Audio Description Customization

Rosiana Natalie
Singapore Management University
Singapore
rnatalie.2019@phdcs.smu.edu.sg

Ruei-Che Chang
University of Michigan
Ann Arbor, MI, USA
rueiche@umich.edu

Smitha Sheshadri
Singapore Management University
Singapore
smithas.2022@phdcs.smu.edu.sg

Anhong Guo
University of Michigan
Ann Arbor, MI, USA
anhong@umich.edu

Kotaro Hara
Singapore Management University
Singapore
kotarohara@smu.edu.sg

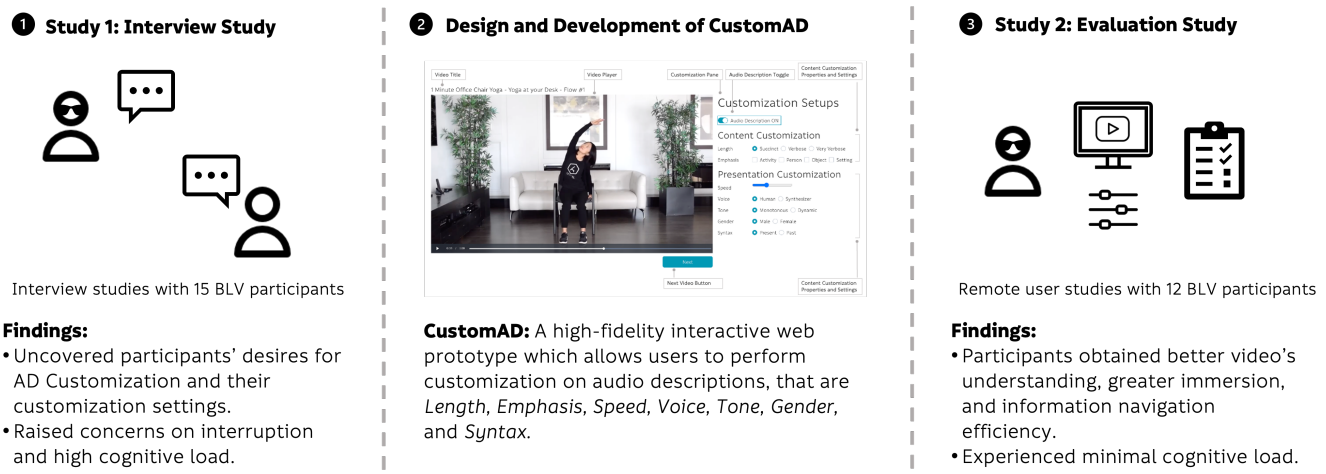

Figure 1: Our research on the customization of audio descriptions consists of three parts: 1) An interview study with 15 BLV participants to uncover their desires and preferences in audio descriptions customizations, 2) the design and development of CustomAD, a high-fidelity prototype that reflects the audio description customization preferences that emerged from the interview study, and 3) an evaluation study on the effectiveness and trade-offs of audio description customization.

Abstract

Blind and low-vision (BLV) people use audio descriptions (ADs) to access videos. However, current ADs are unalterable by end users, thus are incapable of supporting BLV individuals' potentially diverse needs and preferences. This research investigates if customizing AD could improve how BLV individuals consume videos. We conducted an interview study (Study 1) with fifteen BLV participants, which revealed desires for customizing properties like *length, emphasis, speed, voice, format, tone*, and *language*. At the same time, concerns like interruptions and increased interaction load due to customization emerged. To examine AD customization's effectiveness and tradeoffs, we designed CustomAD, a prototype that enables BLV users to customize AD content and presentation. An evaluation study (Study 2) with twelve BLV participants showed using CustomAD significantly enhanced BLV people's video understanding, immersion, and information navigation efficiency. Our work illustrates the importance of AD customization and offers a design that enhances video accessibility for BLV individuals.

Keywords

Accessibility, Blind and Low-vision Individual, Video Accessibility, Audio Description, Customization

ACM Reference Format:
Rosiana Natalie, Ruei-Che Chang, Smitha Sheshadri, Anhong Guo, and Kotaro Hara. 2024. Audio Description Customization. In *The 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24), October 27–30, 2024, St. John's, NL, Canada.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3663548.3675617

ASSETS '24, October 27–30, 2024, St. John's, NL, Canada

ACM ISBN 979-8-4007-0677-6/24/10
https://doi.org/10.1145/3663548.3675617

1 Introduction

Blind and low-vision individuals (BLV) depend on audio descriptions (ADs), verbal narrations of visual content, to comprehend videos [35, 51, 72, 94, 99]. Guidelines for authoring ADs outline the qualities that good descriptions should satisfy like descriptiveness

and succinctness [7, 28, 72, 105]. A growing body of research focuses on enhancing the efficiency of AD production while meeting these quality standards, either by supporting manual authoring with technologies [28, 36, 42, 60, 69–71, 75, 78, 92, 107, 108] or by automating the process [17, 61, 74, 100]. However, the premise of these efforts is that offering one good description for each video will satisfy everyone, which may overlook diverse preferences and needs of BLV individuals when consuming ADs. While such generic "one-size-fits-all" ADs are generally accepted by BLV individuals [22], they may not be optimal.

Despite the increasing interest in technological solutions for AD production, studies aimed to understand the diversity in AD needs and the demand for personalized ADs remain relatively sparse. Prior work has mentioned the inadequacy of generic ADs in fully meeting the needs of BLV individuals and has called for more investigations [54, 100]; still, the investigations were left as future work. Most relevant to the current work are recent studies by Jiang *et al.* and Chmiel and Mazur, which offer evidence showing that AD needs among BLV individuals indeed vary depending on the contexts in which videos are consumed [52] and the levels of visual impairments [22]. While informative, several critical questions remain unanswered. For instance, what specific characteristics of ADs are essential for personalization, and how can we effectively cater to such diverse needs?

To extend the prior study in understanding BLV individuals' AD preferences and explore ways for AD personalization, we ask: *Is enabling BLV end-users to customize ADs desirable as a way to cater to diverse needs, and if so, what customization properties are perceived to be meaningful and important? How could we design a tool to support AD customization?* And, *how does AD customization affect video consumption, and how do BLV individuals perceive its impact?*

To address the first question, we conducted remote semi-structured interviews (Study 1) with fifteen BLV individuals. Participants first watched seven different types of videos with ADs, such as instructional videos and documentaries. We then asked questions to determine their interest in AD personalization, whether customization is a favorable approach to personalization, and which AD properties they consider important to customize. Participants' responses suggested that AD customization could improve experience in consuming videos, enhance AD clarity, and increase video immersion and efficiency of information consumption. Our findings also revealed that both content-related properties (*e.g., length*, *emphasis*) and presentation-related characteristics (*e.g., speed*, *tone*) could improve how BLV individuals consume ADs.

Motivated by the results of Study 1, we designed and developed CustomAD, a high-fidelity web-based prototype that enables users to customize both the content and presentation properties of ADs. Users could adjust content properties such as *length* and *emphasis*, as well as presentation properties including *speed*, *voice*, *tone*, *gender*, and *syntax* by controlling values in form elements (Fig. 4). The CustomAD interface was designed for accessibility, allowing BLV users to navigate using keyboard shortcuts. Any changes to customization settings were reflected immediately in the ADs, allowing the user to assess if the customized AD meets their liking.

Using CustomAD as an apparatus, we conducted a remote evaluation study (Study 2) with twelve BLV participants to investigate the effectiveness of AD customization and the tool's usability. The study was a two (*with vs. without* customization) by three (*entertainment*, *explainer*, and *tutorial* videos) within-subjects design. Participants were asked to use CustomAD to watch six videos, two in each video type. Each video was accompanied by diverse set of AD versions for customization authored by a professional. To measure the effect of customization on video comprehension, we asked participants to identify specific information from the ADs. We assessed their accuracy and the time taken to complete these tasks. Additionally, we collected subjective metrics to evaluate the tool's usability (*e.g.,* Likert-scale responses measuring perceived usefulness and NASA TLX questionnaire to measure customization task load). The study concluded with a brief interview. Our findings showed that that participants' accuracy in performing the information identification task was significantly higher with customization. However, they took longer to complete tasks, as they spent more time interacting with the customization interface. Subjective data suggested that using CustomAD was easy and comfortable, with a manageable cognitive load.

In summary, our work makes the following contributions:

- Findings from the interview study with fifteen BLV individuals that uncover the desire for customizing ADs and important customization properties.
- The design and development of CustomAD that enables customization of ADs to suit BLV individuals' preferences and needs.
- Empirical findings from the evaluation study with twelve BLV participants demonstrating the effectiveness of AD customization.

2 Related Work

2.1 Audio Descriptions

Legislation in many countries mandate provision of audio descriptions (ADs) across various media, including television, cinema, and digital platforms [12, 16, 24, 25, 41, 62, 93]). For example, in the US, CVAA Title 2 [24] requires major broadcast and cable networks to make online videos accessible. The UK's 1996 Broadcasting Act specify minimum numbers or percentages of programs and their duration that must be made accessible. As the response the broadcating act, UK's TV broadcasters voluntarily have offering up to 20% of their airtime with ADs to further enhance accessibility [56]. These efforts have increased the availability of ADs, in turn improved how BLV individuals consume traditional visual media [62].

AD authoring guidelines, developed by organizations like the American Council of the Blind [7], DCMP [28], ADLab [4], and Netflix [72], also, prior works on visual descriptions (*e.g.,* [70, 85, 86, 100]) provide detailed recommendation on how to craft ADs that are useful for BLV individuals. These guidelines and research suggest what to include in ADs (*e.g.,* focusing on the important visual content and progressing from the general to the specific [85, 86], avoiding descriptions that can be inferred from the sound [7, 72], and providing right amount of information [70, 71, 100]), placement (*e.g.,* ensure they do not overlap with dialogues [7, 14, 18, 28, 80, 94, 95]), and style of delivery (*e.g.,* match the tone [7, 14, 80] and vocabulary of the source video, use active voice [7, 18, 28, 80], acceptable speed [7, 28, 80], and avoid editorializing [7]). WCAG 2.0's Success

Criterion 1.2.7 [96] advocates for the implementation of extended ADs, which temporarily pause audio and video in the original content to deliver important visual details, especially when the natural pauses in dialogue are not long enough for a detailed description.

Prior research has explored technological solutions to facilitate AD authoring while adhering to these guidelines. For example, Pavel *et al.* introduced Rescribe, a system that helps novice authors craft inline extended ADs that seamlessly integrate with other video contents [75]. Chang *et al.* developed Omniscribe [20], a tool is designed for both authoring and delivering ADs tailored to 360° content. Chang's approach enhances AD immersion by incorporating spatial audio, vibrations to signal scene transitions, and tracking head movements, introducing ways to present AD effectively for BLV individuals to enjoy 360° videos.

Although these guidelines and technologies advance the goal of making videos more accessible, they implicitly emphasize the existence of an ideal AD that is presumed to suit each video, and so creating a "one-size-fits-all" AD could meet the needs of every BLV individual. This assumption, however, overlooks the diverse preferences, needs, and abilities within BLV individuals. This paper explores the variations in AD preferences and the specific properties that BLV individuals wish to customize. We also evaluate how such customization impacts video comprehension and immersion.

2.2 Audio Descriptions Preferences

Characteristics of ADs identified in prior work as crucial for BLV individuals may also be suitable for customization [17, 53, 63, 69–71, 75, 100, 107]. For example, Jiang *et al.*, in developing Accessible AD, found that BLV individuals appreciate details related to the characters, background settings, and actions in videos [53]. Yuksel *et al.* showed that BLV individuals seek precise direction and measurements, and emphasizing such information in cooking videos to be beneficial [107]. In developing and evaluating a tool that supports authors to create ADs, Pavel *et al.* found BLV individuals value ADs that offer extensive details without overlapping with the original audio track [75]. The relevance of these characteristics like content emphasis and level of detail would vary depending on the audience and context. Thus, they provide a starting point for examining variations in AD preferences and exploring potential customization.

