Co-Designing Systems to Support Blind and Low Vision Audio Description Scriptwriters

Lucy Jiang

Supervised by Dr. Richard Ladner

A senior thesis submitted in partial fulfillment of
the requirements for the degree of

Bachelor of Science
With Departmental Honors

Paul G. Allen School of Computer Science & Engineering
University of Washington
June 10, 2022

Presentation of work given on June 13, 2022

Thesis and presentation approved by ________________________________

Date _____________________

Abstract

Audio description (AD), an additional narration track that conveys essential visual information in a media work, is imperative for improving video accessibility for people who identify as blind or low vision (BLV). The AD creation process includes three primary steps: writing the script, recording the voiceover, and mixing the narration track with existing video audio. Despite being the primary beneficiaries of AD, BLV individuals are limited in how they can contribute to the AD writing process due to technology inaccessibility and societal biases. The BLV community and sighted allies advocate for including BLV individuals in the AD creation pipeline, as their expert end-user perspectives lead to high quality descriptions. In this thesis, we (1) design, prototype, and test AccessibleAD, a system to make writing audio description more accessible, (2) surface qualitative insights on audio description accessibility from interviews with AD audiences and writers, and (3) analyze what visual context and features are desired or necessary for BLV writers when writing AD. From user studies with six members of the BLV community, we uncover accessibility challenges with current AD writing systems and find that BLV AD writers seek to have a detailed objective understanding of character identities, background settings, and actions while crafting engaging AD. This thesis expands on existing literature regarding co-designing systems with BLV end users and builds a foundation for understanding how to expand BLV engagement with audio description writing.

Contents

1 Introduction

1.1 Background and Motivation

For over 300 million blind and low vision (BLV) people across the world [5], audio descriptions are a critical extension to the storytelling occurring on screen. Audio description (AD), as defined by the American Council of the Blind, is "the descriptive narration of key visual elements of live theater, television, movies, and other media to enhance their enjoyment by consumers who are blind or have low vision" [16]. As the quantity of digital video content produced grows each year, the need for audio described videos grows rapidly alongside it [14].

Each portion of the AD pipeline, from writing to production, is essential for conveying human emotion and artistry in entertainment. Audio description companies now consult a handful of BLV audience members when writing AD [20], but there is still a major difference between serving as a consultant and being a scriptwriter. For example, the priorities of audio description script writers may not always align with the needs of BLV audiences, who wish to hear more about characters' races, costumes, and disabilities [27]. Furthermore, disabled people are frustratingly subjected to sanitized entertainment experiences by sighted writers, which often takes the form of being offered PG-rated descriptions for R-rated scenes [28]. Despite the proliferation of audio description across most entertainment platforms, BLV audiences cite the lack of description quality and consistency as a primary grievance [26].

There are many stigmas surrounding involving BLV people in the audio description creation process. BLV and sighted audiences alike hold beliefs that BLV creatives cannot create high quality descriptions and that their "writing is of lower quality because [they] have to synthesize information in a different way" [11], as evidenced through a lack of employment opportunities for BLV writers [19]. There are very few blind and low vision audio description writers employed in the AD industry today. Ren Leach, a blind voiceover talent, confronted this issue by asking: *"If #AudioDescription is intended for #Blind #LowVision people, why is it that the AD creators are not recruiting from the blind community and why are their methods so entrenched in visually based creative processes? I am not a fan of empty promises of inclusion"* [12].

As the primary consumers of AD, blind and low vision individuals must have the same opportunities as sighted people to participate in the creative process of creating audio descriptions. The reasons for this are twofold:

1. BLV users deserve to have fair and equal access to employment opportunities, and
2. AD created or informed by blind and low vision perspectives is of a higher quality and accounts for the needs of BLV audiences.

To understand the perspectives of BLV people on audio descriptions, we conducted semi-structured interviews with eight blind and low vision audio description users. During the interviews, we also asked participants to audio describe two short videos using a prototyped system with pre-written descriptions and video navigation controls. They were permitted to ask quantifiable and yes / no questions regarding the video visuals in a simulation of an automated visual question answering system (VQA), which better illuminated the information that BLV audiences believe is important for having a fuller understanding of a clip. Upon completion of the AD task, participants were asked to rate their satisfaction with the descriptions they had written, as well as with the prototyped system and the overall experience of writing audio description as a blind or low vision individual.

By proposing, designing, and evaluating voice-based query interactions within the context of creating audio descriptions, this work explores how blind and low vision audio description writers can independently participate in the AD creation process. Advancements in knowledge about BLV interactions with VQA can also provide insight into future implementations of this technology, such as in personal voice assistants. Exploring technological solutions in an unconsidered space for increasing BLV involvement can greatly impact future efforts for the necessary inclusion of disabled users.

1.2 Research Questions

Through this study, we work towards understanding how to provide equal access to AD creation and to combat stigmas against blind and low vision AD writers. The primary questions we seek to answer are as follows:

1. What context is necessary or desired for blind and low vision writers to independently write artistic audio descriptions?
2. What features are the most accessible, efficient, and effective for blind and low vision AD writers?
3. How can more blind or low vision creatives become involved in audio description production pipelines?

1.3 Contributions

This research expands on previous work in making audio description more widely accessible, with a specific focus on the creation and curation of accessible videos. This thesis is the first to explore the usage of simulated VQA systems to make the AD writing process more accessible for BLV writers, and is also one of the first to evaluate the impact of technology on audio description writing workflows. It is critical to co-design with BLV writers and BLV audiences to develop the most accessible audio description writing experience for blind and low vision people.

The technical contribution of this thesis is AccessibleAD, a prototype of a system that enhances BLV access to AD writing. Using this system, blind or low vision AD writers can either work with visual question answering technology or with sighted individuals to ask questions and receive additional details about visual scenes. Lastly, we provide recommendations and design considerations for future accessible AD writing systems and address the stigma facing BLV audio description writers, informed by interviews and usability studies with BLV audiences and writers of audio description.

2 Related Work

From the first conceptualization of audio description in the 1970s, to WGBH's creation of Descriptive Video Services in the 1980s, the audio description industry has been on a steady incline for the last half century [14]. In 2015, Netflix released Daredevil, a show with a blind superhero protagonist, without any audio description for BLV viewers [21]. This issue was faced with extreme scrutiny, and was rectified within days of the initial release. Over the last seven years, Netflix alone has described over 1100 titles [17], and many other streaming services have begun offering AD as well. In tandem, the interdisciplinary research area of audio description and accessibility has proliferated and grown significantly.

