency of user-driven ADs. Speech quality was less important and only was a source of relevance for entertainment content. Some participants noted AI voices can sound monotonous, lacking the emotional depth that a human describer would provide ($P12_{LB}$, $P19_B$). For other video genres, participants preferred a faster TTS to browse the video quickly. Overall, the quality of the AI descriptions was far more important to the participants than the quality of TTS, *"so long as the quality of the description itself is high. that's all that matters" ($P19_B$).*

Chapter 7

DISCUSSION

Through detailed participant feedback, we identified the benefits and challenges associated with user-driven descriptions. Although participants appreciated the control offered by a user-driven interaction, they highlighted the cognitive load that it incurs. AI descriptions aided visual understanding but lacked context. Additionally, statistical analysis revealed that concise ADs were requested more often, and the time interval between descriptions varied significantly across genres. We discuss the need for an online AI-based AD platform, the evolving role of describers and BLV users in AI-assisted AD, and the potential of multisensory interactions in AD consumption.

Implications for an AI Description Platform

An online platform for AI-generated descriptions must offer various forms of ADs for watching online videos. In our study, BLV users watched various online content and wanted to use pre-recorded, user-driven, concise, and detailed ADs depending on the video content, social context, and their individual needs when watching a video. While some preferred having pre-recorded AD for entertainment content, others enjoyed having control over the ADs. Also, information that seemed trivial to some could be significant to others, such as the use of colors in descriptions. The diversity of BLV users' opinions and their patterns of AD activations underscores the need for customization in AI descriptions and AD systems. Relatedly, recent research have investigated various AD types and question-answering for BLV users [49, 72, 9, 61]. These AD approaches can be integrated into an online video-sharing platform (e.g., YouTube or YouDescribe) to enable BLV users to effectively watch diverse video

content. With BLV users' permission, such an online platform can collect data on variations in video viewing preferences of BLV users (e.g., AD frequency, type, questions) over time to improve the AI descriptions and learn to time ADs depending on video content and user needs. This crowdsourced approach can further enable new ways for BLV users to skim or re-watch videos or receive video descriptions.

While AI descriptions present an opportunity for increasing video accessibility, some nuances need to be taken into account. Descriptions generated using GPT-4-vision were high in accuracy, but a few descriptions still included object hallucinations (e.g., labeling "pork ribs" as "chicken" in the *Food and Cooking* video) [30]. This highlights the importance of human verification for AI descriptions when factual correctness cannot be determined via the audio track, or for information critical content [62]. Additionally, contextual relevance can be a source of concern, which makes standalone AI descriptions insufficient for long-form content such as films and documentaries. Furthermore, even with prompt engineering, descriptions generated using MLLMs are prone to consistency issues. Although some inconsistencies might not notably affect the BLV user experience, issues such as self-reference to previous descriptions can increase the cognitive load for BLV users and reduce enjoyment. Addressing accuracy, context and consistency will improve the reliability of using AI descriptions.

Change in Roles for Describers and BLV Users

With recent advances in AI descriptions, there is a potential for the roles of sighted volunteers to change from creating descriptions to adjusting the timings of receiving descriptions and removing hallucinations and inaccuracies. Prior work has explored how preset sentence templates can enhance descriptions generated by novice describers for images [42]. Similarly, creating descriptions with automated

feedback [45] and automating video text generation and scene segmentation [72] was easier for novice describers and significantly improved the AD quality for BLV users. While previous efforts have mainly focused on using AI to support novice describers, with newer MLLMs, the role of describers could shift. In terms of providing visual details, most participants felt the AI descriptions were adequate. However, several participants needed help knowing when to request descriptions. Instead of volunteers working to create descriptions, they can work on when to insert descriptions. Although libraries and models exist that help detect silent gaps in audio for the insertion of descriptions, these fail for short-form content where there are fewer or no silent gaps. Some BLV users also had low trust in AI-generated descriptions which worsened with even small AD inaccuracies. Sighted users can easily detect such hallucinations and inconsistencies and correct them to improve BLV user comprehension, trust, and overall experience.

With user-driven AI descriptions, the role of BLV users can also further shift toward from video consumption to active content creation. Some BLV users in our study were excited about the possibilities of user-driven ADs and wanted to save and share ADs (and variations on every re-watching) for the videos. Also, compared to sighted volunteers, low-vision volunteers have a better idea of when descriptions are needed for a video (e.g., on-screen text, visual reference). While there are already BLV content creators [74, 26], a user-driven AD tool can open up the space to a wider range of BLV users to become AD creators, determining when and how ADs should be inserted for different videos based on their lived experience. This role change can make AD creation an inclusive space for BLV describers.