Some studies have explored the diverse preferences and information needs of BLV individuals for visual media [21, 22, 52, 54, 62, 65, 85, 86]. The work by Stangl revealed the varied description needs of BLV individuals, such as contexts of image use, types of sharing platforms, and the goal of the information sought [85, 86]; though their focus was on image descriptions, it is plausible such variations in needs exist for videos's ADs, too. In fact, a study by Lopez *et al.* [65] suggested that BLV individuals demand personalized ADs that are tailored to their unique abilities and interests. Studies that are directly relevant to the current work include recent research by Jiang *et al.* [52] and Chmiel and Mazur [22]. Jiang *et al.* revealed BLV individuals' preferences on levels of details and output modalities for different types of videos [52]. Chmiel and Mazur found that, while BLV individuals generally preferred ADs that adhere to existing guidelines, variations in their residual vision affected preferences on characteristics like character naming [22]. While these studies provide evidence of differences in individual preferences for AD delivery, they primarily focused on whether viewing scenarios and levels of visual impairment affect AD preferences. To complement and extend these findings, we conduct an interview study to explore the preferences for low-level characteristics of ADs, such as length, emphasis, speed, and voice, identifying which features are more desired for AD customization. Furthermore, through the design and evaluation of CustomAD, we investigate the objective effectiveness and subjective usability of end-user customization of ADs, providing insights into user interactions for personalized ADs.

2.3 Customization of User Interfaces

Previous research in Human-Computer Interaction (HCI) has highlighted the benefits of enabling end-users to customize various aspects of their digital environments. This includes customization of menus and toolbars [15, 27, 49], interface layouts [26, 37, 79, 83, 89, 90, 101], content [43, 76], and information visualization [2, 10, 58]. Customization has shown to be effective in supporting the unique needs of individuals with disabilities, too. For instance, blind users have adjusted options for screen readers [3, 8, 84], online discussion forums [91], navigation tools [11], audio books [68], graphical user interfaces [37], and augmented reality [67] to better suit their needs. The past work consistently supports that customization improves digital tools' usability and accessibility [45, 85, 86]. Thus, designing a tool that gives BLV individuals autonomy and control to customize the content and presentation of ADs could be a viable design direction for AD personalization. In this research, we explore what BLV individuals want to customize in ADs beyond basic AD toggling feature that existing video platforms like YouTube [106], Netflix [73], Disney+ [29], and Amazon Prime [6] provide. Our study is the first to design, develop, and evaluate the effect of AD customization on the video-watching experience of BLV individuals.

3 Study 1: Interview Study Method

We conducted remote semi-structured interviews with BLV individuals to investigate whether AD customization is viable for accommodating their diverse AD needs. The study also explored what customization properties are perceived as essential.

3.1 Participants

We recruited N=15 participants through snowball sampling [39]. All the participants completed the demographic questionnaire before the study. Among the fifteen participants (7 female, 8 male), six had congenital visual impairments, and nine acquired their visual impairments later in their life. Seven participants were totally blind, and eight were low-vision. The average age of the participants was 42 years old (SD = 13.63, Md = 40). See the demographic information in Table 1.

3.2 Procedure

We conducted the study remotely over Zoom or FaceTime. Each session consisted of a video-watching activity and an interview session in which we discussed customization viability and preferences. To expose the participants to various videos with ADs and to establish

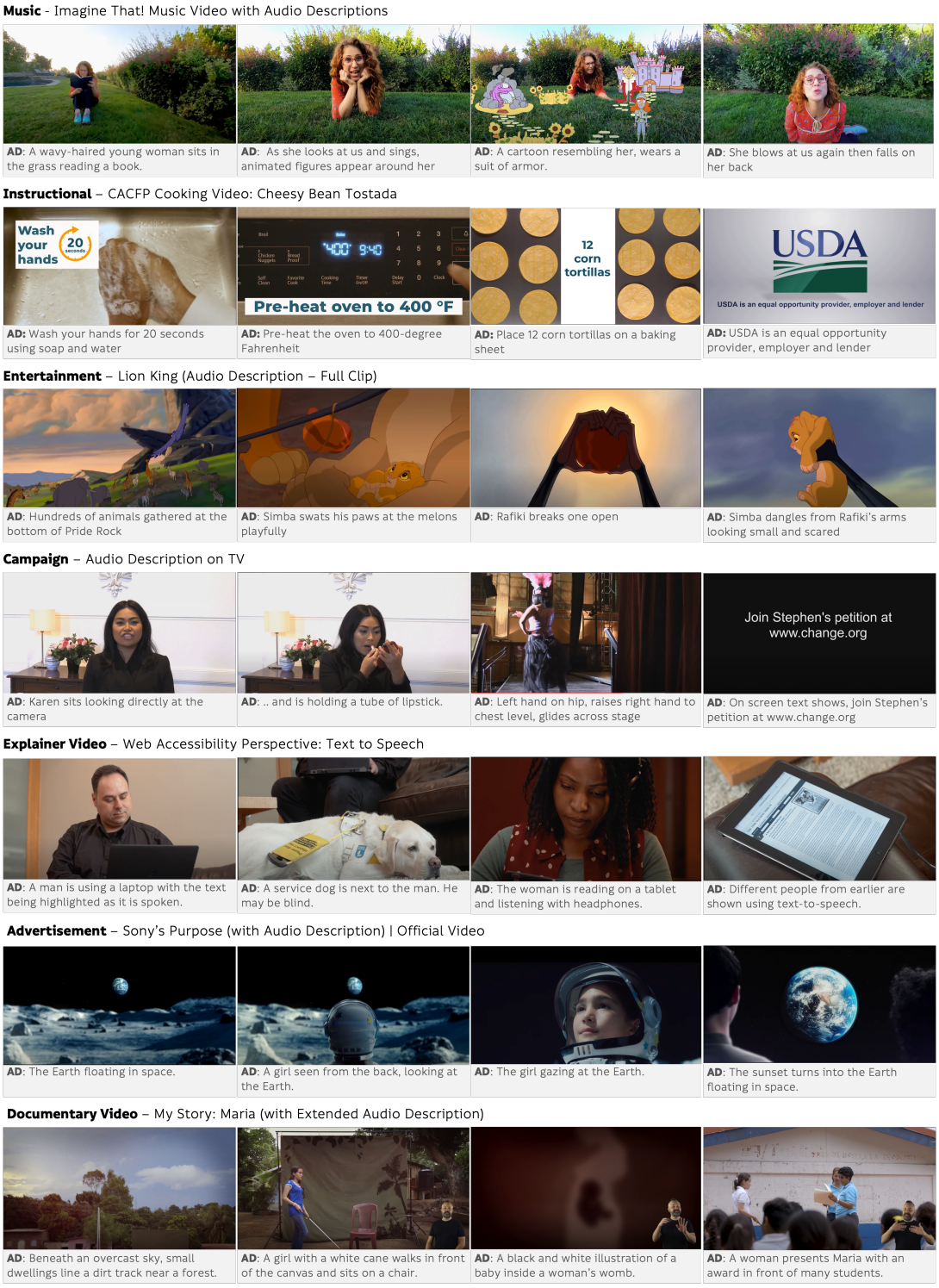

Figure 2: Videos and their ADs used in Study 1. Each row represent a video. Video types are: *music, instructional, entertainment, campaign, explainer, advertisment,* and *documentary.* Each participant watched all seven videos to increase their awareness of different AD contents and styles.

ID	Gender	Age	Primary Occupation	Level of Vision	Visual Onset	Diagnosis	Frequency of Watching
P1∗	Male	34	Freelance	Blind	Acquired	Cataract & Retinal Detachment	Everyday
P2	Male	25	Digital accessibility specialist	Low Vision	Congenital	Glaucoma	Everyday
P3	Female	59	Tour guide	Blind	Congenital	Cataract & Glaucoma	Once a week
P4∗	Male	26	Trainer and consultant	Low Vision	Congenital	Retinal Dystrophy	Undetermined
P5∗	Male	42	Executive	Low Vision	Acquired	Retinitis Pigmentosa	Once a week
P6∗	Male	41	Technology analyst	Low Vision	Congenital	Retinitis Pigmentosa	Everyday
P7	Female	34	Receptionist	Low Vision	Congenital	Phthisis Bulbi	Everyday
P8∗	Male	62	Retired	Low Vision	Acquired	Congenital Sclerocornea & Glaucoma	Everyday
P9	Male	66	Retired	Blind	Acquired	Congenital Cataracts & Glaucoma	Everyday
P10∗	Female	50	Senior manager	Blind	Acquired	Central Retinal Artery Occlusion	Once a week
P11∗	Female	54	Part time tour guide	Blind	Congenital	Retinal Detachment	Undetermined
P12∗	Female	39	Restaurant server	Low Vision	Congenital	Maculopathy	Everyday
P13	Male	40	Call center agent	Low Vision	Congenital	Cone Exstrophy	Everyday
P14∗	Female	26	Administrative assistant	Blind	Congenital	Retinopathy of Prematurity	Everyday
P15∗	Female	29	Civil servant	Blind	Acquired	Glaucoma	Once a week
P16∗	Male	24	Student	Low Vision	Congenital	Retinitis Pigmentosa	Everyday
P17∗	Female	27	Coach and facilitator	Low Vision	Congenital	Aniridia and Glaucoma	Everyday

Table 1: Demographic information of the participants for Study 1 and Study 2. P1 to P15 participated in Study 1. Participants annotated with an asterisk (∗) participated in Study 2 (ten participated in both studies). For level of vision, Blind indicates participants with total blindness and Low Vision represents participants with low-vision and legally blind participants.

a common ground for the subsequent interview about customization preferences, we asked the participants to watch seven videos of different types (Fig. 2). In the interview, we asked participants about their attitudes toward AD customization and their thoughts on how different customization properties could assist them in consuming ADs. For each property, we asked participants to rate their agreement to the statement, *"The customization can help me to consume AD more effectively,"* on a 5-point Likert scale (1 = Strongly Disagree to 5 = Strongly Agree). We also asked the participants for their rationale for the scores. Additionally, we invited participants to suggest any customization properties not covered by our customization properties list described in Section 3.3.2 that they believed could also enhance AD consumption.

Our study procedure was approved by our institution's IRB. Before the study started, the participants gave consent to participate verbally. Each session lasted for about two hours. We compensated participants with $40 for their participation upon completing the session.

3.3 Apparatus

3.3.1 Videos. To provide participants with a comprehensive and diverse experience with AD, we selected videos with a wide range of types for the video-watching activity. These video types included: (a) music video of a song, "Imagine That" [87], (b) instructional videos on making a cheese tostada [33], (c) an entertainment video, a clip from the movie "Lion King" [30], (d) a campaign video promoting ADs on TV in Australia [9], (e) an explainer video on text-to-speech [9], (f) an advertisement video showcasing Sony products [38], and (g) a documentary video featuring Maria, a visually impaired girl living in Nicaragua [19]. These videos were selected from reputable sources and featured high-quality AD crafted by experts. See Fig. 2 for the screenshots of each video's scenes and their corresponding AD.

3.3.2 Customization Properties. To understand customization properties preferred by BLV individuals, we compiled a list of properties by reviewing existing AD guidelines [1, 7, 14, 18, 28, 72, 80]. We identified six primary customization properties: *speed, voice, format, tone, gender,* and *syntax* which primarily pertain to the presentation aspects of ADs. Drawing from prior research, (*e.g.*, [70, 85, 86, 100]), we also explored opportunities to enhance and customize the ADs content themselves. We examined the potential benefits of customizing *length* and *emphasis* as proxies to adjust the verbosity of information and focus on desired content, respectively. We summarized the list of customization properties in Table 2.

4 Study 1: Interview Study Result

We used a content analysis to iteratively code and analyze interview transcripts [31, 46]. We transcribed the interview recordings using Whisper [82]. The first author reviewed all transcripts and generated the initial codebook through open and axial coding. Two authors then independently coded two transcripts, achieving a Cohen's Kappa score (κ) of 0.81. The two authors resolved the disagreement and coded three additional interview transcripts ($\kappa = 0.86$). Three more transcripts were coded, resulting in a $\kappa = 0.94$ agreement. The two authors resolved disagreements and finalized the codebook. One researcher used the final codebook to code the remaining interview transcripts.

