“It’s Kind of Context Dependent”: Understanding Blind and Low Vision People’s Video Accessibility Preferences Across Viewing Scenarios

Lucy Jiang
Cornell University
Ithaca, NY, USA
lucjia@cs.cornell.edu

Crescentia Jung
Cornell Tech
New York, NY, USA
cj382@cornell.edu

Mahika Phutane
Cornell University
Ithaca, NY, USA
mahika@cs.cornell.edu

Abigale Stangl
Georgia Institute of Technology
Atlanta, GA, USA
abigale.stangl@design.gatech.edu

Shiri Azenkot
Cornell Tech
New York, NY, USA
shiri.azenkot@cornell.edu

ABSTRACT

While audio description (AD) is the standard approach for making videos accessible to blind and low vision (BLV) people, existing AD guidelines do not consider BLV users’ varied preferences across viewing scenarios. These scenarios range from how-to videos on YouTube, where users seek to learn new skills, to historical dramas on Netflix, where a user’s goal is entertainment. Additionally, the increase in video watching on mobile devices provides an opportunity to integrate nonverbal output modalities (e.g., audio cues, tactile elements, and visual enhancements). Through a formative survey and 15 semi-structured interviews, we identified BLV people’s video accessibility preferences across diverse scenarios. For example, participants valued action and equipment details for how-to videos, tactile graphics for learning scenarios, and 3D models for fantastical content. We define a six-dimensional video accessibility design space to guide future innovation and discuss how to move from “one-size-fits-all” paradigms to scenario-specific approaches.

CCS CONCEPTS

• **Human-centered computing → Accessibility**.

KEYWORDS

video accessibility, audio description, blind, low vision, scenarios, audio cue, tactile feedback, context-aware, scenario-based design

ACM Reference Format:
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 (CHI ’24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3613904.3642238

CHI ’24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642238

1 INTRODUCTION

As videos become more widespread and stylistically diverse, ranging from documentaries on Netflix to 60-second videos on Instagram, blind and low vision (BLV) people remain excluded from engaging with this growing variety and volume of visual content. Currently, the standard method of making videos accessible is adding audio description (AD), a separate audio track with narration of visual elements [34, 108, 119]. AD is typically created for high-budget content such as movies and TV shows. However, many videos still lack AD, and it is unclear to what extent existing accessibility practices can or should be applied to newer video formats such as short-form video. As user-generated content increases in popularity on platforms like YouTube and TikTok [41, 74, 133], the need to make videos of all types accessible to BLV people grows.

To support video accessibility, researchers and practitioners have aimed to increase the *quantity* of described videos through crowdsourcing platforms and automation (e.g., [6, 11, 61, 84, 126, 132]). Others have focused on improving the *quality* of AD through providing authorship guidelines to help determine what content to include or which tone of voice to use (e.g., [3, 26–28, 58, 59, 117]). However, limited research has explored how nonverbal techniques, which include visual, audio, and tactile enhancements [59, 99], can support video accessibility in a holistic manner.

Furthermore, no work has systematically considered the evolving ways in which people consume videos today. Over the last two decades, video consumption has shifted from only watching on large screens to frequent watching on mobile devices [87, 97]. Currently, people often watch videos on mobile devices or computers and use platforms that allow them to access new content at an unprecedented rate. Video types have also become more diverse — videos can range from 5-minute how-to videos on YouTube, to hour-long documentaries on Netflix, to comedic 30-second short videos on TikTok. Users’ goals for watching these different videos include learning how to do something, being entertained, or keeping up with family and friends.

We conceptualize these video watching contexts as ***viewing scenarios***, which encapsulate (1) video types (e.g., how-to, comedy, music video), (2) viewing platforms (e.g., streaming services, video sharing sites, social networking sites), and (3) users’ information goals (e.g., to learn, to be entertained, to engage with friends). In other words, a scenario is the story of what video is being watched, where a user found the video, and why a user is watching the video. Researchers have established that BLV people's accessibility needs vary for different image viewing scenarios [114]; however, few have explored the nuances of accessibility for the large variety of videos produced today. Thus, there is a gap in understanding how BLV people's video accessibility needs vary across different scenarios.

To address this gap, we investigate the following research question: **What are BLV people's needs and preferences for video accessibility across viewing scenarios?** We considered various approaches to video accessibility, including the content and presentation of AD and augmented output modalities, to build on prior work [59]. We conducted a formative survey with 101 respondents and interviewed 15 BLV participants. Interviews included a discussion of recently watched videos and a co-watching session encompassing multiple video scenarios. Throughout the interview, we probed about current access needs and brainstormed ideas for holistically enhancing video accessibility.

We found that BLV user needs and preferences for video accessibility varied across viewing scenarios. For example, participants wished to know details about actions and equipment to help with learning how to do something from how-to videos on YouTube. In contrast, when watching short-form videos on Instagram or Facebook to engage with friends, participants placed more emphasis on subjects, actions, clothing, and settings. Participants' desired output modalities ranged from standardizing audio cues for indicating scene changes to using physical and tangible 3D models for conceptualizing fantastical character designs. Across scenarios, we also identified that video types strongly correlated with platforms used and users' goals. Based on our findings, we present six dimensions in the video accessibility design space: level of detail, alteration of video time, level of augmentation, modality of presentation, synchronicity of accessible content, and tone and style of approach. We also consider the efficacy and ethical implications of using generative AI to support video accessibility for a subset of dimensions in our design space.

In summary, we (1) contribute novel insights on the variety and specificity of BLV users' preferences for video accessibility across diverse *viewing scenarios*, (2) articulate a six-dimensional design space for holistic video accessibility, and (3) consider how advancements in AI technology intersect with video and content accessibility efforts. Our work challenges the existing paradigm of "one-size-fits-all" audio descriptions. We intend for our design space to serve as a valuable resource for the accessibility community as videos, media, and technology continue to evolve.

2 RELATED WORK

Our work builds on prior work in image and video accessibility, specifically regarding BLV users' visual description preferences and personalized access solutions.

2.1 Image Accessibility

We first draw on image description literature to foreground our understanding of video descriptions. Image description guidelines (e.g., [17, 44, 71, 75, 94, 123]) instruct description creators to give information about an image in relation to its context [17, 94] and describe the predominant content (e.g., objects, people, text, scenery) to aid understanding [10, 71, 111]. Researchers have also investigated creating descriptions through human-authored [5, 9, 45, 105, 123] and AI-supported methods [40, 100, 101].

To augment BLV users' image experiences, researchers have explored methods for making images accessible beyond static textual descriptions. For example, Morris et al. [76] examined approaches for "rich" image descriptions that support interactive image accessibility, while others built systems to enable touch-based image exploration [65, 81]. Researchers have also studied the efficacy of using music, earcons, and tactile elements for the accessibility of artwork [14, 15, 96], museum experiences [2, 69], and data visualizations [8, 103, 106]. Others have researched the nuances of image consumption and visual descriptions on popular social media websites, including Twitter and Facebook [37–40, 72, 77, 124, 130]. They found that nonverbal cues, such as short sounds to indicate the repetition of a meme format [39], could aid in image accessibility while preserving emotion and tone.

Prior work has also emphasized that BLV people's preferences for image descriptions vary based on an image's context [17]. Specifically, preferences differ based on the source or content of the image [4, 21, 49, 60, 63, 78, 79, 111, 114]. For example, Stangl et al. [111, 114] found that BLV people wanted different details for images associated with different sources and user goals. Some details, such as attributes of the primary object in the image, were desired across a variety of scenarios. While some have used artificial intelligence to generate descriptions with different linguistic attributes, such as personality [101, 104] or writing style [21, 35], these investigations often do not consider how a BLV user's current context may influence their photo-viewing experience or their description presentation preferences.

In this paper, we draw on Stangl et al.'s definition of scenarios [114] to inform our study design. We focus on contextual factors that affect BLV users' preferences when consuming *video* content and diverge from current "one-size-fits-all" models of video description to design more personalized video accessibility experiences.

2.2 Video Accessibility

Audio description and video accessibility practices are guided by BLV innovators [118], advocacy organizations serving BLV people [3, 26–28, 117], and industry practitioners [34, 73, 86]. Although a video can be conceptualized as a sequence of frames, the process of AD authorship involves additional intricacies beyond simply linking image descriptions together. Audio description requires creators to understand the video's broader context, distill meaning from multiple frames, and insert descriptions at appropriate times within the auditory and visual narrative of the video [93, 135].

Some prior work in HCI has specifically explored BLV people's audio description detail preferences. In investigations of ViScene, a collaborative AD authoring system that employed novices to increase AD availability, Natalie et al. [82–84] reported that BLV people valued details about clothing, time of day, and location. Additionally, through evaluations of an automated AD system, Wang et al. [126] found that BLV users preferred different details for different video types — for example, they wished to have more detailed descriptions of people in a comedy but not in a DIY video.

Furthermore, Jiang et al. [58] found that BLV AD writers focused specifically on character descriptions (e.g., race, age), background settings, action descriptions, and clarifying audio cues, given their experience as both creators and consumers.

Additional insights about which details, and levels of detail, that BLV users want in AD have emerged through the development of tools to streamline the AD creation process [11, 61, 66, 93, 126, 128, 134]. For example, Yuksel et al. [134] found that BLV participants had preferences for description styles and content for cooking videos, such as precise directions and accurate measurements. Pavel et al. [93] developed Rescribe, an AI-supported AD tool, to investigate the viability of extended-inline AD, a new description format that looped the video's audio beneath narration while the video's visuals were paused. While prior works identify general guidelines for overall AD quality, we examine how BLV users' desired details may vary for different types of videos.