User-Driven AD with Multisensory Interactions

Prior research has looked into using multi-sensory AD approaches for artwork [47, 12, 28] and for AD in movies [65, 3] to improve visualization. Subjects have also highlighted how they would like additional output modalities such as audio cues, tactile graphics, and haptics to enhance the AD experience for different viewing scenarios (e.g., how-to, short-form, comedy, drama) [31]. Though this requires technology that might not always be available to BLV users, several output modalities can still be incorporated. Our participants wished to know when to activate AD for user-driven style AD. Audio cues, or vibrations, can be incorporated into the video to let the BLV user know when they can activate a description. This way, the control of receiving descriptions still lies with the user, and they have a better idea of when there are scene changes or where they can benefit from activating descriptions, effectively reducing the cognitive load of not knowing when to request descriptions or having too many descriptions in the case of predetermined AD. There is limited work done on the intersection of AD and other modalities, especially for different video content. Future work can explore whether other modalities complement or obstruct BLV users' attention to AD.

Limitations and Future Work

Our work has several limitations. First, our study covers a subset of video types that BLV users wish to watch. Within a 75-minute study, we could only test seven video genres. Also, all the videos were relatively short. A lot of variation exists in content within each genre, and the pace of video has also likely impacted the quantitative results. Testing user-driven AD with other genres, videos with different paces within a genre, and longer videos can give further insights into the efficacy of user-

driven descriptions for different video types. Second, we created all the descriptions before the user study to account for latency and description quality. Pre-generating descriptions allowed us to reduce time delays and control for the effect of variations in AI generation on BLV user ratings. This approach also enabled us to process any formatting issues in the generated descriptions. A user-driven AD platform must include automated methods to mitigate these issues in the descriptions generated on the spot. Third, we investigated BLV people's perception of user-driven ADs in a single session. Some participants highlighted they would require more time with this new way of watching videos, and their frequency of use might differ for the content they watch on a regular basis. To gather more information on user-driven AD, a longitudinal study in a more organic setting needs to be conducted. This would provide further insights into how BLV users' interest in the video content and their feelings about user-driven ADs can change over a more extended period and how these factors can impact the frequency of requests for description.

Chapter 8

CONCLUSION

We presented an alternative approach to video accessibility by giving more control to the user through user-driven descriptions with two types of detail. Supported by quantitative results and interviews with 20 BLV participants, user-driven descriptions improved the sense of user control with flexibility of when and how participants receive descriptions but increased cognitive load and FOMO. BLV users found the descriptions compelling in terms of the detail of visual content and saw a potential for AI to describe more types of content. There were also concerns about the misuse of AI and its ability to describe certain kinds of genres. With the rapid advances in AI, we hope our results can inform future work on tailoring AI descriptions based on user preferences and customization needs and open up further possibilities for BLV users to access and create AD content based on their needs and lived experiences. Finally, by integrating AI-generated ADs with assistive technologies, such as service robots [52, 53, 29], we envision a future where BLV users can engage with [25] and interact [58, 59, 5, 4] in physical environments more seamlessly. This integration would enable robots to provide real-time, context-aware descriptions of surroundings, objects, and activities, empowering BLV users to navigate [22, 23, 24] spaces with greater confidence and independence.

REFERENCES

[1] 3Play Media. Audio Description: What It Is and How It Works. https://www.3playmedia.com/learn/popular-topics/audio-description/, 2024.

[2] 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*, 52(6), oct 2019.

[3] Damien Ablart, Carlos Velasco, and Marianna Obrist. Integrating mid-air haptics into movie experiences. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, TVX, page 77–84, 2017.

[4] Ramzi Abou Chahine, Sid Padmanabhuni, Pooyan Fazli, and Hasti Seifi. Designing and evaluating interactive tools for a robot hand collection. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, pages 328–332, 2023.

[5] Ramzi Abou Chahine, Steven Vasquez, Pooyan Fazli, and Hasti Seifi. Clustering social touch gestures for human-robot interaction. In *International Conference on Social Robotics*, pages 53–67. Springer, 2023.

[6] American Council of the Blind. Beginner's Guide to Audio Description. https://adp.acb.org/docs/Beginners-Guide-to-Audio-Description.pdf, 2020.

[7] American Council of the Blind. Audio Description Guidelines and Best Practices. https://www.acb.org/adp/guidelines.html, 2024.