Participants generally responded positively to the customization properties presented in the study. Most participants strongly agreed that customizing *length* ($Mean = 4.20; SD = 0.86$) and *emphasis* ($Mean = 3.93; SD = 1.10$) helped consume AD more

Properties	Customization Properties Explanations	Settings & Examples
Content Customization		
Length [70, 71, 100]	Length of the AD, which also acts as the proxy of the amount of information in the AD. The more verbose the AD, the more information is covered.	**Settings:** Succinct^, Verbose, Very Verbose Succinct: *Elsa looks at Anna sadly.* Verbose: *In a ballroom crowded with richly-dressed men and women, Elsa, a young white woman with blond hair, walks away from Anna, a young, red-haired white woman. Both wear gowns, Elsa's more conservative. She also wears a tiara. Beside Anna is Hans, a formally-dressed brunet white man. Anna rushes towards Elsa, reaching for her hand. She pulls off one of Anna's gloves.*
Emphasis[85, 86]	Information category users could focus on in AD. We allow this property to be empty (<null>), so the AD will have a balanced information emphasis.	**Settings:** <null>^, Activity, Person, Object, Setting Activity × Succinct: *Elsa sadly looks at Anna.* Object × Verbose: *In the ballroom, Elsa walks away from Anna. Elsa wears a conservative gown with a tiara, while Anna wears a tiara as well. Anna rushes towards Elsa, swiftly removing her glove.*
Presentation Customization		
Speed [7, 13, 28, 80]	AD speed in the range of 0.25 to 2, with the increment of 0.25. Values greater than 1.0x denote users speeding up the video. Values below 1.0x indicate users slowing down the video. The default value is 1.0x	**Settings:** A slider is provided for selecting speeds from 0.25x to 2.0x, with increments of 0.25. Default is 1.0x∗
Voice [1, 7, 14, 80]	Voice that reads out AD, which could be human voice or synthesizer voice.	**Settings:** Human^, Synthesizer
Tone [7, 14, 80]	Tone of AD, which could be monotonous or dynamic.	**Settings:** Monotonous^, Dynamic
Gender [72]	Gender of the voice that reads out AD.	**Settings:** Male^, Female
Syntax [7, 18, 28, 80]	Grammatical syntax of the AD. AD could be narrated in present or past tense.	**Settings:** Present^, Past Present: *They look at each other sadly. Hiding her ungloved hands, Elsa walks to the door, her cape trailing behind her. Other guests stare. Guest turn to look.* Past: *They looked at each other sadly. Hiding her ungloved hands, Elsa walked to the door, her cape trailing behind her. Other guests stared. Guest turned to look.*
Format∗ [7, 14, 18, 28, 80, 95]	Inline or extended. Inline ADs are designed to fit naturally within the existing gap between dialogues in a video. Extended ADs are longer and require pausing the video to fit their full duration.	-
Language∗∗	Customize the language of the AD to a language that the BLV viewer understands.	-

Table 2: Summary of customization properties explored in and emerged from Study 1. Settings & Examples column describes the options supported by CustomAD (options marked with '^' are default). (∗) In CustomAD, *format* was automatically adjusted depending on the *length* properties (*i.e.,* when ADs were too long to fit in the available pause, which was usually the case for the *Verbose* and *Very Verbose* ADs, they were presented in the *extended* format). (∗∗) *Language* property emerged from the interview. We did not implement it in CustomAD because the videos were in English and all participants were fluent in English.

effectively. They agreed that adjusting the AD speed was also desirable ($Mean = 3.87; SD = 0.64$). They were neutral about *voice* ($Mean = 3.33; SD = 1.11$), *format* ($Mean = 3.47; SD = 0.92$), *tone* ($Mean = 3.33; SD = 1.18$), and *gender* ($Mean = 3.67; SD = 1.05$). *Syntax* customization was the only property that they perceived not beneficial. Fig. 3 summarizes the Likert-scale responses for each customization property.

In an open discussion, three participants noted they would prefer not to customize ADs due to concerns about potential complexities associated with interacting with the setting user interface. For instance, P3, who self-identified as less tech-savvy, preferred minimal interaction or even no customization despite recognizing the potential advantages of it.

4.1 Participants' Preferences on Customization Properties

4.1.1 Length (Succinct, Verbose, Very Verbose). Most participants (N = 13) stated that they would appreciate the ability to customize the length of ADs, because it would allow them to tailor the amount of information to their individual preferences and satisfy their curiosity about a scene.

> *"I like it because, you know, different people have different preferences and curiosity. So some people might want to know more visual information, some people might want to know less."* - P1

Six participants shared that the need to adjust AD length depended on the type of video and their intention to watch it. For example, participants expected more detailed AD to fully follow the instructions in instructional videos that required complete understanding. P15 said, *"It depends on the intention. For example, if it's like an exercise video, then I will adjust it to be very descriptive because maybe I want to know what's the accurate form [...]. But then if, for example, it's like a music video or something similar, then probably I wouldn't need a very detailed description."* (P15)

4.1.2 Emphasis (Person, Activity, Object, Setting). Nine participants believed that customizing the ADs' information *emphasis* could enhance the clarity and focus of the AD. The option was deemed particularly helpful as different videos have various intended learning outcomes, and participants may want to focus on different aspects of the content.

> *"I think [emphasis customization is] good because everyone have a very different appetite or interest. Let's say for example, I'm going to be watching an instructional video. Then I would want more information on the activities. [...] So, that would help [to follow the instructions]. And then if, let's say, I am watching a documentary, maybe I like to have [AD] to focus more on the people's descriptions."* - P10

On the other hand, three participants were concerned about lacking the knowledge to decide what *emphasis* property setting to choose. Thus, they suggested encouraging the creators to determine the information focus when creating the AD.

> *It's very difficult [to decide] the good setting for [emphasis] customization, especially for people who have not watched the video, you don't even know what you want to focus on [...].*– P8

4.1.3 Speed (Slow, Fast, Original). Participants (N=12) mentioned that adjusting the speed enhanced the audio clarity of the video. For example, when the AD was too fast, the participant could slow it down, which helps to digest the AD better.

Also, the adjustment of AD speed was mentioned by five participants as a potential enhancement to information navigation efficiency while watching videos. It is based on their familiarity with consuming audio content at a faster pace. Increasing the speed enabled participants to quickly grasp the main message and navigate to the desired information in the AD.

Two participants reported that increasing the AD speed could potentially enhance their enjoyment of videos. In particular, when AD was presented in a synthesizer voice by default, which tended to be more robotic and monotonous compared to the human voice. The faster speed helped to maintain their interest and prevent them from losing focus.

4.1.4 Voice (Human, Synthesizer). Nine participants preferred human voice over synthesizer voice and consistently chose the human voice option whenever possible. For this reason, voice customization was deemed not useful. P14 said, *"I think [...] anybody will always prefer a human voice because it makes the whole thing sound natural."*

But, some participants pointed out the advantages of using a synthesizer. For instance, P8 noted that the synthesizer voice could be easier to adjust the pitch or tone without sounding unnatural. Additionally, P4 mentioned that the synthesizer can be easier to comprehend sometimes than the human voice due to the variability in human accents.

4.1.5 Format (Inline, Extended). Nine participants expressed concerns that extended descriptions negatively impacted their viewing experience because they disliked the idea of videos being paused abruptly. For instance, P4 said, *"Sometimes you don't want to cut into the flow of actual video itself. So that the [audio description] can be played seamlessly."* Nevertheless, the participants would be open to an extended format if it added value, such as providing more detailed descriptions during pause. Two participants were more positive; they mentioned that adjusting the format can enhance the clarity of the AD. Specifically, the extended version of AD allowed the participants to have more time to digest the information and understand it better.

Format customization can be beneficial in different ways depending on the type of video (N=6). For example, P2 found that extended versions were particularly helpful for videos where the dialogue or monologue was unrelated to the visual content being displayed. However, in fast-paced videos like movies, participants preferred inline descriptions.

> *"It really depends on the kind of video that you're playing. Okay, for example, some videos, that's movies they're very fast moving you will need something in-line. I mean you cannot pause it and then let the [audio descriptions play]. It will break the flow and I will not enjoy that."* - P3

4.1.6 Tone (Monotonous, Dynamic). Participants believed that customizing the tone to fit into the overall video's mood (*i.e.*, dynamic setting) would enhance the immersion (N=9). P1 said, *"Audio descriptions can be sounded in a more "annoyed" tone to really emphasize that the user is annoyed with the situation. You know, it then conveys more emotion in the audio descriptions."*

Other than conveying emotion, the different tones of AD can make the video more engaging and less monotonous. *"Maybe [in] a love story, when the couple is happy, the describer can speak in a more upbeat tone. Then let's say if they break [up], then the describer described in a very sad tone."* - P11

4.1.7 Gender (Female, Male). Participants were mainly indifferent to gender voice customization as long as there was enough contrast between the AD and the dialogue or monologue narrator (N=6). Aligned with [72], regardless of the gender voice, the AD should contrast with most of the voices in the videos.

4.1.8 Syntax (Present Tense, Past Tense). All but one did not find syntax customization to be meaningful. They believed that customizing the syntax would not have any impact on consuming AD. However, one participant mentioned that having the option to

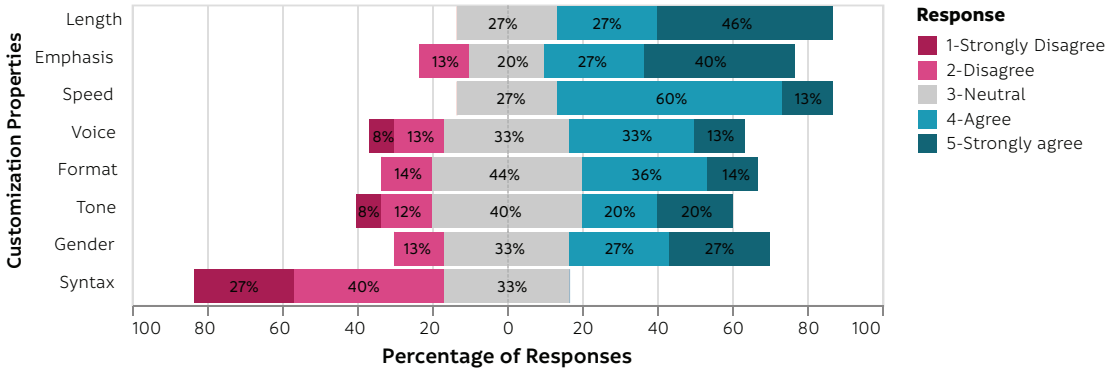

Figure 3: Summary of Likert scale questionnaire responses on customization properties preference by participants. Participants rated how much they agreed that the customization property could help them to consume AD more effectively

present the AD in different syntax helps her to feel the temporal factor of the video, that is to indicate whether things are happening in the present or the past.

4.1.9 Emergent customization property. While we designed our study to focus on eight customization properties (*i.e.*, six presentation and two content customization properties), additional customization of interest emerged. Three participants recommended allowing users to customize the *language* of AD. Currently, AD in online streaming platforms is typically provided in the same language as the video's language. While these platforms may offer dubbing options for the video, the AD remains in the original language. P15 said, *"Maybe language. Because sometimes I want to watch Korean dramas, but because there's no dubbed [audio descriptions], then I can't watch it."* (P15).

4.2 Section Summary

Overall, participants expressed that customizing properties like *length, emphasis, speed, voice, format, tone,* and *language* would positively impact video understanding, audio clarity, immersion, and information navigation efficiency. Also, customization would allow participants to have the flexibility to personalize their experience based on individual preferences. However, some concerns were raised about potential interruptions caused by specific customization properties, such as format. Moreover, there is a need to have intuitive usability and reasonable defaults for those who are not tech-savvy or prefer minimal interaction and customization with the system. Lastly, participants expressed uncertainty about selecting desirable settings, e.g., determining the specific information to focus on for the *emphasis* property.

5 CustomAD: A Hi-fi Prototype for Audio Description Customization

Insights from Study 1 shed light on BLV participants' desire for AD customization, its perceived benefit, and the potential trade-offs tied to different customization properties. These observations has motivated us to conduct an experiment to substantiate utility of AD customization through a controlled study. As a groundwork to conduct the experiment, we designed and developed the high-fidelity web prototype named CustomAD.

CustomAD is a prototype designed to support AD customization for BLV individuals. The system consists of two main parts: a video pane on the left and a customization pane on the right (Fig. 4). Users can play, pause, and seek the video with the video pane. On the right side of the interface, there is the customization setups pane, which displays various customization properties that users can modify to tailor their experience. The interface can also be fully operated with keyboard shortcuts for accessibility (Appendix 1).

Using CustomAD, participants can customize seven properties: *length, emphasis, speed, voice, tone, gender,* and *syntax* (Table 2). These customization properties are grouped into two categories: content and presentation. Content customizations focus on the *length* and *emphasis* properties, while presentation customizations involve *speed, voice, tone, gender*, and *syntax*. For each property, we offer two to four settings to choose from, except for the *speed* property, where the user adjusts the value using a slider; the participant can adjust *speed* property from 0.25x to 2x in increments of 0.25 using the slider. This setting choice was inspired from existing video streaming platform like, YouTube, which allowed a speed range between 0.25x to 2x using slider. For *length*, the settings are *succinct*, *verbose*, and *very verbose*. For *emphasis*, participants can choose to focus the ADs on descriptions of *activity*, *person*, *object*, or *setting*. For the *voice* property, the options are *human* or *synthesizer*. For *tone*, the options are *monotonous* or *dynamic*. For *gender*, the options are *male* or *female*. And for *syntax*, the choices are between *present* and *past*. The default settings for *length, emphasis, speed, voice, tone, gender,* and *syntax* are *succinct*, *null*, *1.0x*, *human*, *monotonous*, *male*, and *present*, respectively. Any changes in customization settings take immediate effect on ADs.