My research builds extensively on literature on audio description writing and visual question answering. Existing literature regarding AD examines user experiences with audio description, the automation of creating audio description, and tools with which sighted writers can create AD. However, they do not evaluate the ways in which BLV writers interact with audio description writing systems; as such, this work is the first investigation into providing equal access to AD creation by designing and creating tools to support BLV audio description writers. By synthesizing BLV perspectives and needs regarding audio description writing, this paper identifies ways in which emerging technology and innovative user interactions can support societal change and the inclusion of the blind and low vision community.

2.1 Audio Description Writing and Preferences

Although audio description is not a new concept, it is an emerging research area within the subfields of HCI and accessibility. Prior work has attempted to improve audio description quality and quantity through multiple avenues, including tools and interfaces to help sighted AD writers [18]. Another system, CineAD [3], automatically generates AD for movies by leveraging data about speech gaps and video content based on the original script and subtitles. Through VerbalEyes [10], a system that generates AD via automated keyframe detection and description, Jiang et al. found that BLV audiences have varied preferences for audio description brevity, voices, and audio mixing, and that automated audio description is serviceable, but not ideal, for BLV audiences.

Regarding the importance of including humans in the process of creating quality audio description tracks, Yuksel et al. [29] reported that a human-in-the-loop machine learning approach was effective at reducing barriers for creating AD. Similarly, ViScene, a collaborative audio description writing tool that enables BLV or sighted script reviewers to contribute to creating non-professional audio descriptions, was shown to decrease costs for creating AD [15]. However, these works are all conducted in a framework that identifies sighted humans as the primary writers of audio description scripts, focusing solely on the benefits of partial automation of AD creation for sighted writers – they do not fully explore how to engage BLV writers in the AD writing process itself.

2.2 Visual Question Answering for Audio Description

More recent work has applied state-of-the-art visual question answering systems to augment BLV audio description experiences. Using two AI-driven tools, NarrationBot and InfoBot, Ihorn et al. [7] examined the impact of combining baseline and on-demand descriptions (through VQA) on BLV audiences' video watching experiences. Their usability studies with 26 BLV people showed that a combination of the two tools was most effective for enhancing their understanding, enjoyment, and agency, but any permutation of the tools was still helpful to end users. This prior work is one of the first to integrate VQA into audio description context. However, this paper does not approach it from an AD writing perspective; my thesis expands on these ideas of providing baseline and on-demand descriptions to examine how to improve the experience of BLV audio description writers. This work specifically focuses on prototyping an accessible interface to facilitate audio description writing and understanding the context desired or needed by blind and low vision AD writers.

3 Study Design

3.1 Preliminary Investigation

We began our background research by conducting a literature review of prior audio description writing system research. While existing AD-related research begins to scratch the surface of the ad-hoc description needs of BLV audiences when watching videos without AD, there is little research into how BLV individuals undergo the process of writing audio descriptions.

After understanding the AD research landscape, we interviewed current BLV audio description writers to better understand their workflows and current processes for writing AD. We conducted interviews with two BLV audio description writers who have contributed to industry-level audio description scripts. From the preliminary interviews, we identified that current AD writing systems are inaccessible and that BLV writers seek to have a detailed objective understanding of characters, background settings, and cross-frame actions prior to crafting artistic descriptions.

3.2 Design Goals

Based on our preliminary research, we derived a set of functionality design goals for a system for blind and low vision audio describers. This system must:

- Be accessible to blind and low vision users. A lack of accessibility in the prototyped system must not hamper participants' audio description writing experience. Blind and low vision individuals have the right to use a system that will be accessible based on their preferred input modalities.
- Support navigation throughout a video. Participants must have a simple and accessible method of navigating through a video to facilitate timecode retrieval and targeted video rewatching.
- Provide context regarding a video's visuals. As the audio description writing process is grounded by semi-objective visual observations, participants must have a means of gathering and understanding the visual aspects of a video.

3.3 System Design and Implementation

3.3.1 Prototype

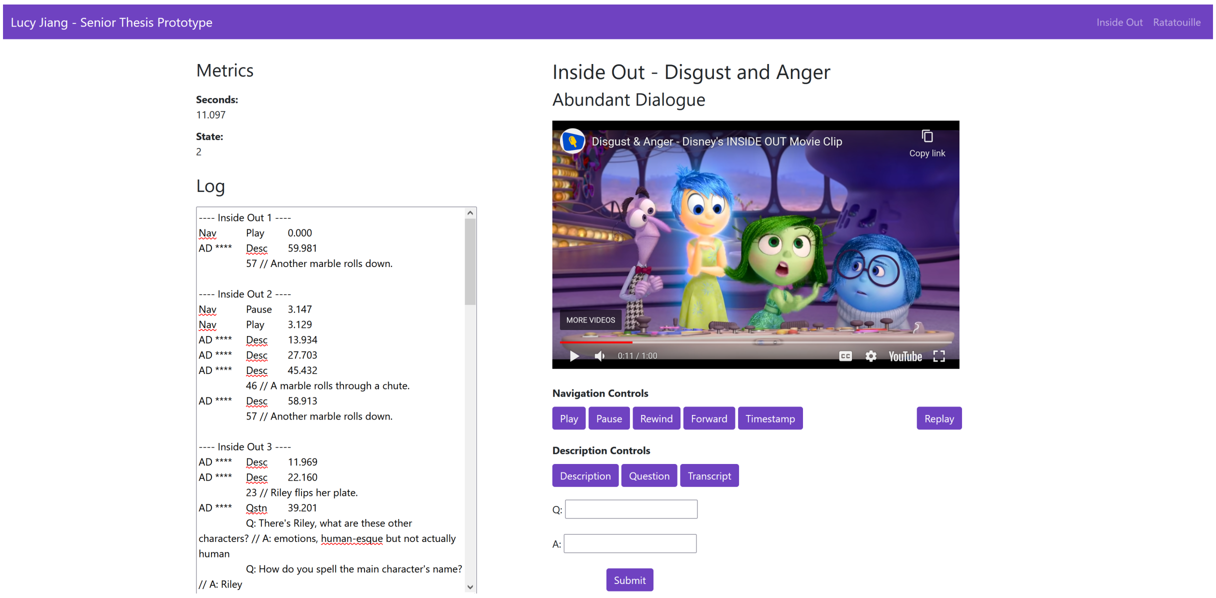
To facilitate the co-design process, we developed a prototype to test with end users. This system, named AccessibleAD, is a web-based platform with features designed to streamline the audio description writing process. During semi-structured interviews and usability tests, we evaluated three key features, which are as follows:

- A written transcript,
- Baseline descriptions (main objects, people, spatial relations and interactions between them, actions and movement, on-screen text, settings, etc.), and
- On-demand descriptions (accessed by asking quantifiable or yes / no questions at any point in the video).

Current machine learning (ML) and VQA technology cannot yet support a quality AD writing experience. As such, all of the features for the prototype were created via a Wizard of Oz method (written by a human but read via a synthesized voice) to better understand users' perspectives on features and workflows before creating a fully automated system.