[8] Anthropic. Meet claude, 2024.

[9] Aditya Bodi, Pooyan Fazli, Shasta Ihorn, Yue-Ting Siu, Andrew T Scott, Lothar Narins, Yash Kant, Abhishek Das, and Ilmi Yoon. Automated video description for blind and low vision users. In *Proceedings of the ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI, 2021.

[10] Carmen Branje and Deborah Fels. Livedescribe: Can amateur describers create high-quality audio description? *Journal of Visual Impairment Blindness*, 106:154–165, 03 2012.

[11] Virginia Campos, Tiago Araujo, Guido Souza Filho, and Luiz Gonçalves. Cinead: A system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society*, 19, 03 2020.

[12] Luis Cavazos Quero, Jorge Iranzo Bartolomé, and Jundong Cho. Accessible visual artworks for blind and visually impaired people: comparing a multimodal approach with tactile graphics. *Electronics*, 10(3), 2021.

[13] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Anhong Guo. Omniscribe: Authoring immersive audio descriptions for 360° videos. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, UIST, 2022.

[14] Agnieszka Chmiel and Iwona Mazur. A homogenous or heterogeneous audience? audio description preferences of persons with congenital blindness, non-congenital blindness and low vision. *Perspectives*, 30(3):552–567, 2022.

[15] Cheng-Yu Chuang and Pooyan Fazli. Clearvid: Curriculum learning for video description. *arXiv preprint arXiv:2311.04480*, 2023.

[16] Victoria Clarke and Virginia Braun. *Thematic Analysis: A Practical Guide*. Sage Publications Ltd, Thousand Oaks, California, USA, 2021.

[17] Google DeepMind. Gemini, 2024.

[18] Marco Del Fabro and Laszlo Böszörmenyi. State-of-the-art and future challenges in video scene detection: a survey. *Multimedia Systems*, 19(5):427–454, 2013.

[19] Described and Captioned Media Program (DCMP). Description Key for Educational Media. `https://dcmp.org/learn/descriptionkey`, 2024.

[20] Benoît Encelle, Magali Ollagnier Beldame, and Yannick Prié. Towards the usage of pauses in audio-described videos. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*, W4A, 2013.

[21] Nazaret Fresno, Judit Castellà, and Olga Soler-Vilageliu. 'what should i say?'tentative criteria to prioritize information in the audio description of film characters. *Researching Audio Description: New Approaches*, pages 143–167, 2016.

[22] Mahmoud Hamandi, Mike D'Arcy, and Pooyan Fazli. Learning to navigate like humans. In *Proceedings of the Workshop on Learning and Inference in Robotics, Robotics: Science and Systems Conference (RSS)*, 2018.

[23] Mahmoud Hamandi, Mike D'Arcy, and Pooyan Fazli. Deepmotion: Learning to navigate like humans. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–7. IEEE, 2019.

[24] Mahmoud Hamandi and Pooyan Fazli. Online learning of human navigational intentions. In *Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10*, pages 1–10. Springer, 2018.

[25] Mahmoud Hamandi, Emre Hatay, and Pooyan Fazli. Predicting the target in human-robot manipulation tasks. In *Social Robotics: 10th International Conference, ICSR 2018, Qingdao, China, November 28-30, 2018, Proceedings 10*, pages 580–587. Springer, 2018.

[26] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang 'Anthony' Chen, Young-Ho Kim, and Amy Pavel. Avscript: Accessible video editing with audio-visual scripts. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI, 2023.

[27] Shasta Ihorn, Yue-Ting Siu, Aditya Bodi, Lothar Narins, Jose M Castanon, Yash Kant, Abhishek Das, Ilmi Yoon, and Pooyan Fazli. Narrationbot and infobot: A hybrid system for automated video description. *arXiv preprint arXiv:2111.03994*, 2021.

[28] Jorge Iranzo Bartolome, Luis Cavazos Quero, Sunhee Kim, Myung-Yong Um, and Jundong Cho. Exploring art with a voice controlled multimodal guide for blind people. In *Proceedings of the International Conference on Tangible, Embedded, and Embodied Interaction*, TEI, page 383–390, 2019.

[29] Jianmin Ji, Pooyan Fazli, Song Liu, Tiago Pereira, Dongcai Lu, Jiangchuan Liu, Manuela Veloso, and Xiaoping Chen. Help me! sharing of instructions between remote and heterogeneous robots. In *Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8*, pages 786–795. Springer, 2016.

[30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023.