The findings from Study 1 informed how we chose the list of customizable properties and how we designed CustomAD. For instance, both *length* and *format* properties are made customizable by manipulating the *length* option; as a user adjusts the option between *succinct*, *verbose*, and *very verbose*, the system automatically switches the AD format between normal and extended AD format. This design decision was appropriate because most participants would want to have an extended AD that pauses a video to fit

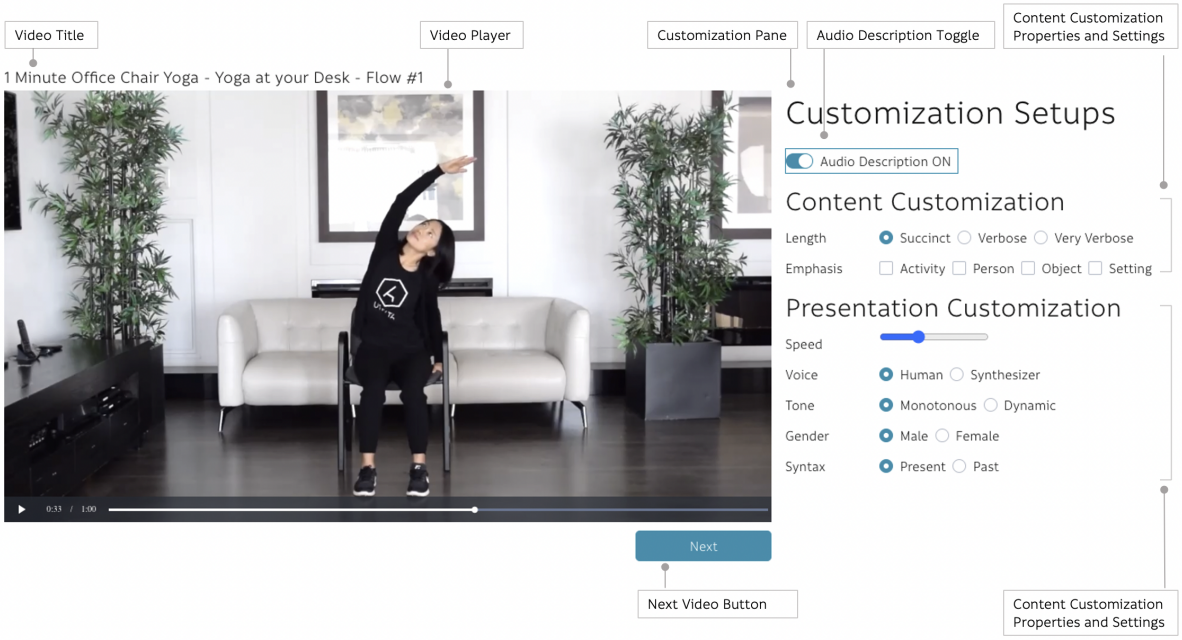

Figure 4: CustomAD interface consists of a video player (left), which allows users to play, pause, and seek the video, and a customization pane (right) where users can customize the properties of ADs. The customization properties are grouped into content settings and presentation settings. The content customization adjusts the script's content as users change the ADs length and emphasis. In presentation customization, users could adjust speed, voice, tone, gender, and grammatical syntax of the ADs to change how the ADs are read out. Users can also toggle the ADs on and off.

an AD only when the long description added substantially more information.

We applied speed customization across the entire video, not just the AD segment to minimize unnecessary silence when the participant is speeding up the AD. Though the majority considered syntax customization unnecessary, we kept this property to address possible long-tail user requirements. The videos we used in Study 2 that we describe below were in English, and all participants were fluent in English. Thus, we left the evaluation the *language* property that surfaced through Study 1 for future research. We gave the participants the option of not to set a value to *emphasis* property and leave it empty for a more balanced emphasis on the information. This decision was also made to address the participants' concern about not knowing what setting they should choose in information emphasis.

6 Study 2: Evaluation Method

To investigate the effectiveness of AD customization for video consumption and to evaluate how BLV individuals perceive its impact, we conducted remote user study with 12 BLV people.

The study is a two (*with-* and *without-customization*) by three (*entertainment*, *explainer*, *tutorial* videos) within-subjects design. We chose video type as our independent variable because videos' visual content and AD presentation varied depending on the video's intended audience, goal, and tone, which in turn could influence the utility of customization. To mitigate potential learning effect, we counterbalanced the sequence in which participants engaged with video customization and encountered different video types.

6.1 Videos

We used three different videos in our study—*entertainment, explainer*, and *tutorial*. We selected these video types because of their popularity among online video viewers [40, 77]. Although the list provided is not specifically for BLV individuals, we believe that what is commonly watched by sighted people should also be accessible for BLV individuals. Beyond their popularity, these three video types also capture the diversity of the videos' delivery styles, objectives, and ways of consumption which is suitable to evaluate the impact of customization on different video types. To see variability within the type of videos, we selected two videos to watch for each type. We chose videos that are about one- to two-minute because it is suitable for the duration of each study session. The videos we used are as follows (see the example scenes for each video in Fig. 5):

- Entertainment Video 1 (EN1): Disney's Frozen "Party is Over" [88], Duration: 49 seconds. A short clip of Disney's Frozen that shows two main characters arguing in a ball reception.
- Entertainment Video 2 (EN2): Frozen Movie Trailer, Duration [50]: 1 min 30 sec. A short animation video that mainly shows the characters, Olaf, a snowman, and Sven, and a reindeer fighting over a carrot, which appears to be Olaf's nose in the snowy open area.

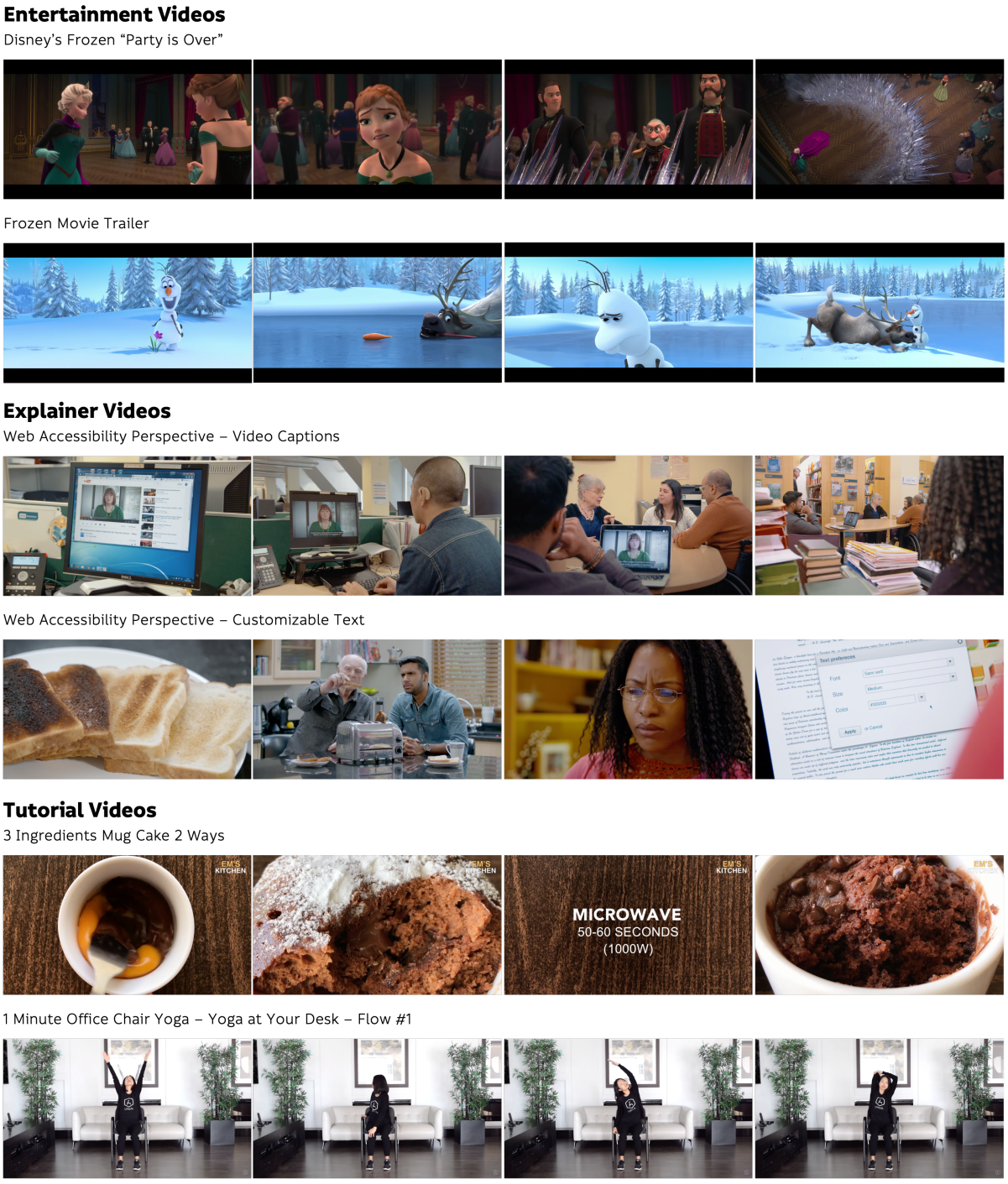

Figure 5: Videos used in Study 2. We used six videos of three types: *entertainment*, *explainer*, and *tutorial*. Each row represents a video.

- Explainer Video 1 (EX1): Web Accessibility Perspective – Video Captions [98], Duration: 1 min 17 seconds. This video explains the importance of video captions. The visuals in the videos illustrate how video captions are used in different circumstances, in which the visuals also complement the audio explanations.
- Explainer Video 2 (EX2): Web Accessibility Perspective – Customizable Text [97], Duration: 1 min 17 seconds. This video explains the importance of having the ability to customize texts in different interfaces. The visuals in the videos illustrate how customization is useful for catering to different user needs and abilities.
- Tutorial Video (Tut1): 3 Ingredients Nutella Mug Cake 2 Ways [59], Duration: 2 mins. This video shows a series of instructions for making two versions of Nutella cakes that are made in mugs. The video shows a top-down view of a pair of hands performing the cooking, all the ingredients, and the tools for cooking. Also, the video shows the text of the ingredient measurements.
- Tutorial Video (Tut2): 1 Minute Office Chair Yoga – Yoga at Your Desk – Flow #1 [104], Duration: 1 min. This video shows a woman who is sitting on the chair she usually uses for work, demonstrating several yoga movements that can be done just by sitting on her office chair.

6.2 Diverse AD Versions for Customization

For each video, we worked with an expert audio describer to create a diverse set of AD versions. The expert has received professional training (*i.e.,* Bonnie Barlow Creative Audio and Visual Describing) on audio descriptions and worked on multiple projects of creating ADs for over two years now. The expert, being a native English speaker, is particularly suitable for our projects as all our videos are in English. The expert generated fifteen distinct AD scripts for each video (see Appendix 2 for all the AD combinations scripts). These encompassed variations driven by the *length* property (*i.e.*, *succinct*, *verbose*, *very verbose*), as well as combinations of the *length* and *emphasis* properties (3 *length* × 4 *emphasis* = 12 combinations, *e.g.,* Succinct × Activity, Verbose × Person, and so on). For several videos, such as EN1, EX1, and EX2, were originally accompanied by the succinct version of AD. As a result, for these specific videos, the expert describers created only the remaining versions of AD.

After receiving the AD scripts from the experts, we generate different versions of AD for Presentation Customization with DupDub.[1] Using this website, we are able to generate different voices, style, gender, and tone which is particularly useful voice, tone, and gender properties.

6.3 Participants

We recruited twelve BLV participants who were familiar with AD (six males and six females, aged between 24 and 62; Mean = 37.92, SD = 12.34). Seven participants' impairments were congenital, while five participants acquired their impairments later in their lives. There are five total blind and seven low-vision participants. Ten participants have also participated in the Study 1 (see Table 1).

[1]https://www.dupdub.com/

6.4 Metrics

To assess the benefits and drawbacks of AD customization, we evaluated several metrics fundamental to effective video consumption. We measured participants' video understanding, time cost and prototype usability, cognitive load, and perceived usefulness, particularly in clarity, immersion, and information navigation efficiency. The primary goal of watching videos is often to actively seek information to gain new knowledge, followed closely by entertainment and enjoyment [40]. Thus, it was important to evaluate how customization plays a role in video understanding and immersion. We also evaluated task completion time and frequency of customization settings changes to better understand the user's interaction pattern with AD customization and its time cost. Also, we evaluated prototype usability, task cognitive load, and perceived usefulness as our Study 1 revealed they were also the main factor that affects the overall experience with AD customization.