Due to the limitations of conducting and observing virtual user studies with the prototype, participants' primary mode of interaction was through voice commands. Users who wished to use keyboard shortcuts were able to send keyboard commands through the "chat" feature native to the meeting platform.

All code for the prototype can be found on GitHub[1].


Lucy Jiang - Senior Thesis Prototype

Inside Out Ratatouille

Metrics

Seconds:
11.097

State:
2

Log

```
---- Inside Out 1 ----
Nav        Play      0.000
AD ****    Desc      59.981
           57 // Another marble rolls down.

---- Inside Out 2 ----
Nav        Pause     3.147
Nav        Play      3.129
AD ****    Desc      13.934
AD ****    Desc      27.703
AD ****    Desc      45.432
           46 // A marble rolls through a chute.
AD ****    Desc      58.913
           57 // Another marble rolls down.

---- Inside Out 3 ----
AD ****    Desc      11.969
AD ****    Desc      22.160
           23 // Riley flips her plate.
AD ****    Qstn      39.201
           Q: There's Riley, what are these other
characters? // A: emotions, human-esque but not actually
human
           Q: How do you spell the main character's name?
// A: Riley
```

Inside Out - Disgust and Anger

Abundant Dialogue

Navigation Controls

Description Controls

Figure 1: Screenshot of the AccessibleAD prototype. Features available to describers include navigation controls (play, pause, rewind, forward, timestamp, and replay) and description controls (description, question, transcript). The log panel on the left side of the screen tracked all user actions during the user study to facilitate later data analysis.

3.3.2 Chosen Videos

Participants were asked to audio describe two short clips from Disney Pixar movies, Inside Out - Disgust and Anger (abundant dialogue) and Ratatouille - Remy Fixes the Soup (minimal dialogue). The difference in dialogue density between the two videos allowed for two distinct AD writing experiences during the study. Brief descriptions of both video clips are included below.

Inside Out follows the journey of five personified emotions of a young pre-teen girl upon moving cities. From IMDb, the synopsis of the movie is as follows:

> *"After young Riley is uprooted from her Midwest life and moved to San Francisco, her emotions - Joy, Fear, Anger, Disgust and Sadness - conflict on how best to navigate a new city, house, and school."* [9]

[1] `https://github.com/lucjia/seniorThesisPrototype`

The one-minute clip [24] presented to interview participants was a flashback of Riley as a young child at dinnertime, during which her parents attempt to feed her broccoli (which becomes her least favorite food). Within her mind, a cavernous purple room with glowing marbles representing memories lining the wall, the five emotions debate on how to approach the unknown substance of broccoli. Disgust, a stylish green sprite with sparkling green hair, first recognizes the broccoli and gags, pressing a button on an emotion-controlling console that causes human Riley to flip her plate and send the broccoli flying. A green marble rolls through a chute to join the memory bank, indicating that a memory associated with disgust has just been created. When Riley's father states that Riley must eat dinner to receive dessert, Anger, a blocky red sprite, reacts explosively. He aggressively slams two levers on the console as flames shoot out of the top of his head, which translates to human Riley crying and throwing a tantrum and a red marble joining the memory bank. Riley calms down once her father places the broccoli onto a spoon and pretends as though he is flying an airplane into her mouth. All five sprites gather together in awe of the "airplane", and a golden (happy) marble rolls into the memory bank.

Ratatouille centers on a Parisian rat who loves to cook. From IMDb, the synopsis of the movie is as follows:

> *"A rat who can cook makes an unusual alliance with a young kitchen worker at a famous Paris restaurant."* [8]

The one-minute 25-second clip [23] presented to interview participants was an early scene in which Remy, the rat, attempts to fix a foul-smelling soup in a dimly-lit restaurant kitchen. He initially sniffs the soup, contemplates what additional ingredients it needs, and then tosses in a small handful of those ingredients. He heads to the window, his method of escape. However, he hesitates, enchanted by the idea of transforming the soup. Gusteau, a late chef and Remy's culinary idol, appears as a ghost to encourage him to take the chance. Resolutely, Remy heads back into the kitchen, and starts first by adjusting the flame underneath the soup pot. He then washes his hands with a droplet of water from the tap and pushes translucent broth and heavy cream into the pot. He joyously sniffs and gathers ingredients to add to the soup, ranging from what appear to be diced garlic, sliced leeks, cubed potatoes, and dried thyme sprigs. Remy stirs and sniffs the soup, feeling satisfied about his handiwork, just before the lights in the kitchen turn on to reveal a shocked young human chef, Linguini. A senior chef storms into the kitchen asking for the soup, and Linguini traps Remy underneath a colander.

3.3.3 Pre-Generated Descriptions

After reviewing each clip, we wrote baseline descriptions with minimal detail, maintaining objectivity as much as possible. The baseline descriptions primarily focused on general actions, and did not include information about characters' races or ages. These descriptions were recorded with IBM's Watson Text to Speech engine to simulate the experience of receiving computer-generated descriptions. In keeping with industry audio description conventions, we intentionally only wrote descriptions for areas in the video that did not contain dialogue. Due to fairly dense dialogue, the Inside Out video contained four pre-written descriptions, as compared to the five descriptions in the officially released AD on Disney+. On the other hand, the sparse dialogue in the Ratatouille video resulted in ten pre-generated descriptions, as compared to the 19 descriptions in the officially released AD on Disney+. Baseline descriptions for Inside Out and Ratatouille are shown in Tables 1 and 2, respectively.

Timestamp	Description
1	Riley sits in a high chair.
23	Riley flips her plate.
46	A marble rolls through a chute.
57	Another marble rolls down.

Table 1: Pre-generated descriptions for Inside Out, including timestamps.

Timestamp	Description
1	Remy sniffs the soup.
5	He tosses ingredients into the pot.
10	He looks back at the pot.
28	He reduces the flame under the pot.
35	Remy washes his hands.
40	Broth and cream are poured into the soup pot.
49	He tosses more ingredients into the pot and smells the soup again.
64	He stirs herbs into the soup.
72	The lights turn on and a chef stares at him.
80	The chef traps Remy.

Table 2: Pre-generated descriptions for Ratatouille, including timestamps.

4 Study Methodology

To explore and understand the needs of blind and low vision audio description writers, we conducted semi-structured interviews and usability studies with each participant in approximately one-hour long sessions, all of which were conducted remotely via Zoom.

4.1 Participants

Eight blind and low vision people with an interest in writing AD, including two people with prior AD writing experience who identify as BLV audio description writers, participated in this study. We recruited participants primarily through the newly established Audio Description community on Twitter and snowball sampling. Participants were screened on the following criteria: they must (1) identify as blind or low vision, (2) have watched audio described videos, and (3) be over 18 years of age. All participants were volunteers and were compensated for their time with a $30 digital gift card. Two interviewed participants did not ask questions during the audio description task or did not write any audio descriptions with timestamps. These interviewees were excluded from the data analysis, resulting in a total of six blind and low vision participants for this study. Participant IDs, visual acuity, and their prior AD writing experience (in participants' own words) are shared in Table 3.