[31] Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. "it's kind of context dependent": Understanding blind and low vision people's video accessibility preferences across viewing scenarios. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–20, 2024.

[32] Lucy Jiang and Richard Ladner. Co-designing systems to support blind and low vision audio description writers. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS, 2022.

[33] Georgina Kleege and Scott Wallin. Audio description as a pedagogical tool. *Disability Studies Quarterly*, 35(2), 2015.

[34] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. Providing synthesized audio description for online videos. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS, page 249–250, 2009.

[35] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2203.05081*, 2022.

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, pages 34892–34916. Curran Associates, Inc., 2023.

[37] Tianhang Liu, Pooyan Fazli, and Heejin Jeong. Artificial intelligence in virtual reality for blind and low vision individuals: Literature review. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, page 10711813241266832. SAGE Publications Sage CA: Los Angeles, CA, 2024.

[38] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. What makes videos accessible to blind and visually impaired people? In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI, 2021.

[39] Xingyu "Bruce" Liu, Ruolin Wang, Dingzeyu Li, Xiang 'Anthony' Chen, and Amy Pavel. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding. *arXiv preprint arXiv:2208.11144*, 2022.

[40] Iwona Mazur and Agnieszka Chmiel. Audio Description Training: A Snapshot of the Current Practices. *The Interpreter and Translator Trainer*, 15(1):51–65, 2021.

[41] Media Access Canada (MediAC). Described Video Best Practices Guidelines, 2012.

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

[43] Rosiana Natalie, Ruei-Che Chang, Smitha Sheshadri, Anhong Guo, and Kotaro Hara. Audio description customization. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*, pages 1–19, 2024.

[44] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebrima H Jarjue, Hernisa Kacorri, and Kotaro Hara. The efficacy of collaborative authoring of video scene descriptions. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS, 2021.

[45] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. Supporting novices author audio descriptions via automatic feedback. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI, 2023.

[46] Netflix Studios. Audio Description Style Guide v2.5. `https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-5`, 2024.

[47] Josélia Neves. Multi-sensory approaches to (audio) describing the visual arts. *MonTI. Monografías de Traducción e Interpretación*, (4):277–293, 2012.

[48] Nguyen Nguyen, Jing Bi, Ali Vosoughi, Yapeng Tian, Pooyan Fazli, and Chenliang Xu. Oscar: Object state captioning and state change representation. *arXiv preprint arXiv:2402.17128*, 2024.

[49] Zheng Ning, Brianna L Wimer, Kaiwen Jiang, Keyi Chen, Jerrick Ban, Yapeng Tian, Yuhang Zhao, and Toby Jia-Jun Li. Spica: Interactive video content exploration through augmented audio descriptions for blind or low-vision viewers. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI, 2024.

[50] Ofcom. Provision of TV Access Services: Guidelines. `https://www.ofcom.org.uk/siteassets/resources/documents/consultations/category-2-6-weeks/178126-review-of-access-services-code/associated-documents/provision-of-tv-access-services-guidelines.pdf`, 2024.

[51] OpenAI. Gpt-v system card. Technical report, OpenAI, 2023.

[52] Utkarsh Patel, Emre Hatay, Mike D'Arcy, Ghazal Zand, and Pooyan Fazli. Beam: A collaborative autonomous mobile service robot. In *2017 AAAI Fall Symposium Series*, 2017.

[53] Utkarsh Patel, Emre Hatay, Mike D'Arcy, Ghazal Zand, and Pooyan Fazli. Setting up the beam for human-centered service tasks. *arXiv preprint arXiv:1710.06831*, 2017.

[54] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the ACM Symposium on User Interface Software and Technology*, UIST, page 747–759, 2020.

[55] C. Pitcher-Cooper, M. Seth, B. Kao, J.M. Coughlan, and I. Yoon. You Described, We Archived: A Rich Audio Description Dataset. *Journal of Technology and Persons with Disabilities*, 11:192–208, May 2023.

[56] Sonali Rai, Joan Greening, and Leen Petré. A comparative study of audio description guidelines prevalent in different countries, 2010.

[57] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models, 2023.

[58] Hasti Seifi, Steven A Vasquez, Hyunyoung Kim, and Pooyan Fazli. Charting visual impression of robot hands. *arXiv preprint arXiv:2211.09397*, 2022.

[59] Hasti Seifi, Steven A Vasquez, Hyunyoung Kim, and Pooyan Fazli. First-hand impressions: charting and predicting user impressions of robot hands. *ACM Transactions on Human-Robot Interaction*, 12(3):1–25, 2023.