We used an information-seeking task as the proxy to gauge video understanding. We curated prompts that covered dimensions like person, activities, settings, emotions, video purpose, and narrative, aimed to stimulate the information-seeking tasks. These dimensions were based on essential visual description elements in videos [85]. Please refer to Appendix 3 for all the prompts. We also collected the user interaction log that consists of the changes in customization and timestamp information. We used the log data to calculate task completion time and customization modification frequency. We also evaluate the system's usability using the SUS questionnaire [55] and exit interviews. We measured cognitive demand via the NASA Task Load Index (TLX) [23], using a 7-point Likert scale, like in previous studies [47, 48]. Additionally, participants completed Likert scale questionnaires and an exit interview to determine customization's perceived utility across video types. They considered the statement: *"Customization in [video types] enhances [clarity/immersion/information navigation efficiency] of audio descriptions."* We further discuss the participant's rationale for each rating in exit interviews.

6.5 Procedure

At the start of the user study, we provided participants with a detailed explanation of the research's aims, the concept of customizing AD, and the upcoming tasks they will perform in the session. Also, we got their verbal consent to participate in the online study. We thoroughly explained the functionalities and the keyboard shortcuts available within the CustomAD system. Subsequently, participants used CustomAD to watch a total of six videos—two from each distinct video type (*entertainment*, *explainer*, and *tutorial* videos). Participants could do AD customizations for three videos, while for the other three videos, they were not able to do AD customizations. To familiarize participants with the CustomAD interface, we introduced a practice video which is an explainer video of color contrast by W3C[2]. This facilitated their engagement and familiarity with the prototype until they felt confident to progress to the primary task.

After the participants were familiar with CustomAD, they started the information-seeking task, followed by answering the NASA

[2]https://www.youtube.com/watch?v=a9kNUv6N8Rk&ab_channel=W3CWebAccessibilityInitiative%28WAI%29

TLX questionnaire after each video, and ended with subjective ratings of perceived usefulness and exit interview. We opted for information-seeking tasks as a proxy for video understanding because information-seeking is often the primary motivation behind watching videos and to gain an understanding of the video to absorb new knowledge [40]. In addition, we believed that answering these questions would also encourage the participants to perform AD customization, something they might have not been familiar with yet. The prompts were given before the participants started watching the video and the NASA TLX questionnaires were administered after every video. After participants were done with the six videos, participants completed the System Usability Scale (SUS) questionnaire and the subjective ratings, along with the exit semi-structured interview. The study lasted for 2 hours. To ensure thorough analysis, we recorded and transcribed screen and audio interactions. We also recorded their interaction log at the backend, which consists of customization properties, setting changes, and timestamp information. Our study was approved by our institution's IRB, and participants consented to participate in the study through verbal consent before the study started. Participants were compensated with $40 for their participation.

7 Study 2: Evaluation Result

We adopted a mixed-methods approach and performed both quantitative and qualitative analyses of our data. We collected the video and Zoom screen recording, the answers to the information-seeking prompts, the survey responses to perform both quantitative and qualitative analyses, and the interaction log with the system. We reviewed the proportion of the questions in which the participants got the correct answer over the total number of questions. We also analyzed the NASA TLX task-load questionnaire Likert-scale results. We reviewed both the session recording and interaction logs to extract participants' interaction with- and without- customization conditions. We transcribed the exit interviews and grouped them according to (1) the perceived usefulness of the customization in general, also in different video types and (2) the usability of the system that supports the customization.

7.1 Correctness of Information Seeking Prompts

Participants' performance in answering the prompts was significantly higher in the *with-customization* condition (Fig 6). We obtained the score by calculating the proportion of the questions in which the participants got the correct answer over the total number of questions per video. On average, scores were 0.83 (SD = 0.16) with customization, compared to 0.42 (SD = 0.21) without customization. This trend extended across different video types. For *entertainment*, *explainer*, and *tutorial* videos, participants using customization achieved averages of 0.79, 0.83, and 0.88, respectively (SD = 0.15, 0.17, 0.14). In contrast, scores for participants when consuming videos without customization were 0.45, 0.37, and 0.43, respectively (SD = 0.29, 0.16, 0.16). A Generalized Linear Mixed-Model (GLMM) analysis with binomial distribution family and logit link function, where the customization and the video type are the fixed effect and participant was the random effect, showed the difference in accuracy ($z = 4.185$; $p < 0.001$) was significant.

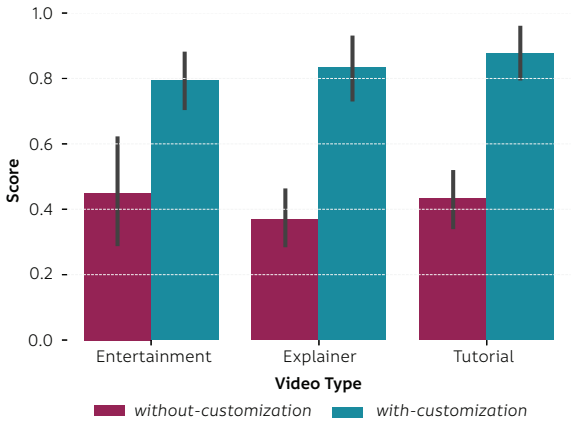

Figure 6: Average correctness scores for completing information seeking tasks for different video types (*entertainment*, *explainer*, *tutorial*) and interface conditions (*without-customization* and *with-customization*). The vertical line on each bar represents standard deviation.

GLMMs were particularly suited for regression analyses involving dependent variables bounded between 0 and 1. Additionally, GLMMs can effectively include the random effect to account for subject-specific variability for *within-subject* study design. This result suggests that customization enabled participants to tailor their AD, resulting in enhanced prompt accuracy, which indicated better video understanding.

7.2 Time and Interaction Analysis

Using the interaction log data, we calculated the duration participants took to watch the video and answer information-seeking prompts. To cater to diverse videos' original duration, we normalized the value to duration per video minute. Table 3 offers a detailed breakdown of task completion time, presented in video minute units. The *with-customization* condition took longer compared to the *without-customization* condition. Specifically, under the *with-customization* scenario, participants averaged 11.03 minutes per video minute (SD = 6.98, Md = 9.39). Conversely, the *without-customization* condition recorded an average of 6.69 minutes per video minute (SD = 6.98). The extended duration in the *with-customization* condition was anticipated as it is attributed to the time invested in the customization process. In addition, customizable properties such as *length* and *speed* properties also extended video duration.

In our evaluation of task completion time across different video types, distinct patterns emerged when comparing conditions *with-* and *without-customization*. With customization enabled, participants took the longest on Entertainment videos, on average, at 13.93 minutes (SD = 10.56 minutes), followed by Tutorial videos at 9.28 minutes (SD = 3.45 minutes), and finally explainer videos at 9.87 minutes (SD = 4.19 minutes). In contrast, without customization, the pattern shifted: Participants took the longest to complete tutorial videos, which was 7.18 minutes on average (SD = 4.05 minutes), followed by entertainment at 6.52 minutes (SD = 4.04 minutes), and explainer videos, which was 6.38 minutes, on average (SD = 4.89 minutes). A GLMM analysis with gamma distribution family and

		Task Completion Time (per video minute)		
Condition	Video Types	Mean	SD	Md.
without-customization	*Overall*	6.69	4.23	5.60
	Entertainment	6.52	4.04	6.25
	Explainer	6.38	4.89	5.16
	Tutorial	7.18	4.05	6.79
with-customization	*Overall*	11.03	6.98	9.39
	Entertainment	13.93	10.56	12.91
	Explainer	9.28	3.45	8.96
	Tutorial	9.87	4.19	8.98

Table 3: Summary of task completion time for *without-customization* and *with-customization* conditions. We present the overall task completion time along with the duration for *entertainment*, *explainer*, and *tutorial* videos. We have normalized the values to duration per video minute (pvm) as durations varied across videos.

log link function, where the customization and video types were the fixed effect and participant was the random effect showed a significant difference in completion time between the *with-* and *without-customization* conditions ($z = 5.733; p < 0.001$).

From the interaction log, we also assessed the frequency of various customizations. The *emphasis* property was modified most frequently (N = 91), followed by *length* (N = 59), *speed* (N = 35), *tone* (N = 18), *gender* (N = 11), *syntax* (N = 11), and *voice* (N = 4). On average, customization patterns among participants were as follows: 8 for *emphasis*, 5 for *length*, 3 for *speed*, 2 for *tone*, and 1 each for *gender* and *syntax*. Participants performed almost no voice customization. These trends mirror insights from our Study 1. Participants had previously expressed a strong preference for *length* and *emphasis* customizations, emphasizing their value for enhancing video understanding. This was evidenced by their leading customization counts in Study 2. Participants held neutral opinions on tone and gender customizations, as these settings were infrequently adjusted in Study 2. Interestingly, though syntax customization was deemed less useful in the Study 1, some participants (N = 5) still opted for it in the Study 2. A closer examination of the data revealed a predominant preference for the "Present" syntax setting ($N_{present} = 8$, $N_{past} = 3$). As for the voice property, despite some customization considerations performed by participants, the manual review indicated a dominant selection of the "Human" voice, matching the preference identified in our Study 1.

During the exit interview, seven participants noted that the extended duration of the videos due to customization might be the greatest drawback to the overall experience. However, they also emphasized that the benefits of customization, such as gaining more visual information to understand the video's content and making the video more engaging, outweighed the additional duration required. For example, P15 mentioned:

> *"With customization, watching videos indeed have become longer, but if that helps me to know more detail of the scene, why not?"* - P15

7.3 Usability

CustomAD's usability in the *with-customization* condition obtained a notable SUS score of 84.23 (SD = 11.92), placing it in the 'Excellent' usability rating. From our semi-structured interviews, a notable portion of respondents (N=10) emphasized the value of keyboard shortcuts for enhanced ease of use.

> *"I like the fact that I can use keyboard shortcut, it makes navigation between customization properties simple and fast."* - P6

Additionally, participants (N=2) appreciated the streamlined, linear navigation and minimalist design, which was tailored and sufficient for task completion and improved video and AD consumption.

> *"From the keyboard shortcut and the voice navigation, I can feel that the system is very simple, clean, and easy to navigate. I think this is very important as customization itself is very complex, so easy navigation between customization properties is very crucial."* - P11

Another notable feature was the instant reflection of customization changes, allowing participants to immediately perceive modifications without unnecessary delay. However, there was feedback from one participant regarding potential disturbance when the menu voice synthesizer played concurrently with the video. Potential solutions suggested included the use of sound beeps in lieu of full-sentence customization readings or employing audio ducking—reducing the background or video volume when vocalizing customization commands. In line with evolving viewing habits, three participants underscored the relevance of extending CustomAD's compatibility to mobile devices, emphasizing their growing preference for mobile-based video consumption over traditional platforms like laptops or TVs.

7.4 Task Load

Overall, participants rated the task load as lower when using customization than without customization while achieving high perceived performance (Fig. 7). We report the mean and standard deviation in tuple with the following format: Mean = (*with-*, *without-customization*), SD = (*with-*, *without-customization*). Specifically, participants reported lower mental load (Mean = (3.67, 4.38); SD = (1.56, 1.69)), temporal load (Mean = (2.52, 3.05); SD = (1.69, 1.36)), effort load (Mean = (3.29, 4.24), SD = (1.31, 1.30), and frustration load (Mean = (2.29, 2.89), SD = (1.35, 1.40)) with customization compared to without customization. Moreover, participants also perceived a higher performance when using customization to answer prompts (Mean = (5.14, 4.05), SD = (1.49, 1.75)). We omitted the physical effort variable in NASA TLX, as the task didn't involve any physical activity. A statistical analysis with Wilcoxon Signed-Rank test showed a significant difference in both effort ($W = 20; p = 0.04 < 0.05$) and performance ($W = 30; p = 0.01 < 0.05$).

Our exit interview results also supported this. None of the participants found the customization process, particularly for information-seeking tasks, to be difficult or excessively demanding. Considering the benefits they received from customization, participants were willing to tolerate the additional effort required.