ID	Visual Acuity	Description
1	Fully blind	None
2	Fully blind	Writes AD by co-watching and asking questions with a sighted person
3	Low vision	None
4	Totally blind	Not much; bits and pieces for work
5	Low partial vision	Has written AD for documentaries
6	Low vision	None

Table 3: Participant IDs, visual acuity, and prior AD writing experience.

4.2 Procedure

The study consisted of three primary parts. First, we asked participants of their opinions and preferences related to any audio description they have consumed in the past, as well as questions about their audio description writing experience. All of the questions in the first part were intended to uncover areas in which current professionally produced audio description is insufficient and identify information that BLV audiences consistently seek in audio described videos. Specific question probes included:

- What is your experience like with audio description? What do you like about it and what do you wish was better?
- What information is important to you in an audio description script?
- What do you think would be helpful for you to have access to when writing audio descriptions?

Second, we shared our screen and asked participants to audio describe two videos using AccessibleAD, which provided BLV users with baseline descriptions and the opportunity to ask quantifiable or yes / no questions at any point during the video. To introduce the task, we explained how participants could interact with the system using voice commands, explained which navigation and description controls were available, and clarified the way in which baseline descriptions could be accessed. Participants were given 15 minutes to describe each video, and it was made clear that they did not need to finish writing their descriptions, nor fix typos or grammar, within this timeframe.

The first video that participants described was a one-minute long clip from the 2015 Disney Pixar movie Inside Out and the latter was a one-minute and 25-second clip from the 2007 Disney Pixar movie Ratatouille. For participants that were unfamiliar with the clips, defined as having given a rating of 5 / 10 or lower when asked how familiar they were with this movie, we provided them with a synopsis from IMDb to give brief context into the clip and the characters present. We observed participants' interactions and experiences with the prototype when describing these videos, and logged all of their actions and questions for further analysis.

Third, we asked follow-up questions about participants' thoughts on the audio description process and how to improve their overall experience with audio description. The interview closed with participants sharing their thoughts on their reactions to blind and low vision people serving as audio description writers, including questions such as *"Are you interested in writing audio descriptions? How has this changed since before engaging in the audio description task?"* and *"What do you think would be helpful for increasing involvement of BLV writers and creatives in audio description production pipelines?"*. Participants also discussed ways that they believed would be helpful to combat current stigmas against BLV people.

4.3 Data Analysis

For the data analysis, we drew on principles of grounded theory to perform thematic analysis on the interview transcripts. We placed a focus on understanding the attributes of audio description that they believed to be important and their experiences with writing audio description with the prototyped system. In identifying themes related to how participants' approaches to audio description, we aggregated the questions that participants asked during the description task by question type (clarifying a sound, quantitative, yes or no questions, etc.) and question content (asking about character identity, background noises, etc.). Lastly, we analyzed quantitative data regarding participant satisfaction with audio description writing to understand the serviceability of the system overall.

5 Study Findings

Observing users' interactions and experiences with AccessibleAD when describing videos identified unspoken needs of BLV audio description writers, and analysis of these interview results aided in iterating on a design for a platform that better includes BLV writers in industry AD pipelines. From participant insights, we synthesize and present key needs of BLV audio description writers, which are primarily in regards to obtaining additional context about the video's visuals and having accessible interactions with the system.

5.1 Context Required for Audio Description Writing

Four of the six study participants sought more detail about specific elements in each video to provide context for their description writing. Through analyzing their questions and reactions to the pre-generated audio description tracks, we identified four major components of context that are necessary for and / or desired by blind and low vision writers.

5.1.1 Character Descriptions

In alignment with the commonly expressed desire for more description of characters' physical appearances, half of the participants (N = 3) asked about the identities of the characters on screen. Questions about character attributes included inquiries about a character's race, age, or even their expressions and body language. For example, P1 emphasized the necessity of describing race for giving insight into a character's actions and the way they may be treated by others. He asked about the races of the characters, which reflected in their final description script.

> *"00: Riley, a Caucasian baby sits in a high chare for dinner with her parents and 5 humonoid charectors representing the emotions of disgust, anger, joy, sadness and fear" [sic]* (P1)

Despite being relatively unfamiliar with the film, P5's questions about color usage in Inside Out revealed how color is utilized to indicate which emotion is being expressed.

> *Q: What does disgust look like? // A: Green sprite, green skin, green hair, long eyelashes*

> *"00:00:22 Green Disgust stands before the others.*
> *00:00:46 Red Anger presses down a lever and a red marvel rolls down a shoot." [sic]* (P5)

5.1.2 Action Descriptions

In Ratatouille, questions about character descriptions were not as abundant as in Inside Out, which can likely be attributed to the main character of the clip being a rat rather than a person. Instead, a majority of participants (N = 4) asked for further details regarding the character's actions. Some questions were intended to fill in gaps in the pre-generated AD, and some participants (N = 2) asked about actions that they missed during a lengthy description gap between 49 and 64 seconds. Other questions sought to gather more context about the actions the pre-generated AD had begun to detail.

While the pre-generated audio description stated, “He tosses ingredients into the pot” at 5 seconds and “He tosses more ingredients into the pot and smells the soup again” at 49 seconds, three interviewees expressed curiosity about the types and quantity of ingredients that were described. For example, after hearing this description, P2 rewound the video to 46 seconds and proceeded to ask about the specific ingredients that Remy gathered. These details were reflected directly in P2’s resulting script.

> *“42: Reemy pours leaks, garlic potatoes, into the pot.” [sic]* (P2)

When describing Inside Out, P2 also asked about the speed and direction of the marble rolling down the chute to clarify their own understanding of the video based on the pre-generated descriptions. However, they did not include this in their final script.

Lastly, one participant inquired about how or why the character was doing the action described. In P5’s descriptions for Inside Out, as a follow up to their questions about how to describe Disgust, they asked, *“What is her expression and what is she doing with her body?”*. These questions about Disgust’s demeanor enabled P5 to show, rather than tell, the audience more about the character’s personality.

5.1.3 Settings

In addition to valuing foreground details, such as characters’ identities and their actions, participants also asked about the background settings of each scene. Two participants mentioned that having information about the scene background or location was important to them in an audio description script. After one pass of the pre-generated audio descriptions for Ratatouille, which ends with a mention of the lights turning on, both P3 and P1 immediately inquired about the setting of the scene. These details were included in the first line of their final description script, indicating their opinion on the importance of this information for establishing context for viewers.