Other research has shown that BLV people wish to interact and engage with video content in ways other than only listening to preset neutral descriptions during the video itself. For example, Stangl et al. [110] and Bodi et al. [6] investigated the viability of providing video access through interactive visual question answering, reinforcing the importance of BLV users having agency in the process of making videos accessible. Others explored the impact of changing the tone or style of verbal descriptions for select video types, finding that alternative AD styles were engaging for BLV users [30, 59, 121, 125]. Additionally, Romero-Fresco et al. [98] conducted a preliminary study on the efficacy of audio introductions for providing additional detail to described films. They identified that BLV participants were in favor of accessing descriptions prior to watching a movie to improve their understanding about the characters, settings, and visual style of the film.

Prior work has also explored using different modalities to make videos accessible. In a study with BLV AD users, sighted AD creators, and BLV AD creators, Jiang et al. [59] identified key video accessibility considerations posed by BLV AD creators due to their unique intersection of experiences. The authors found that the linguistic and aural presentation of AD, sound design, and multisensory aspects contributed towards greater immersion for BLV viewers. Sackl et al. [99] also explored how visual enhancements such as contrast adaptation, color manipulation, and sharpness adjustments could improve video accessibility for BLV users. Others evaluated how spatial audio could augment sports broadcasts [54, 55], short films [67], and 360° videos [18]. We build on existing research to concretely consider how video accessibility preferences can differ across scenarios and understand how to leverage nonverbal output modalities to best suit user needs.

While most prior research on video accessibility has focused on creating universally satisfactory descriptions, some have investigated the diversity of BLV users' preferences and information needs for the same video [19, 20, 59, 126]. For example, Chmiel and Mazur [20] examined AD preference differences between people with different vision levels and experiences with vision loss onset. They concluded that "middle-of-the-road" solutions could support the AD information needs of most users, but recommended future research to study how different levels of detail for AD could best suit individual preferences. Additionally, individual interests in spatialization and multisensory interactions can vary depending on users' abilities (e.g., spatialized audio may not be accessible for d/Deaf and hard of hearing users) and vision levels (e.g., low vision people may not want as much AD detail) [59]. Despite some progress in understanding BLV users' preferences, most work in this area does not include short-form video content [87] or consider how nuances from users' unique video watching *scenarios* may give rise to different description needs [20, 84].

Building on prior work, we explore more holistic approaches for video accessibility to address BLV users' preferences across contexts. In this study, we focus on how a user's *scenario* impacts their video watching experiences.

3 FORMATIVE SURVEY

Before beginning our interview study, we conducted a survey to gain a preliminary understanding of BLV people's video watching behaviors and inform our interview protocol. We also used the survey as a recruiting tool for the interview.

3.1 Survey Methodology

We recruited participants through social media postings, mailing lists, and snowball sampling. Our recruitment notice indicated that participants had to be 18 years old or older, identify as blind or low vision, and have experience regularly watching online videos, which we defined as using a video platform such as a social networking service, a video sharing site, or a streaming service at least twice a week. A total of 101 respondents completed our survey, including 40 who identified as blind and 45 who identified as low vision. We discarded data from 16 people who identified as sighted. For every survey response, we donated $2.50 to the National Federation of the Blind and $2.50 to the American Council of the Blind. Our survey procedure was approved by the Institutional Review Board (IRB) at our university.

Our survey instrument, provided in Supplementary Materials, included both multiple choice and long answer questions and was designed to take about 10 minutes to complete. We asked respondents about which platforms they used to watch video content (e.g., Netflix, TikTok, YouTube, news sites), types of videos they watched (e.g., informational / educational, comedic, how-to, videos from friends or family), and how often they used description services (e.g., only watch videos with AD, use AD whenever available, only for certain types of videos). We also asked how video accessibility could be improved overall, what types of videos would be most useful to have AD, and if there were specific types of videos they would like to watch but are currently inaccessible.

To analyze our survey responses, we calculated the frequencies of video types viewed and platforms used by BLV participants. The first author analyzed open-ended responses and identified common video types and platforms, video accessibility ideas, AD usage patterns, and useful situations for accessible videos.

3.2 Survey Findings and Discussion

According to the survey, the five most popular video types were: informational / educational (77.6% of respondents), comedic (63.5%), how-to / DIY videos (57.6%), lifestyle (52.9%), and news / commentary (52.9%). For the three categories of platforms, the most used video sharing site was YouTube (83.5%), the most used streaming

Table 1: BLV survey respondents' commonly viewed video types and commonly used video platforms. # represents the number of responses and % represents the proportion of responses (of 85 total BLV respondents). VSS = video sharing site, SS = streaming service, and SNS = social networking site.

Video Type	#	%
Informational / educational	66	77.6
Comedic	54	63.5
How-to / do it yourself	49	57.6
Lifestyle	45	52.9
News / commentary	45	52.9
Science fiction / fantasy	44	51.8
Thriller / horror	41	48.2
Action / adventure	41	48.2
Music videos	36	42.4
Videos from friends or family	36	42.4
Romance	34	40.0
Animation	33	38.8
Other (sports, religious, etc.)	26	30.6

Category	Video Platform	#	%
VSS	YouTube	71	83.5
SS	Netflix	63	74.1
	Disney+	33	38.8
	Amazon Prime	31	36.5
	Hulu	30	35.3
	HBO Max	25	29.4
SNS	Facebook	50	58.8
	TikTok	42	49.4
	Instagram	40	47.1
	WhatsApp	20	23.5
	Reddit	12	14.1
Other	News sites	19	22.4

service was Netflix (74.1%), and the most used social networking site was Facebook (58.8%). Aggregated responses for video type and platform are presented in Table 1.

Most participants used multiple devices to watch videos. 90.6% of respondents reported using mobile phones, 71.8% reported using computers, and 64.7% reported using televisions. 4.7% used tablets, and one used a projector.

Survey respondents most frequently mentioned the following video types as ones that acutely required better video accessibility measures: how-to, informational / educational, action, content reliant on visuals (e.g., infographics, maps, or visual effects), and news and weather. Some also wanted more accessible music videos, foreign language videos, videos with minimal dialogue or only music in the background, sports, vlogs, and live events such as theater productions and concerts.

With regards to frequency of watching videos with AD, 31.8% of participants used AD "whenever available," and 11.8% of participants selected that they "only watch videos with audio description" and use description "whenever they are available." For example, one respondent explained their reasoning: *"I will sometimes watch movies or TV shows [without AD] because I have a cited [sic] partner who can assist me, but for genres like a horror [sic], we will not watch a movie if audio description is not available."*

However, others did not use AD as often. 14.1% of participants reported that they only used AD for certain types of videos, in certain situations, or on certain platforms. For instance, another respondent used AD for movies, but not for *"concerts because I don't like the audio description talking in the middle of a song [or for] standup comedy because it is hard to hear the comedian talking during the audio description track."*

Due to our relatively limited sample size, we do not draw any specific conclusions about BLV people's video watching habits and note that further investigation is warranted. However, our survey did allow us to gain insight into the breadth of videos that BLV people watch. We utilized these survey results, specifically the types of videos that they wished could be more accessible, to guide our interview protocol.

4 METHOD

To delve deeper into BLV people's preferences for video accessibility across viewing scenarios, we conducted semi-structured interviews with a subset of our survey respondents. During the interviews, we asked participants to expand upon their survey responses, probed about prior experiences with watching accessible and inaccessible videos, and engaged participants in a video co-watching session to brainstorm video accessibility ideas.

4.1 Participants

We invited 15 BLV survey respondents to participate in our interview study. We selected participants who were at least 18 years old, identified as blind or low vision, were comfortable communicating in English, and had experience regularly watching online videos as per our survey inclusion criteria.

All participants identified as blind (none identified as low vision), but had varying degrees of residual vision. For example, some participants had light perception and could read with magnification, others did not have light perception, and one participant was born sighted and started experiencing vision loss in his 30s. Their ages ranged from 24 to 62 (mean 39.9, SD 11.3). Nine participants identified as women, five identified as men, and one identified as agender. Some participants had extensive experience with AD: one was a hobbyist blind film critic, another worked as an AD consultant and advisor, and yet another was an extended reality sound and media artist with AD creation and production experience. All participants used screen readers and five regularly used Braille displays.

Participant demographic details are presented in Table 2. Our interview protocol was approved by our university's IRB.

Table 2: We present participant pseudonyms and demographics, including gender and ethnicity in participants’ own words. All participants identified as blind; here, we paraphrase their self-disclosed vision details. We also indicate the platforms they use for video watching, their AD usage (e.g., only in certain situations, whenever they are available, only watching videos with AD), and which of the three common scenarios they watched.