> *"If you talk about effort, of course, i need to put more compare to watching the videos without customization. I think this is given. But, i think it is still bearable, i still feel the overall task load is minimal, considering*

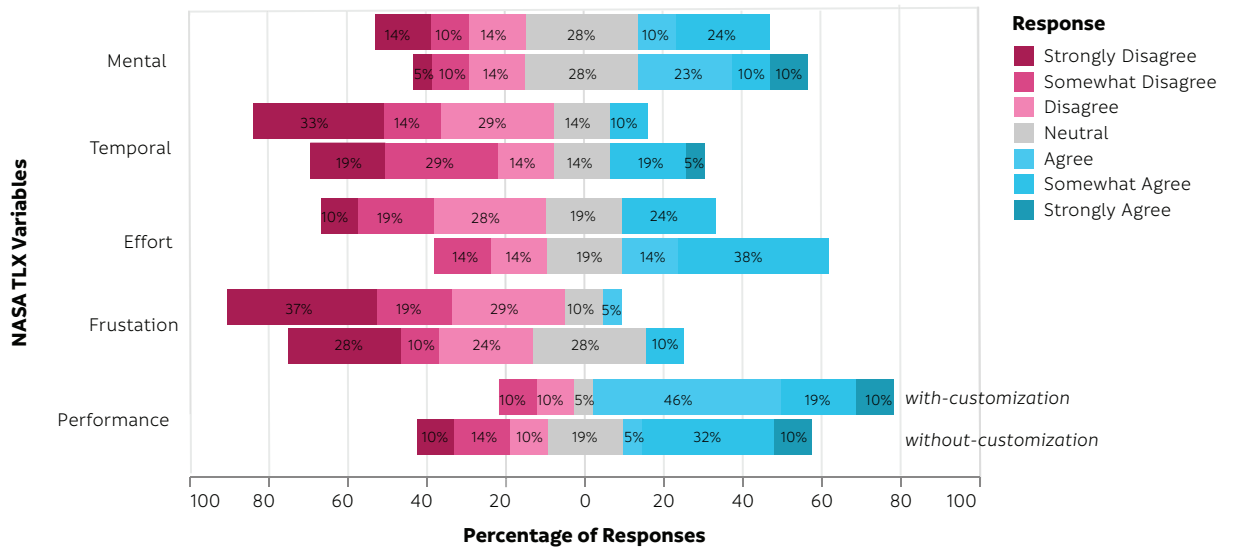

Figure 7: Summary of Likert-scale response for each NASA TLX variable, which are mental, temporal, effort, frustration, and performance. For mental, temporal, effort, and frustration, lower scores are better, while for performance higher the better.

> *the additional benefit, which is to enjoy and understand the video more, I would get from putting a slight more effort."* - P6

These results underscore the positive impact of customization. Interestingly, while the Study 1 highlighted concerns about increased effort and potential troubles with customization, our results indicate the opposite. Participants' reports of reduced task load suggest a favorable potential for adopting customization. The consistently low task load scores highlight the benefit of AD customization.

7.5 Perceived Usefulness of Customization

In terms of the perceived usefulness of customization, participants agreed that being able to customize contributed to better clarity, immersion, and information navigation efficiency of watching the video. Closer inspection at Fig. 8 reveals the rating scale is 6.33, 5.92, 6 for clarity, immersion, and information navigation efficiency, respectively (clarity: SD = 0.65; immersion: SD = 1.16; information navigation: SD = 0.74). This result suggests a broad consensus: AD customization demonstrably improved the video's clarity, immersion, and information navigation efficiency.

Delving deeper into the clarity, participants found advantages from the flexibility of information consumption enabled by the customization. The capacity to select pertinent information, also the ability to customize to an individual's information preference and requirements, helped the participants to consume the AD more effectively.

Participants mentioned the positive impact of customization on video immersion, though two cited interruptions as a drawback. The flexibility of AD length and emphasis emerged as particularly valuable, ensuring a nuanced capture of entire emotions and ambiance. For example, adjustments to the *emphasis* property, especially choosing the "Setting" setting enhanced the understanding of the video's narrative and ambiance better than staying with the default. The exit interviews revealed that customization options such as *emphasis*, *length*, and *tone* were particularly beneficial for participants with residual vision (N = 2). For instance, P16 mentioned that *emphasis* customization allowed him to supplement his limited visual input and foster a more integrative viewing experience by obtaining information about the scene's mood.

> *"with some residual [vision], I can still somehow see visuals related to object and actions, sometimes, but maybe, not so much on the overall mood. Then, I decided to choose "Setting" [in emphasis property], which I think, then, I can obtain a better understanding of the mood, whether it is happy or sad."* - P16

Though customization is perceived beneficial for good immersion of the video, however, certain customizations, such as adjusting the *length* property, require the videos to pause to accommodate extended AD. This customization makes participants feel uncomfortable because the videos' flow is interrupted most of the time. These findings are consistent with insights from the Study 1.

The flexibility of customizing the content (*length* and *emphasis* properties) and presentation, especially change in *speed* property, is deemed beneficial in increasing information navigation efficiency. Our study required the participants to answer prompts and participants (N=5) found that customizing *length* and *emphasis* quickly helped them to obtain the information they needed. For example, P12 mentioned:

> *"I think emphasis property is very useful to quickly get the answer I know. While answering the questions, I*

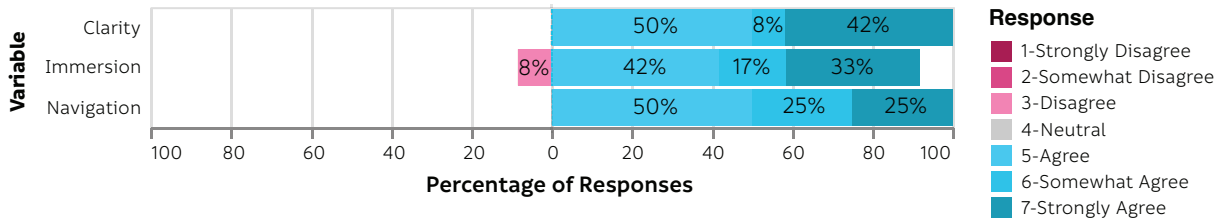

Figure 8: Summary of Likert-scale responses of perceived usefulness of AD customization for enhancing AD clarity, immersion, and information navigation efficiency.

> *was kind of deciding which category of information this question belongs to, for example, if the question is about a human or the person, then I quickly change the emphasis to "Person", then I can quickly get the answer I need. Putting the task aside, in general, I like the fact I can customize the emphasis of information. With that, I can decide what information I want to mostly be informed about, perhaps, it also depends on the video I am watching."* - P12

In addition to *emphasis*, *speed* was also perceived to be useful to increase efficiency (N = 3). Particularly, it is related to how quick or slow a user can obtain the necessary information in video content. For example, P1, who typically preferred to listen to videos at a slower speed, used to watch videos he found too fast multiple times to understand the content. However, by allowing him to adjust the speed, he could choose a pace that he felt comfortable with and could follow along more easily.

> *"I like the ability to adjust the speed. For example, when some videos have very fast audio descriptions, or the video speed is generally too fast, I have to watch the video multiple times to understand the video, but being able to adjust the speed, allows me to choose which speed is more comfortable for me and I can follow, then I don't need to watch the video multiple times"* - P1

8 Discussion and Future Work

Our research aims to enhance the video-watching experience for Blind and Low Vision (BLV) individuals by enabling them to customize Audio Descriptions (ADs). A formative study involving BLV individuals revealed a strong preference for AD customization, highlighting key properties such as *length, emphasis, speed, voice, format, tone* and *language*. Building on these insights, we then sought conclusive evidence on the impact of customization on the BLV video-watching experience. To this end, we developed CustomAD, a high-fidelity web prototype that supported AD customization. CustomAD let BLV individuals to tailor both the content and presentation of ADs, including properties such as *length, information emphasis, speed, voice, tone, gender,* and *syntax*. The evaluation results indicated that CustomAD significantly enhanced the video-watching experience for BLV individuals. The tool deepened comprehension of visual content, increased immersion, and improved efficiency in information navigation.

We discuss the trade-offs of customization and task load and duration of videos. While our Study 1 revealed some participants' concerns regarding potential interruptions and cognitive load, Study 2 alleviated this concern. Our analysis of NASA TLX results showed that participants felt AD customization imposed minimal cognitive strain. This minimal load might also be attributed to the system's high usability, as indicated by its "excellent" rating on the SUS, which likely reduced perceived interruptions and mental demand. Moreover, the benefit of doing customization (*e.g.,* obtaining more detailed information) seems to compensate for the added complexity and possible extended video duration, making customization still favorable.

8.1 Interactions for Adjusting Customization Options

In CustomAD's current design, BLV individuals manually adjust settings to align with their preferences. This autonomy to customize the AD, while offering flexibility to match the ADs to what they prefer, has surfaced challenges; people lacked the contextual or objective knowledge needed to pinpoint appropriate settings. This challenge was particularly pronounced for properties like *emphasis* and *tone*, where the appropriate setting was influenced by the video's context and learning objectives—factors users may not know in advance. Such feedback underscores the need for offering users good defaults and setting recommendations based on a video's context or learning objectives. For example, the system could suggest popular settings chosen by prior viewers, thereby aiding subsequent viewers in their customization choices. In addition, video creators should also play a crucial role in better communicating the intended objective of the videos and assign good AD defaults. Beyond human-centric solutions, there is a compelling opportunity to leverage automation. The system could, for example, analyze a video's content and automatically suggest relevant *emphasis* settings. Additionally, leveraging a video's sentiment and topic, *e.g.,* [102, 103] could inform the preferable AD tone, guiding users between a monotonous or dynamic delivery.

8.2 Generating Multiple Versions of AD

In our study, we opted for professional audio describers to generate diverse versions of AD to explore customization properties. While this approach produced high-quality AD, it induced significant time and financial costs. For instance, the creation of a minute of video content required 15 distinct scripts (*i.e.*, three different AD lengths and the combinations of 3 *length* $\times$ 4 *emphasis* properties), costing around $60 in total per video minute. Moreover, the turnaround time took more than a week, especially for visually complex videos. Given these constraints, there is an apparent need to explore more economical and efficient avenues. Notably, research by Natalie et al.

[71] underscores the viability of novice describers as a cost-effective alternative, although they often face challenges balancing quality dimensions such as succinctness versus sufficiency. An intriguing future direction might involve coupling novice-generated verbose ADs with Large Language Models (LLM) and perhaps also computer vision (CV) models for automatic refinement. For example, LLM could help in summarizing tasks to produce AD in different verbosity levels (*length* property). Also, LLM could help in rephrasing AD that is specific to the description of "Activity", "Person", "Object", and "Setting" (*emphasis* property) depending on the primary focus of the video detected by CV models. Leveraging this human-AI collaboration method could pave the way to address the identified challenges and signify a promising avenue in AD research. Furthermore, the prospect of fully automating AD generation for diverse customization properties is a compelling research trajectory. The growing field of intelligent visual descriptions in recent studies [*e.g.,* [61, 81, 100]] indicates progress, though there is more to be done to fully achieve optimal quality in AD.

8.3 AD Customization in Co-watching Activities

In our findings, customization emerged as an enabler for enhancing video comprehension and immersion, allowing individuals to tailor AD to their preferences. The present study's focus has been predominantly on the effect of customization for solo-viewing scenarios. However, this perspective may not directly translate to co-viewing experiences, where diverse preferences diverge, and individual adjustments can become a point of disagreement. When multiple viewers with distinct preferences watch content concurrently, an intriguing challenge arises: *How might a system balance and accommodate these distinct preferences without prioritizing one viewer's experience over another?* Potential solutions could involve collaborative customization, where participants input their preferences, and an intelligent system aggregates and weighs them to derive a consensus-based customization setting. Or, allowing the use of their individual headset, but still ensuring interactivity to support co-watching experience. However, as AD customization remains a nascent field with limited empirical exploration, our study's insights into its effects on solo-viewing remain relevant. Future endeavors in this domain would benefit from delving into collaborative customization paradigms to enhance co-viewing experiences.

8.4 Toward More Accessible Videos Players and Streaming Platforms

We suggested existing video players and online streaming platforms to incorporate AD customization features on top of toggling AD on and off. Our research highlights both the desire and the benefits of AD customization, which many current platforms overlook. Looking ahead, we envision evolving CustomAD into a plug-in compatible with popular video players like VLC and QuickTime Player, enhancing their AD delivery capabilities. Feedback from participants further underscores the importance of supporting AD customization for mobile devices. This is because most of them are consuming videos on mobile devices daily. In addition, given that approximately 4.18 billion individuals consumed video content on mobile devices in 2020 — a number that is expected to double annually [57] — optimizing CustomAD for diverse devices is essential. Moreover, online streaming platforms such as YouTube, Netflix, and YouDescribe could enhance accessibility for blind audiences by incorporating these customization features. Most video streaming platforms possess existing customization features, like speed, language, and video quality adjustments, and showed that users are already familiar with them. Extending beyond these customization features and implementation, we envision extending customization features to include AD customization features, enriching the viewing experience for BLV individuals.