> *Q: The lights are off for the whole scene? // A: Dimly lit*
> *Q: Commercial kitchen? Professional restaurant? // A: Yes*

> *“00: A rat, Remi is cooking over a stove in a dimly lit resteront kitchin” [sic]* (P1)

More participants asked about the background settings for the Ratatouille clip than the Inside Out clip, but P3 asked a question about the location in which the characters were convening in Inside Out as well.

> *Q: Which place is this? // A: Emotions are in her mind, purple* (P3)

5.1.4 Clarifying Sound Effects

Interview participants also built on the audio cues present in the video, using sound effects and the tones of the dialogue to understand the full meaning of a scene. In the Inside Out clip, notable sound effects included the clinking of marbles and Riley’s temper tantrum. Both of these sounds were alluded to in the pre-generated descriptions as “marbles rolling down a chute" and "Riley flipping a plate", respectively. However, the sound alone was not enough to capture who was completing this action, the purpose of each visual element, or exactly what was happening. For example, during the first play of the video, P2 heard a clinging sound and rewound the video to clarify what it was.

> *Q: The cling I heard at 20 seconds was silverware hitting a plate? // A: Yes*

Additionally, when P5 was writing descriptions for Inside Out, they asked about audio cues from varying character voices and sound effects. The sound of Riley crying was misinterpreted by P5 to be the sound of a screeching cat, and their question helped them clarify their understanding of the actions and emotions of each character.

> *Q: When he says right after you eat this, is he throwing a cat? // A: Anger is red and slamming a lever on a console, fire exploding out of his head, Riley starts crying*

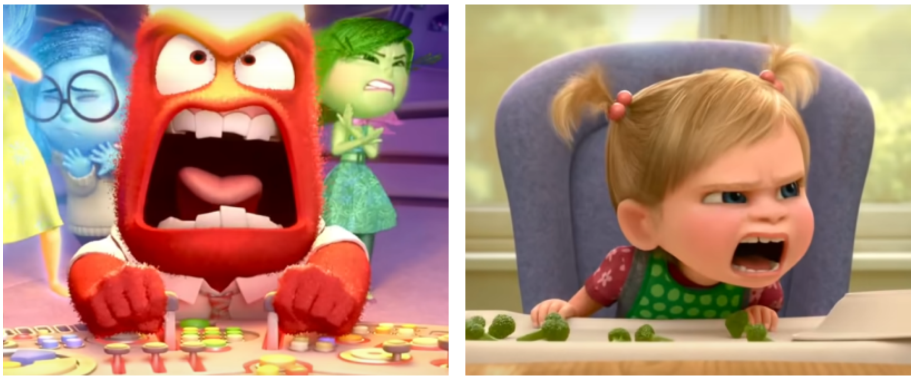


Figure 2: Left: Anger, a blocky red emotion sprite, slams levers on the console as flames erupt from his head. Right: Riley, the toddler, cries after almost being fed broccoli. These two frames are shown in immediate succession in the clip, but the video's audio cues were not fully clear to BLV writers.

5.1.5 Missed Context

The official audio described versions of both clips, available on the Disney+ streaming service, provided slightly more detail in their AD than participants were able to create during the study. This is due greatly to the involvement of trained professional audio description writers and the lack of a time limit for writing the descriptions. The official descriptions for Inside Out [25] were fairly sparse as a result of the dense dialogue, and BLV study participants largely captured the same information in their written descriptions.

For Ratatouille, a clip with significantly less dialogue and therefore more time and space for audio description [22], there were more differences between the official AD and the details noted by BLV participants. The vast majority of missed details were descriptors of minor actions. For example, the soup simmering on the stove was described to be a *"bubbling soup"*, and Remy was characterized as taking a *"satisfied whiff"* after he finished fixing the soup. Additionally, the official AD track's description of Remy as he made the decision to fix the soup indicated one key area that was not queried about by BLV participants. The description stated: *"Remy's ears flatten back against his head as he gazes with* *a determined stare"*, which demonstrates the rat's character and resolve. However, none of the participants asked about nor included this non-audible detail. Although this detail is not critical to understanding the overall meaning of the scene, this prompts questions about ways to surface subtle details to BLV writers if they lack additional audio cues.

5.2 Accessible Features

Due to the inaccessibility of AD writing interfaces, BLV writers often do not have the support that they need to contribute meaningfully to this process, contributing greatly to negative stigmas surrounding BLV audio description writers. In this section, we identify which features are accessible, efficient, and effective for blind and low vision AD writers.

Pre-generated baseline audio description tracks are effective ways to provide blind and low vision AD writers with additional context about the visuals of a video. During the audio description task, most interviewees (N = 5) requested to listen to all of the pre-generated audio descriptions to build an understanding of the video's context. Four of the five participants who requested the full description track to be played preferred for the descriptions to be played within their first or second pass, while the last participant began listening to all descriptions on their fifth pass of the video. The only participant who elected to not listen to the entire set of pre-generated audio descriptions reported that they were already familiar with both clips, giving Inside Out a rating of 8 / 10 and Ratatouille a rating of 6 / 10.

However, pre-generated AD is not enough to provide full information access. When describing what information was important to her in an AD script, P5 stated: *"I just want full access to what someone... just because they have a functioning pair of eyeballs, has access to."* VQA support must also be integrated into AD writing systems to build on the context afforded by audio cues and baseline descriptions, as this allows BLV writers to clarify additional uncertainties and write descriptions based on information that they believe is critical to their and the audience's understanding.

Another pain point that was expressed by a current audio description writer related to the formatting of the final script, a logistical portion of the audio description writing process that is rarely explored. The current process of formatting a script for narration is inaccessible and frustrating to screen reader users, and could be streamlined through automatically attaching timestamps to description tracks when they are written.

Lastly, many interview participants gave feedback regarding their usage and interactions with the prototype. While two participants specifically liked the voice control interface, two participants expressed that they would have appreciated additional input methods, including keyboard shortcuts. Having multiple input methods and control mechanisms enables participants to choose whichever method of use is most natural for them, which can greatly boost efficiency and accessibility in the long run.

5.3 Quantitative Feedback

5.3.1 Description Satisfaction

During the interview, participants were asked to rate their satisfaction with their descriptions for Inside Out and Ratatouille. They gave a rating of 5.42 / 10 for Inside Out on average, with three interviewees rating their satisfaction at 7 / 10. Two participants cited the abundant dialogue of the Inside Out clip as a primary reason why they did not write as many descriptions as they thought would be helpful for full understanding. Additionally, as Inside Out was presented first, the moderate learning curve of the system could have interfered with participants' writing efforts and reflected in their satisfaction ratings.