Pseudonym	Age / Gender	Ethnicity	Vision Details	Platforms	AD Usage	Video
Alice	30 / Female	Chinese American	Completely blind, born legally blind and lost more vision in 20s	SS, VSS, SNS	Situationally	V1
Blair	62 / Female	Caucasian	Born with low vision, now have light perception, limited peripheral vision	SS, VSS	Whenever available	V1
Colin	36 / Male	White	Born with low vision and lost remaining sight at 19	SS, VSS, SNS	Whenever available	V1
Diana	34 / Female	Asian / Pacific Islander	Blind since birth, sees shapes and colors, can read text with significant magnification	SS, VSS, SNS	Whenever available	V1
Emily	29 / Female	White	Has light perception, born blind	SS, VSS, SNS	Whenever available	V1
Felix	40 / Male	Caucasian	No central vision, pockets of peripheral vision, born sighted and started losing sight at 34	SS, VSS	Only watch with AD	V2
Grace	29 / Female	White	Left eye nothing, can read text with significant magnification	SS, VSS, SNS	Whenever available	V2
Haley	40 / Female	Puerto Rican	Only see color and movement, born with vision but lost gradually	SS, VSS	Whenever available	V2
Isaac	48 / Male	White	Blurry tunnel vision, can read large print, gradual vision loss	SS, VSS, SNS	Situationally	V2
Julia	59 / Female	White	Has light perception, no color, born with low vision, experienced significant vision loss at 28	SS, VSS, SNS	Whenever available	V2
Karla	24 / Female	Hispanic	Completely blind (no light perception), born blind	SS, VSS	Whenever available	V3
Layne	41 / Agender	White	One eye with no vision, other eye has no peripheral vision or depth perception	SS, VSS, SNS	Whenever available	V3
Mason	38 / Male	White	Completely blind (no light perception), born blind	SS	Only watch with AD	V3
Nicki	34 / Female	Hispanic	Has light perception, cannot see shadows, lost vision over time	SS, VSS, SNS	Whenever available	V3
Oscar	54 / Male	Caucasian	Born legally blind, gradually lost peripheral vision starting at 30	SS, VSS, SNS	Whenever available	V3

4.2 Procedure

Our study included a virtual 75-minute interview session, conducted via Zoom. The interviews consisted of three parts: a review of the participant’s survey responses, a discussion about several recently viewed videos, and a video co-watching activity.

We first asked participants demographic questions, then reviewed their survey responses and probed about interesting comments. For example, we asked participants to expand on their rationale for only using AD in certain situations or on certain platforms.

During the second part of the interview, we asked participants to recall one or more specific videos they had watched in the last few weeks. This recent critical incident approach [32] allowed participants to reflect on concrete instances and provide richer, more specific insights than when speaking about general behaviors. For each video discussed, we determined the participant’s viewing scenario and asked them about positive and negative aspects of their experience. We then probed participants to explore what could make these videos more accessible, eliciting creative ideas not yet possible with current technology. Our inquiry was based on multisensory interactions for videos proposed by Jiang et al. [59] and Sackl et al. [99]. Referring to the specific video and scenario, we asked questions such as:

- How could the video overall have been made accessible, given that you are watching it in [this scenario]?
- In what other ways would you like to have descriptions of the video?
- Thinking about audio or sound effects more generally, what additional audio could help make the video more accessible?
- What visual enhancements, if any, could help make the visuals more accessible?
- What tactile feedback, if any, could help make the video more accessible?

The third part of the interview was the video co-watching portion, during which we presented participants with multiple videos to encourage them to compare and contrast their preferences *across different scenarios*. To ensure that scenarios were relatable to participants and avoid a limitation of a similar study [114], we developed naturalistic viewing scenarios based on our survey findings. The first scenario was randomly selected from three **common scenarios**, which consisted of a nature documentary, a comedy sketch, and a short-form video (more details provided in Table 3). Although short-form videos were not as commonly reported in our survey, we included this video type to capture participants' thoughts on an unfamiliar but emerging scenario.

Next, we presented participants with one to two **participant-specific scenarios**. We pre-selected scenarios with video types that participants indicated they would like to watch in their survey responses. Our aim was to gather participants' perspectives on how to improve the accessibility of videos that did not have any existing access measures; as such, none of the videos we presented to participants had AD. Since scenarios were assigned based on participant preferences, some video types had more responses (e.g., six participants co-watched various how-to videos while only one co-watched a foreign language film clip). However, this method allowed us to choose scenarios specific to participants' unique viewing interests, which helped with elucidating current frustrations and brainstorming future accessibility measures. Three examples of participant-specific scenarios can be found in Table 4 and a full list is provided in Appendix A.

During this part of the study, the interviewer shared their screen and audio so that participants, regardless of familiarity with the interview platform, did not have to navigate potentially inaccessible user interfaces. Prior to playing each video, we presented the participant with the scenario as follows: *"Imagine you were watching a video such as [video type] on [platform]. Your goal of watching this video is [user goal]."* We played 60-90 seconds of each video, which we found in our pilot sessions to be a sufficient length of time to elicit meaningful feedback while avoiding fatigue. Participants were instructed to say "pause" if they had questions or comments about the video, but most chose not to while watching.

Each time the participant paused the video, we invited them to share their thoughts and asked about the accessibility of the video with questions such as:

- What do you think just happened in the video?
  - If participants had simple questions or misconceptions, we clarified by concisely providing visual information, based on guidance in prior AD work [58].
  - We asked participants about what they thought happened in the video before providing clarifications, as this helped identify mismatches between participants' inferences and the video's actual visuals.
- Was anything confusing or unclear?
- How accessible was the video so far?

Once each video was finished, we asked participants about its overall accessibility and other ways to make it more holistically accessible, using the same questions presented during the second part of the interview.

Participants were compensated with a $30 gift card for their time and contributions.

4.3 Data Analysis

We audio recorded and transcribed all interviews. Three researchers analyzed the data using inductive coding. The first author individually coded two transcripts and two other authors each individually coded one of the two transcripts. Then, the authors discussed code discrepancies, developed a codebook, and split up the remaining interview transcripts for coding. After coding, three authors performed thematic analysis [7] on our interview transcripts to identify overarching themes and patterns across video watching scenarios.

4.4 Positionality

Members of the research team identify as sighted and low vision, and have varying degrees of experience with using AD in their everyday lives. The first author, who conducted all interviews and spearheaded analysis, is a sighted person who has experience as an amateur AD creator and frequently uses AD when it is available.

5 FINDINGS

Our findings are presented in terms of scenarios that participants encountered. The scenarios are anchored in video types, as we identified that video types were strongly correlated with the platform used and a user's goals. We briefly begin each section with common barriers to watching videos in each scenario. Then, we describe the specific details, levels of detail, and output modalities that participants found helpful for accessibility. Lastly, we highlight similarities that persisted across all scenarios.

5.1 How-To Video: Learning How to Do Something on Video Sharing Sites

Participants frequently watched how-to videos with the goal of learning how to do something (N = 9). However, seven mentioned the lack of detailed narration as a frustrating accessibility barrier. For instance, using demonstrative pronouns (such as "this") during narration was a primary source of confusion. Diana, who often watched knitting how-to videos, needed to search for videos that *"feature[d] the person actually saying what they're doing, as opposed to just being like, 'And then you go like this.'"* Videos were also inaccessible when they only played music and did not include any narration at all. Karla recounted her frustrations with coming across how-to videos with background music and text on screen: *"that just sounds like music to me."*

Table 3: The three common video probes and further details, including each video's scenario (type, platform, and user goal) and our rationale for choosing the videos.

		V1: Nature Documentary	V2: Comedy Sketch	V3: Short-Form Video
Scenario	Type	Informational / educational	Comedic	Lifestyle
	Platform	Streaming service	Video sharing site	Social networking service
	Goal	Learning about a concept	Entertainment	Engaging with others
Details	Synopsis	Aerial and close-up footage of wooded forests and forest floors, with narration	A man engages in a conversation with a woman at a cafe to try to guess her age	From a hotel room balcony in Paris, an influencer waves to tourists gathered outside
	Rationale	Familiar video format, text on screen, sparse narration, cinematic shots	Familiar video format, text on screen, multiple characters, visual gags	Unfamiliar video format, text on screen, no dialogue, sound effects
	Source	Netflix [85]	YouTube [95]	Instagram [68]
	Length	90 seconds	90 seconds	11 seconds

Table 4: Three of our 17 unique participant-specific video probes. Videos were hosted on YouTube but represented different scenarios and platforms. The full list of participant-specific videos is provided in Appendix A.

		Tennis Match	Cooking Tutorial	Pop Music Video
Scenario	Type	Sports	How-to	Music video
	Platform	Streaming service	Video sharing site	Video sharing site
	Goal	Entertainment	Learn how to do something	Entertainment
Details	Synopsis	A highlight reel of a professional women's tennis match	A video explaining how to cook four different meals	Music video for *You Belong With Me* by Taylor Swift
	Rationale	Minimal narration / sound	Text on screen, no narration	No dialogue, only the song

5.1.1 Details about Actions and Equipment. Ten participants explained that providing more details would help with learning how to do something. Five participants specifically mentioned that actions should be explained in *"excruciatingly painful detail"* (Layne). Grace, who often watched cooking how-to videos, emphasized that such details were critical for success: *"[if] this is a recipe that I would hope to replicate, I'm going to want those step-by-step very detailed instructions."* As a former chef, Blair also enjoyed watching cooking videos. She preferred cooking videos that presented detailed information in a "technical" structure, which resembled her formal training: *"[the chef] will give you your ingredients, then your method and your technique. And it's all just very logical"* (Blair). For DIY videos, Colin gave a hypothetical description which identified actions and specific corners of a piece of furniture instead of using vague instructions such as "this" or "that": *"If they are nailing nails into something, [the AD should say,] 'He gets out three nails for this project, and nails them in corner A, corner B, and corner C.'"* Mason also wanted to know more about which ingredients and cooking utensils to use to confidently replicate a recipe.