[60] Joel Snyder. Audio description: The visual made verbal. *International Congress Series*, 1282:935–939, 09 2005.

[61] Abigale Stangl, Shasta Ihorn, Yue-Ting Siu, Aditya Bodi, Mar Castanon, Lothar D Narins, and Ilmi Yoon. The potential of a visual dialogue agent in a tandem automated audio description system for videos. In *Proceedings of the International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS, 2023.

[62] Hariharan Subramonyam, Roy Pea, Christopher Lawrence Pondoc, Maneesh Agrawala, and Colleen Seifert. Bridging the gulf of envisioning: Cognitive design challenges in llm interfaces, 2024.

[63] Synthesia. Video statistics in 2023. `https://www.synthesia.io/post/video-statistics#:~:text=Video%20made%20up%2082%25%20of,10%25%20through%20text`, 2023.

[64] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. Making short-form videos accessible with hierarchical video summaries. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI, 2024.

[65] Lakshmie Narayan Viswanathan, Troy McDaniel, Sreekar Krishna, and Sethuraman Panchanathan. Haptics in audio described movies. In *IEEE International Symposium on Haptic Audio Visual Environments and Games (HAVE)*, pages 1–2, 2010.

[66] 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 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 2021.

[67] Lindsay Kolowich Wiegand. How video consumption is changing in 2023 [new research]. `https://blog.hubspot.com/marketing/how-video-consumption-is-changing`, 2023.

[68] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, CHI, page 143–146, 2011.

[69] World Wide Web Consortium (W3C). Understanding Success Criterion 1.2.3: Audio Description or Media Alternative (Prerecorded). `https://www.w3.org/TR/UNDERSTANDING-WCAG20/media-equiv-audio-desc-only.html`, 2008.

[70] Ilmi Yoon, Umang Mathur, Brenna Gibson, Tirumalashetty Pooyan Fazli, and Joshua Miele. Video accessibility for the visually impaired. In *International conference on machine learning AI for social good workshop*, volume 1, page 1, 2019.

[71] YouDescribe. Youdescribe: Video description for youtube. `https://youdescribe.org/`, 2024.

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

[73] Beste F Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A Miele. Increasing video accessibility for visually impaired users with human-in-the-loop machine learning. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–9, 2020.

[74] Lotus Zhang, Simon Sun, and Leah Findlater. Understanding digital content creation needs of blind and low vision people. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS, 2023.

APPENDIX A

PERMISSION FROM CO-AUTHORS

This is to certify that I, Maryam Saadat Cheema, have obtained permission from all co-authors to include the work presented in this thesis. The contributions of my co-authors, Hasti Seifi and Pooyan Fazli, have been duly acknowledged within the thesis, and their insights and feedback have been instrumental to the project's realization. Each co-author has been informed of their inclusion in this thesis and has expressed their consent for me to incorporate our collective efforts herein. This consent encompasses all aspects of the work as presented in the original document and any derivative use within this thesis.

APPENDIX B

ASU IRB DEFERRAL FOR THE USE OF HUMAN SUBJECTS

DEFERRAL

Pooyan Fazli
HIDA: Arts, Media and Engineering, School of (AME)
-
pooyan@asu.edu

Dear Pooyan Fazli:

On 10/15/2024 the ASU IRB reviewed the following protocol:

Type of Review:	IRB Site
Title:	From Human-Powered to Automated Video Description for Blind and Low Vision Users
Investigator:	Pooyan Fazli
IRB ID:	STUDY00018023
Funding:	Name: HHS: National Institutes of Health (NIH), Grant Office ID: FP00036161
Grant Title:	FP00036161;
Grant ID:	FP00036161;
Documents Reviewed:	• BLV Participant - Recruitment Email.pdf, Category: Recruitment Materials; • BLV participants - Interview questions and study user interface, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • BLV_consentform_video evaluation.pdf, Category: Consent Form; • consent_AMT-UCSC Format.pdf, Category: Consent Form; • consent_form.pdf, Category: Consent Form; • consent_templatesighted.pdf, Category: Consent Form; • Determination - 8947_Fazli_ From Human-Powered to Automated Video Description for Blind and Low Vision Users.pdf, Category: Other; • Form-Local-Context-Review_3.21.24.pdf, Category:

	Other; • IRB Approval, Category: Other; • MTurk Qualtrics Survey.pdf, Category: Measures (Survey questions/Interview questions /interview guides/focus group questions); • NIH Proposal, Category: Sponsor Attachment; • Research Protocol, Category: IRB Protocol;

The ASU IRB has deferred review and oversight of this project to UCSB IRB and the associated IRB protocol number is HS-FY2023-132.