Participants rated their satisfaction with their Ratatouille descriptions at 6.92 / 10 on average, with all participants giving their Ratatouille descriptions the same rating or higher than their Inside Out descriptions. Half of the participants acknowledged that the minimal dialogue in Ratatouille afforded them more space and flexibility to write descriptions, but some still wished for greater detail and accuracy in their descriptions.

Table 4 lists the full set of participant satisfaction ratings regarding their written AD.

	P1	P2	P3	P4	P5	P6	Average
Inside Out	3.5	7	7	4	7	4	5.42
Ratatouille	7.5	7	9	4	8	6	6.92
Average	5.5	7	8	4	7.5	5	

Table 4: Participant ratings of their satisfaction for the descriptions that they wrote for Inside Out and Ratatouille. P1 shared approximate ratings, "3 to 4" and "7 to 8", which are listed in the table as 3.5 and 7.5.

5.3.2 Overall Satisfaction

When asked to rate their satisfaction with the system overall on a scale from 1-10, participants gave a rating of 6.58 / 10 on average. Two interviewees found the task to be difficult given the time constraints due to their lack of prior experience with writing AD, but a majority of participants liked the overall system design and its intuitive features. P2, a current audio description writer, noted that they took some time to get used to the system interface. However, they also acknowledged applications of this system towards their current AD work: *"after I learned how the software works and everything, I love it"*.

Table 5 lists the full set of participant satisfaction ratings regarding the audio description writing experience for Inside Out and Ratatouille.

	P1	P2	P3	P4	P5	P6	Average
Satisfaction	5.5	9	7	8	5	5	6.58

Table 5: Participant ratings of their satisfaction with the system and overall writing experience. P1 shared an approximate rating of "5 to 6", which is listed in the table as 5.5.

6 Discussion

This work presents the first systematic investigation of BLV writers' needs in writing audio description and is the first to codesign audio description writing platforms with BLV users. In the following sections, we discuss the technical requirements for implementing automated AD writing systems, potential harms and technology misuse, and the implications of developing technology for blind and low vision audio description writers.

6.1 Design Considerations

6.1.1 Visual Question Answering Systems

The tasks of being able to generate video descriptions and answers to visual questions are nontrivial, and both are actively expanding research areas. Regarding design considerations for VQA systems in particular, it is important to utilize VQA datasets that contain relevant image and question training data. VizWiz-VQA [6] is a dataset composed of pictures taken by BLV users, questions that BLV users had about the pictures, and answers sourced by sighted users. This serves as an example of an existing dataset that can be directly applied to automating the answering of queries by BLV audio description writers, as the dataset is tailored to the types of questions that BLV individuals typically ask.

Furthermore, the VizWiz-VQA-Grounding dataset [4] can be helpful in AD contexts as well. Grounding refers to locating and returning the area within an image that is used to answer the question. As found through the user study, BLV audio description writers seek additional context about character descriptions, action descriptions, and settings. These aspects of context can be conveyed through direct visuals, but can also be implied through the framing of the video or can be enhanced by knowing the spatial relations of characters or actions. For example, when describing Inside Out, P2's questions about the direction and speed of the marble could be more easily answered by training a model on the VizWiz-VQA-Grounding dataset.

Advancing research in machine learning, computer vision, and deep learning can greatly improve outcomes of automated audio description generation, but it is unlikely that removing humans from the loop altogether will lead to high quality AD. As such, the need for BLV involvement in audio description writing persists. For developing platforms to make writing AD accessible to BLV writers, it is critical that the VQA systems in place are trained on datasets that also prioritize the needs of BLV users.

6.1.2 Potential Harms with Automated Audio Description

As with any technology, it is important to address ethical considerations related to the misuse or abuse of the technology. While creating pre-generated audio descriptions can be helpful for both BLV and sighted writers, automated audio descriptions may contain incorrect information that may not be easily detected by BLV writers or audiences. When ML is used for automatically generating baseline descriptions, we propose reporting a confidence level alongside the descriptions to signal their approximate accuracy.

Additionally, for studios which hope to increase the quantity but not the quality of the audio description that they produce, they may misuse AD technologies as "weapons for compliance" [7]. Studios, such as Amazon Prime Video, are notorious for using text-to-speech technology in lieu of human voice talents, a production decision that is criticized by BLV audiences for being jarring and unenjoyable for entertainment content [13]. Blind and low vision viewers state: *"audio description is too important to treat it like an afterthought. We need our content providers to treat AD as if it was as vital as the main audio track"* [1]. If current human processes are further replaced by automation, this may lead to thoroughly unusable final products even if these poorly created accommodations will enable studios to adhere to increasingly strict accessibility regulations.

6.2 Advancing the Audio Description Industry

Despite its recent and rapid growth, the audio description industry is still largely inaccessible. The "curb-cut effect" states that designing accessible technologies for disabled people can have universal benefits [2]. Pre-generated descriptions alone often do not provide enough context for BLV writers to create engaging and artistic descriptions. Visual question answering support, whether through a sighted assistant or state-of-the-art machine learning and computer vision technology, can greatly help with filling in gaps and giving writers more freedom and flexibility.

While sighted description writers may not expressly need pre-generated descriptions or visual question answering systems, these adaptations for understanding and accessing a video in an alternative way can be greatly beneficial for any describer to identify key visual elements or to break a video into more manageable segments. Introducing technology, such as VQA, into the audio description industry can be incredibly valuable as long as AD creators prioritize description quality and the needs of the BLV community.

6.3 Stigma

All participants expressed their staunch support for increasing the involvement of blind and low vision creatives in audio description production pipelines. Regarding the personal impacts of being able to write audio descriptions, P6 cited staying up to date as an important reason for BLV writers to be included, while P3 noted that he thought writing audio descriptions would be a valuable and positive way for him to contribute to his community. However, despite recognizing great value in technological augmentations of AD, P2 also remarked that changing the societal perception of BLV writers is just as critical – *"these tools are not a solution to each and every problem; they are just a tool"*.

Furthermore, participants also shared the broader impact of being involved in AD production processes as a blind or low vision person. As someone who does not watch videos without audio description anymore, P5 advocated for the increased agency of blind and low vision writers in description workflows, and expressed her frustration with existing audio description scripts. Despite the good intentions of sighted description writers, they mentioned how *"there's a real, big kind of historical problem where... [disabled peoples'] experiences need to be sanitized"* (P5), leaving BLV audiences with unequal information about gory or otherwise explicit scenes that sighted viewers had access to. While a shifting culture and recent amendments to already published descriptions, such as those for the Netflix show Bridgerton [28], have begun to remediate these issues of unequal access, it is critical for blind and low vision writers to helm these culture shifts and push for greater audio description quantity and quality.