5.1.2 Output Modalities. Participants suggested various techniques to improve the accessibility of how-to videos; for example, four participants mentioned that a separate resource with additional information would be helpful. Nicki wanted to know more about products used in how-to videos, and found it helpful to include this information through narration, the video description, or comments. Similarly, when watching a home exercise tutorial, Karla thought it would be valuable to have a separate resource, such as a website or transcript with a list of the workout moves in the video.

> "Having a lot more audio and textual feedback would be helpful... Having a link that you can click, or a list of different workout stuff that they're going to do, gives you some time to prep and be like, 'Okay, now I know what I'm doing,' so I can work out along with the video." (Karla, 24F)

Three participants also suggested that audio cues could be added to indicate a change in instruction (Karla) or a timer (Mason) to know when one step was complete and another was starting. However, Mason explained that while audio cues could be useful, their meaning needed to be properly explained: *"I'd have to know what the sound is for and why. [If it's] just a random sound, I'm just going to go, 'That's weird,' and ignore it."* Emily was also interested in maintaining diegetic audio in how-to videos (e.g., the sound of an electric mixer in a baking video), as the sounds could help her determine if the task necessitated electric tools and estimate approximately how long they would need to be used.

Participants thought that having tactile information could help with their understanding as well. Mason often watched how-to videos to learn how to fix items around his home, such as a hot water heater, and wished to simultaneously read a description of the actions on a Braille display while watching the video. Additionally, Karla watched how-to videos on weaving and thought having tactile graphics throughout different stages of completion would be helpful: *"having a picture of the finished product, or of the product as it's going through [the steps] might actually be helpful."* However, Grace cautioned against unexplained tactile cues, mentioning that a vibration would be confusing *"because [she] would think [her] phone was ringing."*

5.2 Informational and Educational Video: Learning a New Concept on a Streaming Service or Video Sharing Site

Informational and educational videos were present in eight participants' viewing rotations. Alice, a student and hobbyist viewer of psychology lectures, mentioned that vague references to visual aids excluded her from forming a full understanding of educational content. Layne, who frequently watched philosophy video essays on YouTube, thought that the extensive narration common in educational videos left limited time to insert inline AD, and cautioned against descriptions that were more *"disruptive"* than helpful.

5.2.1 Details about Visual Aids, Settings, and Subjects. Participants desired details that would help them better conceptualize abstract information. Six found that the narration and lecturing inherent to most informational videos was *"accessible to a point"* (Blair); however, to support their understanding, they wished to have more information about visual aids or graphics, text on screen, settings, and subjects. Isaac described how he relied on a separate app to improve his access to infographics: *"if I'm watching [a video] on my phone and it [has] an infographic, I'll pause it, I'll put on image recognition ... and try to see if it'll recognize any of the text, and then try to fill in the blanks that way."*

During the co-watching sessions, participants' questions during and after watching the video uncovered varied description needs. For documentaries, the setting was important to participants. When asked about what happened in the video probe, Colin mentioned how the narration helped him *"get the context that it's large trees in the dark environment"* but wished to know *"the finer details, like how large the trees might be."* Blair was also curious about *"what kind of trees"* were being shown, and did not want details about clothing unless they were relevant to the educational aim of the video. Emily highlighted the value of specific descriptions for learning more about species in a documentary with the following example:

> "Instead of just saying 'a blue bird,' maybe say its size and beak size and wingspan... If the documentary [showed] the differences between male and female birds, a sighted person who watched that five minutes ago could now tell that a male bird and a female bird are flying toward each other. But if the describer just said 'two birds flying towards each other,' that's not going to work." (Emily, 29F)

5.2.2 Output Modalities. As with other scenarios, six participants were open to adding output modalities to supplement their video watching experience.

Participants thought audio cues would help with video comprehension. They appreciated that ambient sounds in the video's original audio could convey the mood and context of the documentary, which gave them *"more of a feel for the environment"* (Colin). Colin was also open to learning more through *"an audio track or an alternate link to click to for more information."* Three participants proposed additions to the soundscape, suggesting that scene changes could be cued with a sound effect that was distinct from other noises in the scene, so as not to *"blend in with the birds of the movie"* (Emily).

Three participants specifically mentioned tactile graphics as a helpful tool for understanding concepts such as scale or structure. For example, Colin wished to have *"[embossed] pictures of those forests"* from the documentary, while Grace thought *"having tactile versions of the grid or the map would be super cool."* However, others found tactile graphics to be hard to interpret, and instead preferred 3D models to indicate what *"the animals, or... some of the rocks, or what the soil would feel like"* (Emily). While participants differed on which tactile modalities they wanted, they agreed that tactile elements were broadly helpful for learning.

> "It would be cool to have it be more tactile because that's how you learn. Descriptions might be great, but describing a part of a cell is not as good as seeing a picture or feeling it more hands-on." (Alice, 30F)

5.3 Short-Form Video: Engaging with Friends and Pop Culture on a Social Networking Site

Not all participants were active social media users, almost always due to the inaccessibility of platforms like Instagram and TikTok. All participants were aware of the short-form video format, and five were frequent viewers. Seven participants noted that the prevalence of text on screen (e.g., a product review video with product details listed as text on screen) was a major barrier to understanding and engaging with Instagram Reels, TikToks, and Instagram Stories.

5.3.1 Details about Subjects, Actions, Clothing, and Settings. Five participants were primarily interested in knowing more details pertaining to subjects (both people and pets) and their actions. Common questions that arose when discussing short-form videos included "characters" and their actions. For example, when watching an animal video, participants expressed that they at least wanted to know, *"What is the animal? And what are they doing?"* (Layne).

While this information about "who" and "what" was critical, three participants wanted details about clothing and settings. For example, Karla felt that she required *"a full picture of the surroundings and the clothing"* to completely understand the context, given the short video duration. As someone who was unfamiliar with this video format, Mason also emphasized the importance of these details for understanding the short-form common scenario (V3).

5.3.2 Output Modalities. Seven participants wished to access additional detail or context in the video caption or through an external resource. Though AD is typically synchronous with a video, some wished to know more *before* watching. Five participants often used the video caption (a text description posted by the video's original creator) as a tool that could be referenced to *"put [the video] into context"* (Layne), helping viewers infer the overall tone of the piece.

Karla explained how even short captions helped her form an idea of what a video may involve:

> "[If] the caption is 'girls' night,' you can kind of guess what's happening — you're probably going to go out shopping or stay in and watch a movie... If the caption is 'When girls' night goes terribly wrong,' you can assume that they were relaxing and then something happened." (Karla, 24F)

However, others such as Grace reported a lack of connection between the caption and the actual content of the video. She mentioned that *"[creators] don't actually say what happens in the video... the caption might be like, 'Oh, I was so shocked' or a bunch of hashtags"* (Grace). In these cases, she was frustrated about taking extra time to read a caption that did not provide the context she wanted.

Audio, such as background music, could also improve accessibility. Karla and Nicki mentioned that the practice of using popular "sounds" as templates for short-form videos helped their understanding. As with captions, they found that *"certain music... could tell a little bit about the video"* (Nicki), including the video's tone and content. Karla explained how familiar music helped her infer the tone of TikToks:

> "If people use [music] in the right context, some people will put the 'oh no' song... because there's something that makes you go, 'Oh no.' ... Sometimes if it's more of a slower piano beat, I'm like, 'Okay, it's probably something sad, or it's something serious.' But if it's really fast, kind of a bouncy type thing, I'm like, 'Okay, it's probably something intended to be a little bit lighter.'" (Karla, 24F)

For short-form videos, three participants generally thought descriptions were more valuable than additional output modalities for providing context. However, some thought having *"different audio cues for different environments"* (Karla) could reduce the amount of information conveyed verbally. Oscar also shared how audio cues could help with his frustration about not knowing when a short-form video was automatically replaying. Tactile cues were also useful; for dancing videos on TikTok, Karla suggested having *"a pattern on the Braille display that moves along with the dancing [movements]"* to better convey the action in a nonvisual way.

5.4 Music Video: Seeking Entertainment on a Video Sharing Site

> "I've heard this song many times... but I've never really known what happens exactly in the video. How does it start? ... Who's in the video? ... What's happening in the story?" (Nicki, 34F)

Watching music videos for entertainment was a unique scenario, as participants were often familiar with the music itself but were excluded from enjoying the stories in the accompanying videos. In fact, two participants shared that they were uncertain if music videos contained visuals at all: *"all I'm hearing, as a blind person who can't really tell what's happening on the screen, is the music. It's almost like I'm just playing the song without any context"* (Nicki). Four other participants attempted to infer action based on the video's audio or comments, often with limited success.

5.4.1 Details about People, Actions, Settings, Clothing, and Visual Effects. Especially for familiar songs, participants were interested in having access to details that gave them an idea of the story presented through the video. Five participants primarily wished to know who was present in the music video and what they were doing. For example, Nicki mentioned how the song lyrics in Taylor Swift's song, *You Belong With Me*, suggested that the video could contain certain characters (e.g., a cheerleader). However, she was curious about which other characters were in the scene and wished to know more about what the characters were wearing, as *"sometimes what people are wearing can tell you a little bit more about the context"* (Nicki). Similarly, Blair wanted to have enough information to *"use [her] imagination and make pictures for [herself]."*

> "I want to know what the outfits are. I want to know the dancing or the setting, the scenery. I want them to set the stage for me — really, literally set the stage for me — the hair, the makeup, everything." (Blair, 62F)

Lastly, Alice felt *"privy"* to details, including visual effects or flashbacks, that conveyed the cultural and political commentary behind a music video. For example, she recalled wanting to know more about the political context behind a controversial country music video. However, because the video lacked descriptions, she found it difficult to participate in broader discourse about the video.