Sincerely,

IRB Administrator
cc:
Pooyan Fazli
Maryam Cheema
Hasti Seifi
Chaoyu Li

APPENDIX C

UCSC IRB APPROVAL/EXEMPTION FOR THE USE OF HUMAN SUBJECTS

Gmail

HS-FY2023-132 - Renewal: Expedited - Approval (1 Yr Admin Check-In)

1 message

do-not-reply@cayuse.com <do-not-reply@cayuse.com> 15 October 2024 at 14:29
To: skurnia@ucsc.edu
Cc: pooyan@asu.edu

Office of Research

Study #: HS-FY2023-132
Study Title: From Human-Powered to Automated Video Description for Blind and Low Vision Users
Principal Investigator: Sri Kurniawan
ASU Site PI: Pooyan Fazli
Determination: Approved
Determination Date: October 15, 2024
Valid through: October 14, 2025

Dear Sri Kurniawan:

The request to renew the above-referenced human subjects research study was reviewed by the UC Santa Cruz Institutional Review Board (IRB) Chair and/or IRB Chair's designee(s) and approved on October 15, 2024. The study renewal was determined to meet the criteria for expedited review under:

- Category 6 - Collection of data from voice, video, digital, or image recordings made for research purposes.
- Category 7 - Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

In accordance with this approval, the specific conditions the IRB Chair/Chair's designee approved for the conduct of this research are listed below.

- The study meets the criteria for approval as outlined in §45 CFR 46.111.
- The consent procedures provide subjects with the required elements of informed consent as outlined in §45 CFR 46.116.
- The subject's informed consent is documented for sighted subjects in accordance with §45 CFR 46.117.
- The study meets the criteria to waive the requirement to obtain a signed consent form for Amazon Mechanical Turk (MTurk) subjects and blind or low vision (BLV) subjects in accordance with §45 CFR 46.117.

This approval is based on the degree of risk, and will require an Administrative Check-In on October 14, 2025.

This approval is limited to the activities described in the approved Cayuse Human Ethics study, and extends to the performance of these activities at each respective site identified in the study. The research must be conducted in accordance with the IRB-approved study. If applicable, informed consent must be obtained and documented using only the current IRB-approved stamped documents.

This approval does not constitute funding or other institutional required approvals. Should this study involve other review committees, such as the Environmental Health & Safety (EH&S) Institutional Biosafety Committee (IBC) and/or the EH&S RSSC (Radiation Safety and Surveillance Committee), it is the Principal Investigator's responsibility to ensure that all required approvals are in place prior to conducting research involving human subjects or their related specimens.

Please note the following Principal Investigator responsibilities:

- It is the Principal Investigator's responsibility to submit a Renewal Submission in Cayuse Human Ethics, at least one month prior to the administrative check-in date (October 14, 2025), in order to continue conducting study

activities beyond that date. Please note, failure to submit an administrative check-in may result in a processing hold being placed on all submissions from the Principal Investigator.

- It is the Principal Investigator's responsibility to submit any proposed changes to study activities, via a Modification Submission in Cayuse Human Ethics, to the IRB for review and approval prior to being implemented.
- It is the Principal Investigator's responsibility to promptly notify the IRB of any reportable events (adverse events/protocol deviations/unanticipated problems/subject complaints/other) that occur during the research, including any breach in confidentiality or data security that places participants or others at a greater risk of harm. Notify the IRB via an Incident Submission in Cayuse Human Ethics.
- Cayuse Human Ethics studies must be closed when all activities involving human subjects are completed, including interaction/intervention with participants or analysis of identifiable data. If the principal investigator leaves the University prior to expiration of the study, the study must be closed or transferred to another eligible UC Santa Cruz PI. Student-led studies must be closed before graduation. Closure of student-led studies which remain open after graduation are the responsibility of the faculty advisor.

The UC Santa Cruz Institutional Review Board operates under a Federalwide Assurance approved by the DHHS Office for Human Research Protections, FWA00002797. Our DHHS IRB Registration Number is IRB00000266.

Please note that the IRB has the prerogative and authority to ask further questions, seek additional information, require further modifications, or monitor the conduct of research and the consent process, if applicable. We wish you the best as you conduct your research.

Should you have any questions, please do not hesitate to contact us.

Best regards,

Office of Research Compliance Administration (ORCA)
University of California, Santa Cruz
orca@ucsc.edu | UCSC ORCA