8.5 Broadening Customization to Other Audio-Visual Based Technologies

Our study, while focused on traditional videos, highlights potential applications of customization for assisting BLV individuals across various visual-audio-based technologies. The core principles shared by conventional videos with AD and emerging technologies—reliant on visual and audio explanations for accessibility, such as AR/VR [20, 44, 64], remote sighted assistance (e.g., Be My Eyes [32], Seeing AI [5], Be Specular [34], and navigational tools (e.g., [66])—indicate a potential for applying AD customization in these technologies. For example, in the context of 360° videos, often used for BLV entertainment [20], education [44], or navigation [64] purposes, customization could allow BLV users to tailor information emphasis and verbosity. Such customization capabilities may enable BLV individuals to filter out irrelevant information and focus on areas of interest, addressing the immersion and cognitive load trade-off in a full 360° viewpoint [20]. Similarly, customization may benefit remote-sighted assistance technologies, where users can specify details like information emphasis and length for clearer and more focused assistance to understand the surroundings. The customization capabilities can also extend to the presentation of instructions, including adjustable properties such as speed, voice, and gender. Moreover, these technologies should consider expanding customization options to include properties such as viewing direction and the frequency of descriptions, which would be based on the distance at which the descriptions are triggered and read out.. Additionally, for audio-based navigation context, the ability to adjust instruction audio speed and length may significantly improve real-world mobility for BLV individuals. Customization adaptation not only preserves the essence of ADs in traditional video contexts but also broaden its utility to enhance world interaction for the BLV community.

9 Limitations

Study 2 assessed the value of customization for information-seeking tasks rather than pure content comprehension. This design choice was deliberate; we aimed to simulate scenarios of information seeking and to actively prompt participants towards customization. While this was essential in designing the study, our approach might deviate from the typical viewing behavior of a BLV individual. Future research could delve deeper, examining additional effectiveness, benefits and trade-offs of customization.

Our study was conducted in a controlled environment, limiting our ability to assess the effectiveness of customization in more

naturalistic settings. Furthermore, the videos used in the study were relatively short. Although participants engaged with six videos for an average of eleven minutes each, totaling around an hour of exposure to CustomAD, we acknowledged that studying CustomAD usage with longer videos could provide additional insights into how customization settings persist or evolve over time.

Field studies conducted in more organic, longitudinal settings — where BLV individuals customize their viewing experiences without prompts — might shed the light the nuanced interplay between customization, video comprehension, and viewing experience. Additionally, exploring community-driven and automated methods for dynamic customization in longer videos would be a valuable avenue for future research.

10 Conclusion

In this paper, we explored the potential benefits and challenges of audio description (AD) customization through an interview and an evaluation of a high-fidelity prototype for AD customization, CustomAD. The results of the interview study with 15 BLV participants highlighted the perceived benefits of customization, such as enhanced clarity and understanding of video visuals, improved immersion, and increased information navigation efficiency. Despite these advantages, participants expressed concerns regarding possible interruptions and the challenges of determining appropriate customization settings due to a lack of prior knowledge of the video's context and learning objectives. Recognizing these trade-offs, we designed and developed a high-fidelity prototype, called CustomAD, which supports customization of *length, emphasis, speed, voice, tone, gender*, and *syntax* properties. We conducted an evaluation using CustomAD with 12 BLV participants. This evaluation affirmed that customization empowers BLV users to enhance their understanding of videos, experience greater immersion, and incur minimal mental load. We concluded the study by discussing key insights and suggesting enhancements for CustomAD and similar AD customization platforms. In particular, we see the potential in leveraging both crowdsourcing and automation to streamline the process of producing customized AD versions, informing the customization setting selection, effect on the co-watching experience, and integrating customization in existing video players and online.

Acknowledgments

This research / project is supported by the Ministry of Education, Singapore under its Academic Research Fund Tier 2 (Project ID: T2EP20220-0016). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] 3PlayMedia. 2020. Beginner's Guide to Audio Description. https://go.3playmedia.com/hubfs/WP%20PDFs/Beginners-Guide-to-Audio-Description.pdf. Accessed: 2021-01-13.

[2] G. Abla, E.N. Kim, D.P. Schissel, and S.M. Flanagan. 2010. Customizable scientific web portal for fusion research. *Fusion Engineering and Design* 85, 3 (2010), 603–607. https://doi.org/10.1016/j.fusengdes.2010.02.030 Proceedings of the 7th IAEA Technical Meeting on Control, Data Acquisition, and Remote Participation for Fusion Research.

[3] NV Access. 2023. NVDA Version 2023.3.4. https://www.nvaccess.org/download/. Accessed: 2023-09-14.

[4] ADLab. 2024. ADLab Audio Descriptioon: Lifeling Access for the Blind. http://www.adlabproject.eu/Docs/adlab%20book/. Accessed: 2024-03-14.

[5] Seeing AI. 2024. Seeing AI: Talking Camera for the Blind. https://www.seeingai.com/. Accessed: 2024-09-24.

[6] Amazon. 2023. Amazon Prime. http://www.amazon.com. Accessed: 2023-09-14.

[7] Audio Description Project American Council of the Blind. 2017. Guideline for Audio Describers. https://www.acb.org/adp/guidelines.html. Accessed: 2020-11-6.

[8] Apple. 2023. Chapter 1. Introducing VoiceOver. https://www.apple.com/voiceover/info/guide/_1121.html. Accessed: 2023-09-14.

[9] Vision Australia. 2015. Audio Description on TV. https://youtu.be/ULgLZn91TMo. Accessed: 2023-11-6.

[10] Michael A. Beam and Gerald M. Kosicki. 2014. Personalized News Portals: Filtering Systems and Increased News Exposure. *Journalism & Mass Communication Quarterly* 91, 1 (2014), 59–77. https://doi.org/10.1177/1077699013514411 arXiv:https://doi.org/10.1177/1077699013514411

[11] Raju Shrestha Bineeth Kuriakose and Frode Eika Sandnes. 2022. Tools and Technologies for Blind and Visually Impaired Navigation Support: A Review. *IETE Technical Review* 39, 1 (2022), 3–18. https://doi.org/10.1080/02564602.2020.1819893 arXiv:https://doi.org/10.1080/02564602.2020.1819893

[12] Australian Communications Consumer Action Network Media Access Australia Blind Citizen Australia, Vision Australia. 2024. Blindness Sector Report on the 2012 ABC Audio Description Trial. https://www.bca.org.au/wp-content/uploads/2018/04/Blindness_Sector_Report_on_the_2012_ABC_Audio_Description_Trial.doc. Accessed: 2024-03-14.

[13] Danielle Bragg, Katharina Reinecke, and Richard E. Ladner. 2021. Expanding a Large Inclusive Study of Human Listening Rates. *ACM Trans. Access. Comput.* 14, 3, Article 12 (jul 2021), 26 pages. https://doi.org/10.1145/3461700

[14] Northern German Broadcasting. 2023. Audio description guidelines. https://www.ndr.de/fernsehen/barrierefreie_angebote/audiodeskription/Audio-description-guidelines,audiodeskription142.html. Accessed: 2023-04-09.

[15] Andrea Bunt, Cristina Conati, and Joanna McGrenere. 2007. Supporting interface customization using a mixed-initiative approach. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (Honolulu, Hawaii, USA) *(IUI '07)*. Association for Computing Machinery, New York, NY, USA, 92–101. https://doi.org/10.1145/1216295.1216317

[16] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* (2008).

[17] Virginia P Campos, Tiago MU de Araújo, Guido L de Souza Filho, and Luiz MG Gonçalves. 2020. CineAD: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society* 19, 1 (2020), 99–111. https://doi.org/10.1007/s10209-018-0634-4

[18] Media Access Canada. 2023. DESCRIPTIVE VIDEO PRODUCTION AND PRESENTATION BEST PRACTICES GUIDE FOR DIGITAL ENVIRONMENTS. http://www.mediac.ca/DVBPGDE_V2_28Feb2012.asp. Accessed: 2023-04-09.

[19] EndTheCycle CBM. 2023. My Story: Maria (with Extended Audio Description). https://youtu.be/JUIJ_aNxsG8. Accessed: 2023-11-6.

[20] Ruei-Che Chang, Chao-Hsien Ting, Chia-Sheng Hung, Wan-Chen Lee, Liang-Jin Chen, Yu-Tzu Chao, Bing-Yu Chen, and Anhong Guo. 2022. OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 15, 14 pages. https://doi.org/10.1145/3526113.3545613

[21] Agnieszka Chmiel and Iwona Mazur. 2016. Researching preferences of audio description users—Limitations and solutions. *Across Languages and Cultures* 17, 2 (2016), 271–288. https://doi.org/10.1556/084.2016.17.2.7

[22] Agnieszka Chmiel and Iwona Mazur. 2022. A homogenous or heterogeneous audience? Audio description preferences of persons with congenital blindness, non-congenital blindness and low vision. *Perspectives* 30, 3 (2022), 552–567. https://doi.org/10.1080/0907676X.2021.1913198 arXiv:https://doi.org/10.1080/0907676X.2021.1913198

[23] Lacey Colligan, Henry WW Potts, Chelsea T Finn, and Robert A Sinkin. 2015. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics* 84, 7 (2015), 469–476. https://doi.org/10.1016/j.ijmedinf.2015.03.003

[24] Federal Communications Commission. 2020. 21st Century Communications and Video Accessibility Act (CVAA). https://www.fcc.gov/consumers/guides/21st-century-communications-and-video-accessibility-act-cvaa. Accessed: 2020-11-6.

[25] Barry J Cronin and Sharon Robertson King. 1990. The Development of the Descriptive Video Servicesm. *Journal of Visual Impairment & Blindness* 84, 10 (1990), 503–506.

[26] Ionut Damian, Birgit Endrass, Peter Huber, Nikolaus Bee, and Elisabeth André. 2011. Individualized Agent Interactions. In *Motion in Games*, Jan M. Allbeck and Petros Faloutsos (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 15–26.

[27] Vagner Figueredo de Santana, Rosimeire de Oliveira, Leonelo Dell Anhol Almeida, and Marcia Ito. 2013. Firefixia: an accessibility web browser customization toolbar for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility* (Rio de Janeiro, Brazil) *(W4A '13)*. Association for Computing Machinery, New York, NY, USA, Article 16, 4 pages. https://doi.org/10.1145/2461121.2461137

[28] Described and Captioned Media Program. 2020. Described and Captioned Media Program (DCMP). http://www.descriptionkey.org/quality_description.html. Accessed: 2019-03-19.

[29] Disney+. 2023. Disney+. http://www.disneyplus.com. Accessed: 2023-09-14.

[30] Jane Doe. 2016. Audio Description - Full Clip. https://youtu.be/7-XOHN2BWG4. Accessed: 2023-11-6.

[31] James W Drisko and Tina Maschi. 2016. *Content analysis*. Oxford University Press, USA.

[32] Be My Eyes. 2024. Be My Eyes. https://www.bemyeyes.com/. Accessed: 2024-09-24.

[33] USDA Food and Nutrition Service. 2022. CACFP Cooking Video: Cheesy Bean Tostada Ages 6-18, with Audio Description. https://youtu.be/9H8Ch1tcaCs. Accessed: 2023-11-6.

[34] American Foundation for the Blind. 2024. BeSpecular: A New Remote Assitant Service. https://www.afb.org/aw/17/7/15313. Accessed: 2024-09-24.

[35] Louise Fryer. 2016. *An introduction to audio description: A practical guide*. Routledge.

[36] L. Gagnon, C. Chapdelaine, D. Byrns, S. Foucher, M. Héritier, and V. Gupta. 2010. A computer-vision-assisted system for Videodescription scripting. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 41–48. https://doi.org/10.1109/CVPRW.2010.5543575

[37] Krzysztof Gajos and Daniel S. Weld. 2004. SUPPLE: automatically generating user interfaces. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 93–100. https://doi.org/10.1145/964442.964461

[38] Sony Global. 2023. Sony's Purpose (with Audio Description) | Official Video. https://youtu.be/7Tiem2QBS0U. Accessed: 2023-11-6.

[39] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

[40] Google. 2023. The Latest YouTube Stats on When, Where, and What people watch. https://www.thinkwithgoogle.com/consumer-insights/consumer-trends/youtube-stats-video-consumption-trends/. Accessed: 2023-09-12.

[41] Joan Greening and Deborah Rolph. 2007. Accessibility: raising awareness of audio description in the UK. In *Media for All*. Brill, 127–138.

[42] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. 2020. Captioning Images Taken by People Who Are Blind. In *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham, 417–434.