5.4.2 Output Modalities. Participants shared a myriad of ways to make music videos accessible beyond traditional AD methods. Two participants preferred inline descriptions, and two others wished to have extended descriptions as their goal of understanding the story *"scene by scene [without feeling] pressed for time"* (Nicki) was prioritized over listening to the music itself. One of the inline AD and one of the extended AD advocates recommended having a separate resource to reference for more details, such as a descriptive "prologue" to set the scene of the music video prior to watching it.

Regarding how the descriptions interacted with the music, two participants viewed music and lyrics to be different from dialogue. They suggested that the music could be ducked or even omitted in favor of AD during repetitive sections such as the chorus.

Three participants were also interested in tactile elements such as Braille and haptics, preferring tactile cues over additional audio. Emily explained that Braille could be helpful for providing AD during musical sequences, noting that audio ducking requires producers to *"turn the song down"* whereas Braille could allow users to hear the song while getting descriptions. Additionally, Isaac thought haptic feedback could improve his immersion with a music video. For example, he suggested: *"when somebody gets punched, your phone can vibrate... just to amplify the experience, especially if you're watching it on your phone where you don't have the subwoofer."*

5.5 Live Video: Seeking Information or Entertainment on a Video Sharing Site or Social Networking Site

Eight participants reported that they watched live videos, such as news, sports, and live streams, to seek information or be entertained. Though these videos typically involved live narration, making them somewhat accessible, they often lacked detailed visual descriptions compared to videos with AD added in post-production.

5.5.1 Details about Visual Aids, People, Actions, and Clothing. For news, descriptive details about visuals were crucial for participant safety. Such visual information applied to infographics as well as live-action clips. For example, during a weather broadcast, Grace was frustrated by usage of generic demonstrative pronouns (such as "this") to refer to specific locations on the weather map, as impending weather conditions could require viewers to take action.

> "I struggle with maps [when] a newscaster [says], 'it [will] rain and it's going to go this way.' I don't know if that's hitting near me. So I like it when they talk through a map. But in general, it can be hard because they're just making blanket statements about a state. And you don't know like, 'Okay, is it coming closer to where I am?'" (Grace, 29F)

Blair was particularly interested in live sports. While she sometimes listened to radio coverage of sports to get more description, for videos she desired *"play-by-play"* (Blair) commentary and wished to know more details such as the players' clothing, actions, and facial expressions. She explained that these details were *"all part of the anticipation of tennis"* (Blair).

5.5.2 Output Modalities. Audio cues could also indicate a scene change during news broadcasts. For example, Emily wished to have AD for news footage, such as overhead shots of scenery, but noted that an audio cue would suffice for indicating the program had switched back to a shot of the news anchor in the studio.

Unlike non-interactable news and sports broadcasts, online platforms such as Facebook, YouTube, and Twitch support live streams where creators can directly respond to BLV viewers' requests. For example, Haley recounted her experiences of asking questions during a kitten sanctuary live stream. While the live streamers often described what the kittens were doing, Haley sometimes utilized the chat to request that they *"describe the kittens themselves."* She found that the creators and other viewers in the chat usually responded positively to her request for access in real time.

5.6 Personal Video: Engaging with Friends or Family on a Social Networking Site

Many participants used social media, most often Facebook, to connect with and watch videos from friends and family (N = 6) or for entertainment (N = 4). As such, video subjects often included people or pets that the viewer knew personally, and the content ranged from pets playing to children's recitals. Participants reported that these videos were most often inaccessible when they did not include much speech or dialogue.

5.6.1 Details about People, Pets, Actions, Settings, and Clothing. Participants wished to know more about people or pets and action. Three participants emphasized that the people or pets in the video were most important to describe: *"if someone is showing a video of their cat, they first want you to focus on their cat and what their cat is doing"* (Emily). Similarly, Nicki emphasized that knowing about the people helped her stay updated with her family: *"for example, if there's a baby in the family, [I want] to see their progress, how much they're growing, what they are accomplishing."*

Three participants were also interested in settings and clothing, but noted that this was less critical. For videos that already implied the people and actions through dialogue, describing the setting was helpful. For example, Felix recalled watching a video of his friend's child performing in a play, and mentioned that he wanted AD to give *"the context of, where's this person singing? What's on stage? What's the setting here?"* Nicki explained that finer details about clothing and colors helped her feel more connected to the emotions captured in the videos.

> "Maybe someone just got their hair done or dyed, and I want to know what color they dyed their hair... If it's a video of someone at the beach, I want to know what color the water looks like, I want to know, is the sky sunny or cloudy? Things like that still bring joy to me." (Nicki, 34F)

5.6.2 Output Modalities. For videos from friends and family, participants did not expect nor want professionally produced descriptions or output modalities. Instead, they wished for the videos to have descriptive narration during the actual filming process of the video, such as *"saying, 'Oh, we're at our local park and so and so is going down the slide for the first time'"* (Nicki), or through accompanying context clues in the video caption. Karla also mentioned that her friends and family would sometimes preface a video's content when sharing it as a proactive way to provide access. For times when friends or family forgot to provide additional context, Emily suggested that platforms could *"remind people to add description."*

5.7 TV Show or Movie: Seeking Entertainment on a Streaming Service

Across multiple genres of television and movies, a majority of participants (N = 11) were interested in knowing more about characters' appearances, actions, clothing, facial expressions, and settings, findings that are in accordance with most existing AD guidelines. However, participants' preferences also illuminated differences across genres, most commonly through their proposed output modalities. Here, we present participants' suggestions in groups based on similar findings.

5.7.1 Science Fiction, Fantasy, and Animation.

> "There's something about *Wall-E*, which is one of my favorite films of all time, that just does not translate... As somebody who could see *Wall-E* and now cannot, I can tell you that the audio description just doesn't have it. It tries really hard to capture the magic, but there's something that is missing out of the little expressions that the characters have that's so hard to describe." (Felix, 40M)

Genres with fantastical elements, such as science fiction, fantasy, and animated content, were often difficult to describe within the constraints of dialogue gaps. Felix, a blind movie critic who lost the majority of his sight at 34 years old, was especially emphatic about how separate resources and additional output modalities could help with his understanding and enjoyment. Separate resources were valuable for including detailed explanations of characters and clothing, such as *"what a stormtrooper wears"* (Felix). He highlighted the impact of having a prologue to set the scene of a show:

> "I've seen films that are so immers[ed] into a fantasy or sci-fi realm… where nothing has a basis in reality… We could have an additional… immersive audio description [prologue] to describe the world in which we're about to live. … That way we can really focus on the story, the characters, the plot, the relationships — what's happening." (Felix, 40M)

Having physical 3D models could also assist in conveying the unique designs and nonverbal expressiveness of animated characters. When discussing animated character design, Felix mentioned: *"I know we can't roll a Wall-E into people's homes, but I almost wish we could."* Emily was also a staunch supporter of utilizing 3D models for video accessibility, mentioning that having access to a mermaid doll helped her conceptualize what Ariel from *The Little Mermaid* looked like. Regarding the practicality of actually obtaining these 3D models, she acknowledged that not all viewers would have access to 3D printers. To make this more feasible and to reduce the technical overhead, she suggested implementing *"a rental program in libraries [such as] the National Library Service for the Blind"* (Emily) or making the 3D models available at movie theaters.

5.7.2 Comedy. Five participants prioritized details that could provide access to *"sight gags"* (Isaac), such as humorous text on screen, facial expressions, or even graphic clothing that contributed to punch lines. Grace recalled her experience watching the show *Lizzie McGuire* and how *"a little cartoon would come in a thought bubble and say her thought[s],"* suggesting that a similar *"cartoon guy voice… [with] rising inflection on the end"* could not only read out text on screen, but also present it in a humorous way.

Others stressed the importance of ensuring that sound effects, both in the original work and added for accessibility purposes, conveyed the right tone. When referring to the video probe from the interview (V2), Isaac mentioned how the video successfully indicated that the ticker on screen was going up and down: *"the slide whistle was very, very helpful… the Foley in the show itself did a good job of describing what was going on."*

Three participants thought tactile elements were not necessary for comedy videos. However, they did mention that it was especially important to not over-explain a joke or prematurely spoil the humor. When talking about the timing, Julia advocated, *"I'd want to go along with it as it's happening in order to make it really funny and really entertaining… I want to be in the moment with everything else."*

5.7.3 Historical, Romance, Reality, and Drama. Six participants watched historical, romance, reality, and drama content for entertainment and relaxation. For reality shows, two participants shared that they were often interested in clothing for the sake of getting a better understanding of a person's character. In particular, when watching a reality dating show, Diana mentioned how her wife would often pause the show and describe a *"ridiculous bathing suit"* or other visual details ad hoc, in a more subjective manner than traditional AD. While she did not want to spend too much time familiarizing herself with a show she had just started watching, she was interested in having a separate resource to reference once she became invested with a show. She recounted her positive experiences with sharing these additional descriptions with friends:

> "I actually typed up [my wife's] descriptions of everybody, and sent it to my other blind friends who are watching the show. They were like, 'Oh my god, this is so great.' I wish there was a thing where, on shows like this, you could choose to go in and access this additional description… because a lot of the time it's just not feasible to put that in there. And obviously, that is kind of subjective on some level, so maybe that's not something that a production company would feel comfortable providing, but it really does enhance the experience to know: is this person who acts really vain actually super hot?" (Diana, 34F)

Regarding character appearances, Felix mentioned the harms of omitting race and ethnicity from AD: *"a person's race is not revealed unless it's not white, and so we're all left to assume that unless told otherwise, everybody is white."* While not all participants were interested in clothing and appearance, they generally appreciated if AD could include race, citing shows such as *Bridgerton* as a good example of how describing characters' races could highlight equitable representation in media.