P4, who has been using audio description since the early 1990s, shared a similar viewpoint as the other participants. He noted that he had previously only written descriptions for videos he was extremely familiar with, such as videos that he created himself. However, he was interested in broadening his audio description writing repertoire, and was encouraged by the straightforward prototype that provided greater access to visual content. He stated:

> *"Blind people can author audio description scripts. Adaptations are required, but it's no different than modifications which allow people with disabilities to accomplish all manner of tasks and jobs. There is room for blind people to fill these roles, and in fact, we should be filling these roles given that we know best what blind people want in AD."* (P4)

The disability rights rallying cry, "nothing about us without us," rings truer than ever as technology and digital video content becomes more ingrained into every aspect of everyday life. The stigma against blind and low vision audio description scriptwriters, held by BLV and sighted people alike, must be eliminated. Originally created by BLV people for BLV people, audio description has proliferated and grown to become a major industry. This industry must respect, empower, uplift, and employ blind and low vision creatives in the audio description creation process to achieve full parity, equality, and excellence.

6.4 Limitations

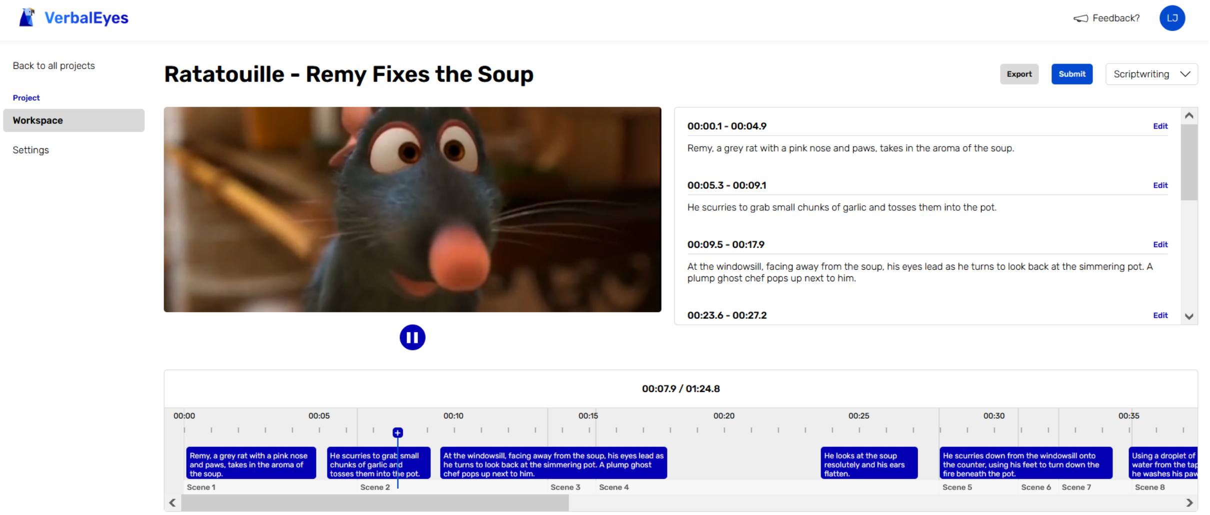
Our work in understanding blind and low vision writers' experiences with a prototyped audio description system has several limitations. Firstly, due to the virtual nature of the interview setup, as well as to facilitate data collection on questions and actions, we operated the prototype based on participants' vocalized commands instead of allowing participants to navigate the system on their own. As such, participants did have as much freedom or flexibility that they could have had when using the prototype on their own, which could have negatively impacted participants' satisfaction with the system.

Secondly, as the question and answer system was executed by a human rather than a computer in a Wizard of Oz fashion, the detailed answers provided for the questions that participants asked are unrepresentative of what is currently possible by state-of-the-art visual question answering systems. This thesis does not fully evaluate how automated visual question answering systems will perform in context, and instead focuses more on the types of questions that can arise in audio description writing scenarios.

Lastly, the small number of participants (N = 6) included in this study limits the generalizability of the results. Next steps for this work include deploying a system that can be navigated by BLV users on their own devices and integrating automated visual question answering systems to fully explore independent writing possibilities.

6.5 Future Work

Exploring technological solutions in an unconsidered space for increasing BLV involvement can greatly impact future efforts for the necessary inclusion of disabled users. Regarding next steps, we are integrating our research findings to the development of VerbalEyes, an end-to-end audio description creation platform made with BLV creatives in mind. By including accessible features such as keyboard shortcuts and accessible script formatting, we hope to streamline the processes of experienced and novice BLV audio description writers. Blind and low vision talents may also become involved in AD pipelines as quality control consultants, narrators, audio mixers, and directors. This work informs future explorations in co-designing accessible technologies with end users, which can be in the area of audio description, closed captioning, or other means of information access. Alongside making technical strides, we also plan to continue advocating for the involvement of BLV talents in AD to work towards eliminating the harmful and discriminatory stigma surrounding disabled creatives.



Ratatouille - Remy Fixes the Soup

00:00.1 - 00:04.9 Edit

Remy, a grey rat with a pink nose and paws, takes in the aroma of the soup.

00:05.3 - 00:09.1 Edit

He scurries to grab small chunks of garlic and tosses them into the pot.

00:09.5 - 00:17.9 Edit

At the windowsill, facing away from the soup, his eyes lead as he turns to look back at the simmering pot. A plump ghost chef pops up next to him.

00:23.6 - 00:27.2 Edit

00:07.9 / 01:24.8

Remy, a grey rat with a pink nose and paws, takes in the aroma of the soup.

He scurries to grab small chunks of garlic and tosses them into the pot.

At the windowsill, facing away from the soup, his eyes lead as he turns to look back at the simmering pot. A plump ghost chef pops up next to him.

He looks at the soup resolutely and his ears flatten.

He scurries down from the windowsill onto the counter, using his feet to turn down the fire beneath the pot.

Using a droplet of water from the tap he washes his paw

Figure 3: Screenshot of the VerbalEyes studio interface, featuring the Ratatouille video.

7 Conclusion

Despite being critical to providing video access for blind and low vision audiences, a vast majority of today's video content lacks audio description. BLV writers are limited in the ways they can participate and contribute to current audio description pipelines due to harmful stigmas and inaccessible AD writing technology. This thesis introduces the usage of visual question answering models in the context of creating audio descriptions, uncovers perspectives from the BLV community regarding the visual context that is important for them to write artistic audio descriptions, and explores ways in which advancing technology can be used to reduce stigmas and further disability inclusion. Our findings indicate that the blind and low vision community is interested in becoming involved in the AD creation pipeline, and that they are frustrated by the lack of opportunities and access to do so. The context that is most important to BLV creatives, both as writers and viewers, takes the form of character descriptions, action descriptions, settings, and clarifying sound effects. Accessible features that are necessary for access include baseline pre-generated audio descriptions, visual question answering support, and multiple modes of input. This investigation also discusses ways to reduce the stigma against BLV creatives, potential ramifications of this technology and these designs, and ways in which this research work can be translated into real-world applications and impacts. This work extends previous audio description and accessibility research to provide new insights into co-designing technology to support the BLV community and to push for societal change.