[43] Kari Halsted and James Roberts. 2002. Eclipse help system: an open source user assistance offering. In *Proceedings of the 20th annual international conference on Computer documentation*. 49–59. https://doi.org/10.1145/584955.584964

[44] Tania Heap, Regina Kaplan-Rakowski, and Audon Archibald. 2023. Experiencing Virtual Reality for Perspective-Taking of Blind and Visually Impaired Learners. *Available at SSRN 4595370* (2023).

[45] Jaylin Herskovitz, Andi Xu, Rahaf Alharbi, and Anhong Guo. 2023. Hacking, Switching, Combining: Understanding and Supporting DIY Assistive Technology Design by Blind People. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 57, 17 pages. https://doi.org/10.1145/3544548.3581249

[46] Daniel J Hruschka, Deborah Schwartz, Daphne Cobb St. John, Erin Picone-Decaro, Richard A Jenkins, and James W Carey. 2004. Reliability in coding open-ended data: Lessons learned from HIV behavioral research. *Field methods* 16, 3 (2004), 307–331. https://doi.org/10.1177/1525822X04266540

[47] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. https://doi.org/10.1145/3586183.3606735

[48] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang 'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 796, 17 pages. https://doi.org/10.1145/3544548.3581494

[49] Amy Hurst, Scott E. Hudson, and Jennifer Mankoff. 2007. Dynamic detection of novice vs. skilled use without a task model. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 271–280. https://doi.org/10.1145/1240624.1240669

[50] IMSTVUK. 2013. Frozen - Trailer with Audio Description). https://youtu.be/O7j4_aP8dWA. Accessed: 2023-11-6.

[51] World-Wide Web COnsortium Web Accessibility Initiative. 2023. Making the Web-Accessible. https://www.w3.org/WAI/. Accessed: 2023-11-6.

[52] Lucy Jiang, Crescentia Jung, Mahika Phutane, Abigale Stangl, and Shiri Azenkot. 2024. "It's Kind of Context Dependent": Understanding Blind and Low Vision People's Video Accessibility Preferences Across Viewing Scenarios. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 897, 20 pages. https://doi.org/10.1145/3613904.3642238

[53] Lucy Jiang and Richard Ladner. 2022. Co-Designing Systems to Support Blind and Low Vision Audio Description Writers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (Athens, Greece) *(ASSETS '22)*. Association for Computing Machinery, New York, NY, USA, Article 74, 3 pages. https://doi.org/10.1145/3517428.3550394

[54] Lucy Jiang, Mahika Phutane, and Shiri Azenkot. 2023. Beyond Audio Description: Exploring 360° Video Accessibility with Blind and Low Vision Users Through Collaborative Creation. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY, USA) *(ASSETS '23)*. Association for Computing Machinery, New York, NY, USA, Article 50, 17 pages. https://doi.org/10.1145/3597638.3608381

[55] Patrick W Jordan, Bruce Thomas, Ian Lyall McClelland, and Bernard Weerdmeester. 1996. *Usability evaluation in industry*. CRC Press.

[56] Ivana Katsarova. 2018. The audiovisual media services directive. *Briefing EU [Legislation in Progress,] European Parliament* (2018).

[57] Irina Kegishyan. 2023. Mobile Video Statistics. https://www.yansmedia.com/blog/mobile-video-statistics. Accessed: 2023-04-09.

[58] Heather Kennedy-Eden and Ulrike Gretzel. 2012. A Taxonomy of Mobile Applications in Tourism. *e-Review of Tourism Research (eRTR)* 10 (01 2012), 47–50.

[59] Em's Kitchen. 2021. 3 Ingredient Nutella Mug Cake 2 Ways. https://youtu.be/sItYaC1z_d0. Accessed: 2023-11-6.

[60] Masatomo Kobayashi, Kentarou Fukuda, Hironobu Takagi, and Chieko Asakawa. 2009. Providing synthesized audio description for online videos. In *Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility* (Pittsburgh, Pennsylvania, USA) *(ASSETS '09)*. Association for Computing Machinery, New York, NY, USA, 249–250. https://doi.org/10.1145/1639642.1639699

[61] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 2603–2614. https://doi.org/10.18653/v1/2020.acl-main.233

[62] Hoi Ching Dawning Leung. 2018. *Audio description of audiovisual programmes for the visually impaired in Hong Kong*. Ph. D. Dissertation. UCL (University College London).

[63] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 272, 14 pages. https://doi.org/10.1145/3411764.3445233

[64] Alice Lo Valvo, Daniele Croce, Domenico Garlisi, Fabrizio Giuliano, Laura Giarré, and Ilenia Tinnirello. 2021. A navigation and augmented reality system for visually impaired people. *Sensors* 21, 9 (2021), 3061. https://doi.org/10.3390/s21093061

[65] Mariana Lopez, Gavin Kearney, and Krisztián Hofstädter. 2018. Audio Description in the UK: What works, what doesn't, and understanding the need for personalising access. *British journal of visual impairment* 36, 3 (2018), 274–291.

[66] Microsoft. 2024. Miscrosoft Soundscape: A map delivered in 3D Sound. https://www.microsoft.com/en-us/research/product/soundscape/. Accessed: 2024-09-24.

[67] Mario Montagud, Pilar Orero, and Anna Matamala. 2020. Culture 4 all: accessibility-enabled cultural experiences through immersive VR360 content. *Personal and Ubiquitous Computing* 24, 6 (2020), 887–905. https://doi.org/10.1007/s00779-019-01357-3

[68] Sarah Morley. 1998. Digital talking books on a PC: a usability evaluation of the prototype DAISY playback software. In *Proceedings of the Third International ACM Conference on Assistive Technologies* (Marina del Rey, California, USA) *(Assets '98)*. Association for Computing Machinery, New York, NY, USA, 157–164. https://doi.org/10.1145/274497.274527

[69] Rosiana Natalie, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 87, 4 pages. https://doi.org/10.1145/3373625.3418030

[70] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebrima H Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. The Efficacy of Collaborative Authoring of Video Scene Descriptions. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New

York, NY, USA, Article 17, 15 pages. https://doi.org/10.1145/3441852.3471201
[71] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting Novices Author Audio Descriptions via Automatic Feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 77, 18 pages. https://doi.org/10.1145/3544548.3581023
[72] Netflix. 2020. Audio Description Style Guide v2.1. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-1. Accessed: 2020-11-6.
[73] Netflix. 2023. Netflix. http://www.netflix.com. Accessed: 2023-09-14.
[74] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding Language-Image Pretrained Models for General Video Recognition. In *Computer Vision – ECCV 2022*, Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer Nature Switzerland, Cham, 1–18.
[75] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864
[76] Peter Pawlowski. 2010. Basic Player Whose Appearance and Functions can be Customized Freely 'Foobar 2000'v1. 0 is Unveiled. *Windows Forest, Japan, Jan* 12 (2010), 3.
[77] Pictory. 2023. What are the Most Popular Genres on YouTube in 2023? https://pictory.ai/blog/what-are-the-most-popular-genres-on-youtube-in-2023?el=0035&htrafficsource=pictoryblog&hcategory=video. Accessed: 2024-04-18.
[78] Able Player. 2020. Able Player: Fuly Accessible cross-browser HTML Media Player. https://www.3playmedia.com/services/features/plugins/3play-plugin/. Accessed: 2020-11-6.
[79] Antoine Ponsard and Joanna McGrenere. 2016. Anchored Customization: Anchoring Settings to the Application Interface to Afford Customization. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 4154–4165. https://doi.org/10.1145/2858036.2858129
[80] Audio Description Project. 2023. Recommendation of the Federal Communications Commission disability ... https://adp.acb.org/docs/DAC%20Recommendation%20on%20Audo%20Description%20Quality%20Adopted%20October%2014%202020.pdf
[81] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[82] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*. PMLR, 28492–28518.
[83] Batul Saati, May Salem, and Willem-Paul Brinkman. 2005. Towards customized user interface skins: investigating user personality and skin colour. *Proceedings of HCI 2005* 2 (2005), 89–93.
[84] Freedom Scientific. 2023. JAWS. https://www.freedomscientific.com/products/software/jaws/. Accessed: 2023-09-14.
[85] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. "Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376404
[86] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. https://doi.org/10.1145/3441852.3471233
[87] Rena Strober. 2020. Imagine That! Music video with AUDIO DESCRIPTION. https://youtu.be/UXz9AtO_kl0. Accessed: 2023-11-6.
[88] Walt Disney Animation Studios. 2013. Disney's Frozen "Party Is Over" Clip). https://youtu.be/jNuZC5_9pQQ. Accessed: 2023-11-6.
[89] S Shyam Sundar and Sampada S Marathe. 2010. Personalization versus customization: The importance of agency, privacy, and power usage. *Human communication research* 36, 3 (2010), 298–322. https://doi.org/10.1111/j.1468-2958.2010.01377.x
[90] Jieun Sung, Torger Bjornrud, Yu-hao Lee, and D. Yvette Wohn. 2010. Social network games: exploring audience traits. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (Atlanta, Georgia, USA) *(CHI EA '10)*. Association for Computing Machinery, New York, NY, USA, 3649–3654. https://doi.org/10.1145/1753846.1754033
[91] Mohan Sunkara, Yash Prakash, Hae-Na Lee, Sampath Jayarathna, and Vikas Ashok. 2023. Enabling Customization of Discussion Forums for Blind Users. *Proc. ACM Hum.-Comput. Interact.* 7, EICS, Article 176 (jun 2023), 20 pages. https://doi.org/10.1145/3593228
[92] Terril Thompson. 2019. Audio Description using the Web Speech API. https://terrillthompson.com/1173. Accessed: 2020-11-6.
[93] JF Vera. 2006. Translating audio description scripts: the way forward? Tentative first stage project results. In *MuTra 2006 Audio Visual Translation Scenarios: Conference proceedings*. 148–181.
[94] W3C. 2022. Descriptions of Visual Information. https://www.w3.org/WAI/media/av/description/. Accessed: 2022-11-6.
[95] W3C. 2023. Extended Audio Description (Prerecorded) (Level AAA). https://www.w3.org/TR/WCAG20-TECHS/G8.html.. Accessed: 2023-04-09.
[96] W3C. 2024. Extended Audio Description (Prerecorded): Understanding SC 1.2.7. https://www.w3.org/TR/UNDERSTANDING-WCAG20/media-equiv-extended-ad.html. Accessed: 2024-03-14.
[97] W3C Web Accessibility Initiative (WAI). 2016. Web Accessibility Perspectives: Customizable Text - Audio Described Version. https://youtu.be/L4WLeVc5l5k. Accessed: 2023-11-6.
[98] W3C Web Accessibility Initiative (WAI). 2016. Web Accessibility Perspectives: Video Captions - Audio Described Version. https://youtu.be/4qIordU8vT8. Accessed: 2023-11-6.
[99] Agnieszka Walczak and Louise Fryer. 2018. Vocal delivery of audio description by genre: measuring users' presence. *Perspectives* 26, 1 (2018), 69–83.
[100] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 277, 12 pages. https://doi.org/10.1145/3411764.3445347
[101] Chunlei Wu, Camilo Orozco, Jason Boyer, Marc Leglise, James Goodale, Serge Batalov, Christopher L Hodge, James Haase, Jeff Janes, Jon W Huss, et al. 2009. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome biology* 10, 11 (2009), 1–8. https://doi.org/10.1186/gb-2009-10-11-r130
[102] Ting Wu, Junjie Peng, Wenqiang Zhang, Huiran Zhang, Shuhua Tan, Fen Yi, Chuanshuai Ma, and Yansong Huang. 2022. Video sentiment analysis with bimodal information-augmented multi-head attention. *Knowledge-Based Systems* 235 (2022), 107676. https://doi.org/10.1016/j.knosys.2021.107676
[103] Sumit K Yadav, Mayank Bhushan, and Swati Gupta. 2015. Multimodal sentiment analysis: Sentiment analysis using audiovisual format. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. 1415–1419.
[104] UTKATA Office Chair Yoga. 2020. 1 Minute Office Chair Yoga - Yoga at your Desk - Flow #1. https://youtu.be/vgf21Tqfwwg. Accessed: 2023-11-6.
[105] YouDescribe. 2020. YouDescribe Audio Description Guideline. https://youdescribe.org/support/tutorial. Accessed: 2020-11-6.
[106] YouTube. 2023. YouTube. http://www.youtube.com. Accessed: 2023-09-14.
[107] Beste F. Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A. Miele, and Ilmi Yoon. 2020. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 47–60. https://doi.org/10.1145/3357236.3395433
[108] 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. 2020. 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* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382821