5.8 Similarities Across Scenarios

While participants generally preferred different details for different scenarios, with Layne even commenting that for *"a lot of internet media, it's kind of context dependent,"* some recurrent suggestions illuminated universal video accessibility needs. Many of these similarities echoed existing guidelines, but participants also contextualized how these ideas could be helpful for emerging scenarios. For example, having access to text on screen was critical for accessibility. Most thought providing text-to-speech or screen reader functionality for text on screen could be helpful, but others cited their disdain for *"the TikTok automated voice"* (Layne) and wanted a human narrator instead. Six participants also wanted diegetic audio to contextualize a video and alleviate the need for verbal descriptions, and ten suggested adding new audio effects for increased comprehension. However, some participants expressed concerns about augmentative audio overshadowing the creative vision of the original video: *"at some point I wonder, is that people messing with it, or what the creator had in mind?"* (Alice). Four participants also wanted transcripts, and two others specifically advocated for dubbing for foreign language videos.

Others suggested adapting a video's visual style to enhance accessibility. Five participants with residual vision found competing colors, flashes, and fast action detrimental when watching videos, and wished to have increased contrast, *"minimal background[s]"* (Grace), and less rapid action to make videos easier to comprehend. Isaac, who found *"flat and simple"* animation styles in shows such as *The Simpsons* easier to view, shared his enthusiasm for changing the stylistic appearance or *"stripping some of the detail"* of a video: *"if you're able to choose a filter and what these characters could look like based on your vision and the way you would prefer to see things… that would be cool."*

Though our results focus more on the diversity of BLV people's perspectives, rather than the generalizability of insights across scenarios, we present some similarities regarding desired details and output modalities in tabular form in Appendix B.

6 DISCUSSION

Our findings detail how different scenarios give rise to varied video accessibility needs. BLV users were in favor of both verbal descriptions and nonverbal output modalities, such as audio cues to indicate scene changes for news, tactile elements to give a sense of character design for science fiction movies, or visual enhancements to increase contrast for fast-paced videos. Though most prior research on image and video descriptions has focused greatly on one outcome for end users, we build on work by Stangl et al. [114] to go beyond universal design and "one-size-fits-all" descriptions. To our knowledge, we are the first to explore varied preferences for video accessibility across a wide set of scenarios.

As technology advances, viewing habits change, and content evolves, it is essential to break away from only adding AD based on traditional guidelines and consider more holistic video accessibility. During our study, participants mentioned a variety of enhancements, ranging from 3D models for unfamiliar concepts to additional resources providing detailed descriptions. The myriad of ideas shared by participants illuminates an emerging design space with more depth and breadth than current AD practice.

In this section, we define a video accessibility design space to provide video creators and video platform designers with an expanded toolkit for making videos more holistically accessible. We then discuss the potential and implications of generative AI's applications to video accessibility and personalization.

6.1 Defining the Design Space for Video Accessibility

Prior HCI work on video accessibility has focused predominantly on providing universal access through audio description — concise, objective narrations spoken during gaps in dialogue. However, since the introduction of AD between the 1960s and 1980s, the video watching landscape has shifted dramatically. Now that we often view videos on our personal devices, many of which are handheld, videos and devices can simultaneously provide other types of output, including haptic and tactile feedback. Additionally, as shown through our findings, participants were interested in using augmentative outputs to convey information for different video scenarios. Our notions of video accessibility should expand to consider the affordances of today's video viewing scenarios and the full context of a user's experience.

Below, we distill the ideas mentioned by participants into six continuous or categorical dimensions to articulate a design space for video accessibility. We also present the design space in Table 5. The first two dimensions are grounded in traditional AD practice [3, 26, 27, 117], while the other four are not yet as common. Similar to other design spaces, these dimensions represent an infinite possibility of different video accessibility solutions.

6.1.1 ***Level of Detail (continuous)****: minimal detail (concise) ⟷ extreme detail (verbose).* To create more effective baseline descriptions, video accessibility creators can leverage known intrinsic qualities of the video, such as the video's type and generally the video's platform, to determine which details to include. Though user goals are often tied to the type and platform, different users may watch the same video for a variety of purposes, and detail preferences vary between BLV users. Our findings can serve as a guide for AD creators to determine which details are of particular interest to BLV audiences for different video types and platforms.

However, given that user goals vary across scenarios, we emphasize that there is no singular "ground-truth" — detail levels should be dynamic and personalizable. BLV users should be able to indicate if they want more or less AD detail through a slider or menu. This setting could be saved universally for a user, but more ideally, should be variable across scenarios (e.g., a system should consistently provide many details for a how-to video but few details for a science fiction film if those are the detail levels a user indicates). Prior work has emphasized the importance of personalizable settings for closed captioning to direct focus and avoid distractions [16, 43, 53]; we recommend for video accessibility systems to also support high degrees of flexibility.

6.1.2 ***Alteration of Video Time (continuous)****: no increase to source material duration ⟷ extension of source material duration.* While inline descriptions, which require no increase to source material duration, are the standard in the AD industry, some audio describers have chosen to extend the duration of source material for particular video types (e.g., fast-paced movie trailers [22]) to provide more time for AD delivery. For short-form videos such as TikToks or Instagram Reels, participants generally wanted additional descriptions to augment the limited AD given during a video's restrictive dialogue gaps. However, while extended AD may allow for the inclusion of extra information, video creators should also consider the additional time and labor that BLV people may have to incur as a result — even a 30-second extension may double the amount of time a BLV person spends on a video.

6.1.3 ***Level of Augmentation (continuous)****: no accessibility measures added ⟷ any number of accessibility measures added after a video's initial production or release.* Most videos were found to benefit from some degree of augmented accessibility measures. Some videos are already accessible by nature, meaning that the source video's audio inherently conveys some information to the user through dialogue, narration (including voice-over narrations added during the editing process), diegetic audio, or other audio effects. Other videos are made accessible afterwards through the addition of AD. Participants also referenced videos that were completely inaccessible (e.g., how-to videos with only music in the background), which required multiple augmentations for complete access. The degree to which a video is understandable from its audio can help determine how extensively to augment the video, which can be through AD and / or other modalities.

6.1.4 ***Modality of Presentation (categorical)****: spoken descriptions, audio cues, visual enhancements, Braille, tactile graphics, 3D models, haptics, etc.* Additional audio elements can improve video accessibility. Similar to how Netflix's title card includes a recognizable sound effect [120], different platforms may adopt a set of distinctive earcons to efficiently indicate common information. For example, streaming services may wish to standardize a sound indicating that credits are rolling, while content creators may wish to designate a specific earcon for encouraging viewers to "subscribe" or "follow" their content. Video creators can reference podcasts and radio sportscasts as strong examples of descriptive and rich audio

Table 5: Our six-dimensional video accessibility design space.

	Dimension	**Endpoints / Examples**
Continuous	Level of Detail	Minimal detail (concise) ⟷ extreme detail (verbose)
	Alteration of Video Time	No increase to source material duration ⟷ extension of source material duration
	Level of Augmentation	No accessibility measures added ⟷ any number of accessibility measures added after a video's initial production or release
Categorical	Modality of Presentation	Spoken descriptions, audio cues, visual enhancements, Braille, tactile graphics, 3D models, haptics, etc.
	Synchronicity of Accessible Content	Before, during, or after watching a video
	Tone and Style of Approach	Excited, sad, judgmental, first-person perspective, cartoonish, etc.

experiences that leverage vocal performance and sound design to engage listeners [62], and may also draw on prior research exploring how sound design can enhance the accessibility and aesthetics of auditory websites [136]. Some films have already adopted such practices of using sound design techniques (e.g., sound effects, 3D audio) to enhance the experience for BLV audiences [25, 52].

Prior research has found that augmentative tactile elements are helpful for improving access to artwork [14, 15, 109] and theater experiences [122]. We found that participants wanted tactile feedback in formats that were not available through just a smartphone — they also wished to have additional materials such as 3D models and Braille output. However, many participants acknowledged that this relied heavily on having (1) the requisite technology, such as 3D printers or refreshable Braille displays, (2) Braille literacy, and (3) tactile graphicacy.

However, it is important to consider that not all participants were interested in additional output modalities — some wished to use them part of the time, some thought they would be easily confused with other cues such as their phone ringing, and some acknowledged the high learning curve. We recommend for audio cues to be included in a tertiary audio track to allow BLV users full control over whether and when they would like to hear audio cues in addition to AD. For emerging devices, such as extended reality (XR) headsets, video accessibility creators can borrow from prior literature in XR accessibility and game design (e.g., [12, 59, 64]) to determine if additional modalities such as smell or taste are appropriate for information presentation.