8 Acknowledgements

I am incredibly grateful to Dr. Richard Ladner for advising this project, for mentoring me as an undergraduate researcher despite being a Professor Emeritus, and for inspiring and encouraging me to pursue accessibility research – this thesis and work would not exist without his support over the last four years. From sparking my interest in accessibility research through a lecture during Early Fall Start, to organizing the Study Away Silicon Valley program, the impact that he has made on my UW journey is immeasurable.

Thank you to Dr. Leah Findlater, who welcomed me into the Inclusive Design Lab and empowered me to work on research beginning in Winter 2019 despite it being my second quarter at UW. Thanks also to Lotus Zhang and Emma McDonnell, two inspiring PhD students and mentors who have challenged me to explore new ways of approaching accessibility research. Thank you also to Dr. Amy Zhang, who inspired me to consider academia and professorship as a future path.

I am grateful to the study participants for their time, energy, and invaluable perspectives regarding audio description and overall video accessibility. I have learned so much from the amazing community in the HCI and accessibility research groups on campus, including UW CREATE and DUB, and I am extremely thankful to have had these opportunities. This research has been supported by a UW CREATE Student Mini-Grant and the Mary Gates Endowment.

Lastly, thank you to my friends and family for their support and encouragement in research and beyond. I am especially thankful for Daniel Zhu, for being an excellent research partner, for pushing me to think critically about my research questions, for reading and providing feedback on draft after draft, and for always being there. I would also like to thank Sophia Hwang, for being the best friend I could have ever hoped for over the last 10+ years. It is a deep honor and privilege to be able to work on projects that I love, and I would not be able to do it without each and every one of them.

References

[1] Everett Bacon. 2022. *Post in the Audio Description Discussion Facebook Group.* https://www.facebook.com/groups/AudioDescriptionDiscussion/permalink/1771528552994084/

[2] Angela G. Blackwell. 2017. *The Curb-Cut Effect.* https://ssir.org/articles/entry/the_curb_cut_effect

[3] Virginia P. Campos, Tiago M. U. de Araújo, Guido L. de Souza Filho, and Luiz M. G. Gonçalves. 2020. CineAD: a system for automated audio description script generation for the visually impaired. 19 (2020), 99–111. https://doi.org/10.1007/s10209-018-0634-4

[4] Chongyan Chen, Samreen Anjum, and Danna Gurari. 2022. Grounding Answers for Visual Questions Asked by Visually Impaired People. (2022). https://doi.org/10.48550/arXiv.2202.01993

[5] The International Agency for the Prevention of Blindness (IAPB). 2022. *Magnitude and Projections.* https://www.iapb.org/learn/vision-atlas/magnitude-and-projections/

[6] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Grounding Answers for Visual Questions Asked by Visually Impaired People. (2018). https://doi.org/10.48550/arXiv.1802.08218

[7] Shasta Ilhorn, Yue-Ting Siu, Aditya Bodi, Lothar Narins, Jose M. Castanon, Yash Kant, Abhishek Das, Ilmi Yoon, and Pooyan Fazli. 2022. NarrationBot and InfoBot: A Hybrid System for Automated Video Description. (2022). https://doi.org/10.48550/arXiv.2111.03994

[8] IMDb. 2007. *Ratatouille.* https://www.imdb.com/title/tt0382932/

[9] IMDb. 2015. *Inside Out.* https://www.imdb.com/title/tt2096673/

[10] Lucy Jiang and Daniel Zhu. 2021. VerbalEyes: A Large-Scale Inquiry into the State of Audio Description. (2021). https://doi.org/10.48550/arXiv.2111.03994

[11] Robert Kingett. 2021. *I blindly described a game trailer.* https://blindjournalist.wordpress.com/2021/07/25/i-blindly-described-a-game-trailer/

[12] Ren Leach. 2022. *Post on Twitter.* https://twitter.com/renleach/status/1524379908373303296

[13] Byron Lee. 2020. *Post in the Audio Description Discussion Facebook Group.* https://www.facebook.com/groups/AudioDescriptionDiscussion/permalink/1435708533242756/

[14] 3Play Media. 2022. *The Ultimate Guide to Audio Description.* https://www.3playmedia.com/learn/popular-topics/audio-description/

[15] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian L.Y. Chan, Ebrima H. Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. The Efficacy of Col-

laborative Authoring of Video Scene Descriptions. (2021). https://doi.org/doi/abs/10.1145/3441852.3471201

[16] The American Council of the Blind. 2022. *The Audio Description Project.* https://adp.acb.org/

[17] The American Council of the Blind. 2022. *Audio Description via Netflix.* https://adp.acb.org/netflix.html

[18] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. (2020). https://doi.org/10.48550/arXiv.2010.03667

[19] Thomas Reid. 2021. *Flipping the Script on Audio Description – Going Social.* http://reidmymind.com/flipping-the-script-on-audio-description-going-social/

[20] Roy Samuelson. 2021. *Audio Description Snobbery.* https://acbvoices.wordpress.com/2021/04/16/audio-description-snobbery/

[21] Justin Sciuletti. 2015. *Netflix adds audio descriptions for visually impaired to 'Daredevil' and other shows.* https://www.pbs.org/newshour/arts/netflix-adds-audio-descriptions-visually-impaired-daredevil-shows

[22] Pixar Animation Studios. 2007. *Ratatouille.* https://www.disneyplus.com/movies/ratatouille/4zRnUvYGbUZG

[23] Pixar Animation Studios. 2010. *Ratatouille Cooking Scene.* https://www.youtube.com/watch?v=jwLKPDJqldw

[24] Pixar Animation Studios. 2015. *Disgust Anger - Disney's INSIDE OUT Movie Clip.* https://www.youtube.com/watch?v=AQ3hjymiCCg

[25] Pixar Animation Studios. 2015. *Inside Out.* https://www.disneyplus.com/movies/inside-out/uzQ2ycVDi2IE

[26] Debi Tate. 2021. *Post on Twitter.* https://twitter.com/ddtate/status/1462293581524344836

[27] Salamishah Tillet. 2021. *'Bridgerton' Takes On Race. But Its Core Is Escapism.* https://www.nytimes.com/2021/01/05/arts/television/bridgerton-race-netflix.html

[28] Robbie Whelan. 2022. *'Bridgerton' Is About to Get Saucier.* https://www.wsj.com/articles/bridgerton-superfans-embrace-audio-option-that-narrates-steamy-on-screen-action-11648223396

[29] Beste F. Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Sueng 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. (2020), 47–60. https://doi.org/doi/abs/10.1145/3441852.3471201