6.1.5 ***Synchronicity of Accessible Content (categorical)****: before, during, or after watching a video.* Primers or prologues were particularly helpful to access prior to watching exercise videos, music videos, and content with fantastical characters or extensive world-building (e.g., science fiction, fantasy, historical fiction). However, for unexpected parts of a video, such as a surprise character appearance, it was more favorable to access additional descriptions afterwards to avoid spoiling the surprise. Additionally, some participants only wished to reference separate resources and descriptions after becoming invested in a show. This extends findings from preliminary research on the positive impact of having audio introductions for select feature films [98] and prior work on image exploration, which found that BLV users frequently accessed overview menus at the beginning and end of exploring images [81].

6.1.6 ***Tone and Style of Approach (categorical)****: excited, sad, judgmental, first-person perspective, cartoonish, etc.* While most existing AD is generally presented in a neutral, third-person tone [28, 70], researchers have found that changing the tone or style of verbal description presentation can improve viewers' immersion in a video [30, 59, 121, 125]. We explored how this could extend to a variety of different scenarios. For example, one participant wished to have subjective and somewhat judgmental descriptions for reality television, while others wished to have descriptions that matched the tone of the piece overall. Aside from spoken descriptions, other output modalities could also take different tones and styles; for example, tactile graphics and 3D models could present users with a more cartoonish representation of a video's visuals to abstract away unnecessary details.

6.2 Design Recommendations and Examples

Across the design space, our study yielded several recommendations for current scenarios.

- Video creators should provide outside resources for BLV audience members to refer to. This can take the form of a vlogger posting a visual description of themself or a blockbuster movie providing a detailed introduction to their characters.
- Entertainment videos (e.g., music videos, historical shows, reality television, etc.) should focus more on describing subjects, such as people and animals, which can be done through a variety of modalities. Additional information, including costumes, can help contextualize scenes, especially those with any fantastical or historical elements.
- On the other hand, for a how-to video, subject appearances are less important. Precise information about actions and equipment should be prioritized to aid in the goal of learning how to do something.

Different points in this design space have been lightly explored in real-world settings. For example, one short-form video creator went viral for creating AD in a narrative and poetic style. When a TikTok user requested for other social media users to provide access to the videos and memes regarding the Montgomery Riverfront brawl in August of 2023 [88, 107], one of the responses he received was particularly distinctive for its creative and calming description of the chaotic event unfolding from the perspective of the co-captain's hat [89]. In an interview by the Washington Post, the responder, a sighted content creator, noted that *"Slater's request inspired him to try a kind of oral storytelling that transcended sensory experiences, in the style of a folk tale"* [56]. Additionally, a Netflix original series, *All the Light We Cannot See*, was one of the first television shows to be released with an official "audio introduction" [127]. The show, starring a blind lead actress and telling the story of a blind French girl and a German soldier during World War II, leveraged the audio introduction to describe character appearances, clothing, movement styles, and settings. These examples demonstrate the success of two newer dimensions, reinforcing the viability of this design space for different naturalistic scenarios.

As video viewing scenarios continue to change, new preferences may arise. We encourage future work to innovate new techniques for description presentation, explore various detail levels, and consider novel user interface designs that enable the personalization of video accessibility.

6.3 Applying Generative AI to Explore the Video Accessibility Design Space

To move towards user-centered and holistic video accessibility, we propose leveraging generative AI to explore the design space for different scenarios. Generative AI can serve as a tool for designers, creators, and end users to adjust video accessibility on-demand.

Some research has focused on the development of datasets and NLP techniques for video understanding and accessibility [42, 47, 48, 115, 131, 138]. Others have developed AI-based tools to support accessibility practices [6, 11, 66, 93, 110, 126, 134, 135]. Major advancements in multi-modal language models such as OpenAI's GPT-4V [90, 91] and Google's Gemini [24, 116] show that AI is already capable of generating image descriptions, and some video descriptions, that attain high levels of quality [135] and BLV user satisfaction [23, 110]. However, prior efforts do not specifically consider short-form video, and they primarily aim to automatically generate text descriptions (e.g., [11, 31]).

To effectively train AI models for these varied use cases, it is crucial to create comprehensive datasets that reflect a wide range of scenarios, information display preferences, and output modalities. If these datasets will contain sensitive information pertaining to BLV people — for example, to capture the scenario of watching videos from friends and family — we must also recognize and consider BLV users' visual privacy concerns [46, 112, 113, 137].

Our study evidences the importance of scenario-based approaches [13] for video accessibility; however, given the size of the design space, it is infeasible to thoroughly explore all possible designs. As such, below we address current capabilities and limitations of AI systems for three dimensions and suggest improvements for the future of video accessibility.

- **Level of Detail**: Prior work has investigated the potential for visual question answering systems to enable users to query for details that they wish to know [6, 58, 110]. As AI advances, it may one day be possible to provide end users with high degrees of flexibility for which details and what level of detail they would like through automatically generated descriptions. While participants did not want completely AI-generated descriptions for professionally produced content, they thought it was desirable for user-generated content created with little to no budget, such as TikToks, Instagram Reels, and videos shared by friends and family.
- **Modality of Presentation**: Our findings show that having additional modalities for conveying visual information, such as 3D models or Braille, were welcomed by participants. Existing open-source resources and workshops are an excellent starting point for becoming familiar with tactile graphics and objects [33, 102]. However, given the large amount of possibilities for modality, which include tactile graphics, 3D models, and audio cues, we suggest that generative AI can be valuable for quickly prototyping a wide variety of modalities to determine which would be the best fit for a specific scenario. For example, some participants wished to know the scale of the forest when watching the nature documentary. Some suggested having a tactile graphic, whereas others found 3D models more helpful for learning. Building on prior tactile graphics and tactile display research (e.g., [51, 80, 92]), non-technically savvy users could utilize generative AI to create scalable vector graphic files for tactile graphics and produce STL code for 3D printing.
- **Tone and Style of Approach**: One of the strengths of current AI systems is their ability to mimic existing writing styles and tones [21, 35], a capability that has been applied to both fiction (e.g., [29]) and scientific writing (e.g., [50]). Aside from changing the style of textual descriptions, tones and styles can also be altered for other output modalities. As some participants with residual vision mentioned, fast-moving visuals were often inaccessible, and some wished to change the visual style of the entire video. To cater to individual users' levels of vision or stylistic preferences, generative AI can aid video accessibility creators and consumers in iterating upon different options to find the tone or style that works best for them.

Recent advances in AI models have led to excitement about potential uses of AI for video accessibility. However, we also recognize that there can be significant ethical harms associated with AI-generated and personalized video accessibility, and we caution against the unregulated deployment of such technologies. Currently, BLV people typically receive AD from a trusted friend or a company that undergoes multiple iterations of quality control. If video accessibility becomes largely automatically generated, and if there are limited methods for assessing the quality of the output, the impacts of biases and misinformation perpetuated by AI can become magnified [1, 57, 129].

As we learned from our study, BLV people watch videos for a wide range of purposes, ranging from entertainment to learning critical information that can impact their safety. Misinformation on a medical video, for example, could be life-threatening. Additionally, consider that a user might want to access a breaking news story about climate change and extreme weather events, or a video reporting on vaccine efficacy. Should their descriptions be personalized and adopt a tone similar to the user's favorite news publication, or should they be more neutral? Should this differ for videos found on news outlets versus social media? Long-term, how would partisan audio description and video accessibility design influence people's views, political or otherwise?

BLV end users inevitably have varied information goals and preferences for video accessibility. As AI continues to rapidly advance, the potential for end users to have personalized agents that can learn and remember their preferences also grows. In line with its impacts on text and image generation, AI is likely to play a big role in video, and video accessibility, generation as well. We encourage future work to leverage the capabilities of generative AI, with a human in the loop, to achieve greater video accessibility at scale while mitigating potential risks and harms.

6.4 Limitations and Future Work

In this study, we investigated a small set of specific scenarios based on findings from our formative survey. Our survey had a relatively small number of respondents, so our survey findings should not be generalized. Due to limitations of our survey platform, the video types and platforms listed were not randomized. We acknowledge that this could have affected survey results, which in turn could have impacted our scenarios for the interview study. While studying specific scenarios slightly limits the generalizability of our findings, we highlight variation in BLV users' preferences across a diverse set of different scenarios and provide a foundation for future video accessibility work. Additionally, as all interview participants self-identified as blind, we did not have an opportunity to thoroughly understand how visual enhancements could improve video accessibility for people with low vision — future work could specifically focus on low vision people's experiences.

We encourage researchers to continue critically examining the ethical and societal impacts of personalized and holistic video accessibility. For example, how do BLV users' lived experiences and cultural backgrounds influence their preferences for styles and tones? What are the implications of video accessibility personalization in terms of reinforcing echo chambers or biases, especially for news and social media content? Furthermore, given the large variety and volume of video content, it is infeasible to manually create personalized accessible videos for all possible scenarios or user preferences. Future work can explore how generative AI can be applied to AD creation in practice, understand to what degree biases can manifest in AI-generated video accessibility, including representational harms [4, 36], and investigate how these biases can impact BLV users' perspectives and trust of AI systems long-term.

Lastly, this work was conducted via virtual interviews; due to logistical and software constraints, we did not examine BLV users' video access preferences in-the-wild. We encourage future work to investigate this area with longitudinal and in-situ studies, such as a diary study to capture insights during participants' actual video viewing sessions, to better capture the breadth of scenarios experienced by BLV users.

7 CONCLUSION

Through a formative survey and a semi-structured interview study, we investigated BLV users' preferences for video accessibility across diverse video scenarios. These preferences included varied levels and types of details provided, as well as additional output modalities such as audio cues, tactile elements, and visual enhancements. We identified preferred details and output modalities for different scenarios, such as watching short-form videos on Instagram to engage with friends, how-to videos on YouTube to learn how to do something, and science fiction movies on Netflix for entertainment. To our knowledge, we are one of the first to (1) contribute empirical insights capturing the diversity of BLV users' video accessibility preferences, (2) consider a wide range of video viewing scenarios, and (3) present a design space to guide future accessible video creation and innovation. Understanding BLV users' accessibility preferences across viewing scenarios can help us move towards more personalized and holistic video accessibility paradigms.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their valuable feedback. We also thank Daniel Zhu, Kelly Jiang, Sharon Heung, Danielle Montour, and Emma McDonnell for their support and input throughout this project. We are grateful to our study participants for their involvement and insights. This work was partially supported by a gift from Meta (Meta Platforms, Inc.).

REFERENCES

[1] Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in ChatGPT: implications in scientific writing. *Cureus* 15, 2 (2023). https://doi.org/10.7759/cureus.35179

[2] Saki Asakawa, João Guerreiro, Daisuke Sato, Hironobu Takagi, Dragan Ahmetovic, Desi Gonzalez, Kris M Kitani, and Chieko Asakawa. 2019. An independent and interactive museum experience for blind people. In *Proceedings of the 16th International Web for All Conference*. 1–9. https://doi.org/10.1145/3315002.3317557

[3] Audio Description Coalition. 2009. *Standards for Audio Description and Code of Professional Conduct for Describers*. https://www.perkins.org/wp-content/uploads/elearning-media/adc_standards.pdf

[4] Cynthia L Bennett, Cole Gleason, Morgan Klaus Scheuerman, Jeffrey P Bigham, Anhong Guo, and Alexandra To. 2021. "It's Complicated": Negotiating Accessibility and (Mis) Representation in Image Descriptions of Race, Gender, and Disability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3411764.3445498

[5] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, and Tom Yeh. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342. https://doi.org/10.1145/1866029.1866080

[6] Aditya Bodi, Pooyan Fazli, Shasta Ihorn, Yue-Ting Siu, Andrew T Scott, Lothar Narins, Yash Kant, Abhishek Das, and Ilmi Yoon. 2021. Automated Video Description for Blind and Low Vision Users. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7. https://doi.org/10.1145/3411763.3451810

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[8] Craig Brown and Amy Hurst. 2012. VizTouch: automatically generated tactile visualizations of coordinate spaces. In *Proceedings of the Sixth International Conference on Tangible, Embedded and Embodied Interaction*. 131–138. https://doi.org/10.1145/2148131.2148160

[9] Michele A Burton, Erin Brady, Robin Brewer, Callie Neylan, Jeffrey P Bigham, and Amy Hurst. 2012. Crowdsourcing subjective fashion advice using VizWiz: challenges and opportunities. In *Proceedings of the 14th international ACM SIGACCESS conference on Computers and accessibility*. 135–142. https://doi.org/10.1145/2384916.2384941

[10] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. 2018. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1209–1218. https://doi.org/10.1109/CVPR.2018.00132

[11] 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 (2020), 99–111. https://doi.org/10.1007/s10209-018-0634-4

[12] Priscilla Maria Cardoso Garone, Sérgio Nesteriuk, and Gisela Belluzzo de Campos. 2020. Sensory design in games: Beyond visual-based experiences. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Communication, Organization and Work: 11th International Conference, DHM 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22*. Springer, 322–333. https://doi.org/10.1007/978-3-030-49907-5_23

[13] John M Carroll. 2003. *Scenario-based design.* MIT Press. https://arl.human.cornell.edu/linked%20docs/Scenario-Based%20Design%20John%20Carrol.pdf

[14] Luis Cavazos Quero, Jorge Iranzo Bartolomé, and Jundong Cho. 2021. Accessible visual artworks for blind and visually impaired people: comparing a multimodal approach with tactile graphics. *Electronics* 10, 3 (2021), 297. https://doi.org/10.3390/electronics10030297

[15] Luis Cavazos Quero, Jorge Iranzo Bartolomé, Seonggu Lee, En Han, Sunhee Kim, and Jundong Cho. 2018. An interactive multimodal guide to improve art accessibility for blind people. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 346–348. https://doi.org/10.1145/3234695.3241033

[16] Anna C Cavender, Jeffrey P Bigham, and Richard E Ladner. 2009. ClassInFocus: enabling improved visual attention strategies for deaf and hard of hearing students. In *Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility*. 67–74. https://doi.org/10.1145/1639642.1639656

[17] Diagram Center. 2019. *Image Description Guidelines.* http://diagramcenter.org/table-of-contents-2.html

[18] 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*. 1–14. https://doi.org/10.1145/3526113.3545613

[19] 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

[20] 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

[21] Cesc Chunseong Park, Byeongchang Kim, and Gunhee Kim. 2017. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 895–903. https://doi.org/10.1109/CVPR.2017.681

[22] Social Audio Description Collective. 2021. *Spider Man No Way Home Trailer with Expanded Audio Description.* https://www.youtube.com/watch?v=5CuHs-yVMLw

[23] Milagros Costabel. 2023. *I'm Totally Blind. Artificial Intelligence Is Helping Me Rediscover the World.* https://slate.com/technology/2023/10/ai-image-tools-blind-low-vision.html

[24] Google DeepMind. 2023. *Gemini.* https://deepmind.google/technologies/gemini/#build-with-gemini

[25] Dianna Delling. 2024. *This 'pictureless' film is visionary cinema for those who can't see.* https://www.mastercard.com/news/perspectives/2024/australia-touch-film/

[26] Described and Captioned Media Program. 2023. *Audio Description Tip Sheet.* https://dcmp.org/learn/227-audio-description-tip-sheet

[27] Described and Captioned Media Program. 2023. *Description Key.* https://dcmp.org/learn/descriptionkey

[28] Described and Captioned Media Program. 2023. *Description Key - How to Describe.* https://dcmp.org/learn/617-description-key---how-to-describe

[29] Josh Dzieza. 2022. *The Great Fiction of AI.* https://www.theverge.com/c/23194235/ai-fiction-writing-amazon-kindle-sudowrite-jasper

[30] Deborah I Fels, John Patrick Udo, Peter Ting, Jonas E Diamond, and Jeremy I Diamond. 2006. Odd Job Jack described: a universal design approach to described video. *Universal Access in the Information society* 5 (2006), 73–81. https://doi.org/10.1007/s10209-006-0025-0

[31] Anna Fernández-Torné and Anna Matamala. 2015. Text-to-speech vs. human voiced audio descriptions: a reception study in films dubbed into Catalan. *The Journal of Specialised Translation* 24 (2015), 61–88. https://core.ac.uk/download/pdf/78531939.pdf

[32] John C Flanagan. 1954. The critical incident technique. *Psychological bulletin* 51, 4 (1954), 327. https://doi.org/10.1037/h0061470

[33] Chancey Fleet. 2017. *Announcing Dimensions: Community Tools for Creating Tactile Graphics & Objects.* https://www.nypl.org/blog/2017/10/18/dimensions-tactile-graphics-objects

[34] Louise Fryer. 2016. *An introduction to audio description: A practical guide.* Routledge. https://doi.org/10.4324/9781315707228

[35] Chuang Gan, Zhe Gan, Xiaodong He, Jianfeng Gao, and Li Deng. 2017. Stylenet: Generating attractive visual captions with styles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3137–3146. https://doi.org/10.1109/CVPR.2017.108

[36] Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. 2023. An Autoethnographic Case Study of Generative Artificial Intelligence's Utility for Accessibility. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–8. https://doi.org/10.1145/3597638.3614548

[37] Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M Kitani, and Jeffrey P Bigham. 2019. "It's almost like they're trying to hide it": How User-Provided Image Descriptions Have Failed to Make Twitter Accessible. In *The World Wide Web Conference*. 549–559. https://doi.org/10.1145/3308558.3313605

[38] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–10. https://doi.org/10.1145/3373625.3417027

[39] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B Chilton, and Jeffrey P Bigham. 2019. Making memes accessible. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 367–376. https://doi.org/10.1145/3308561.3353792

[40] Cole Gleason, Amy Pavel, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. 2020. Twitter A11y: A browser extension to make Twitter images accessible. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–12. https://doi.org/10.1145/3313831.3376728

[41] Cristos Goodrow. 2017. *You know what's cool? A billion hours.* https://blog.youtube/news-and-events/you-know-whats-cool-billion-hours/

[42] Google. 2019. *YouTube-8M.* https://research.google.com/youtube8m/

[43] Benjamin M Gorman, Michael Crabb, and Michael Armstrong. 2021. Adaptive Subtitles: Preferences and Trade-Offs in Real-Time Media Adaption. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11. https://doi.org/10.1145/3411764.3445509

[44] W3C Accessibility Guidelines Working Group. 2022. *Using alt attributes on img elements.* https://www.w3.org/WAI/WCAG21/Techniques/html/H37.html

[45] Danna Gurari and Kristen Grauman. 2017. Crowdverge: Predicting if people will agree on the answer to a visual question. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3511–3522. http://doi.org/10.1145/3025453.3025781

[46] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. 2019. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 939–948. https://doi.org/10.1109/CVPR.2019.00103

[47] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023. AutoAD II: The sequel-who, when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13645–13655. https://doi.org/10.1109/ICCV51070.2023.01255

[48] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023. AutoAD: Movie Description in Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18930–18940. https://doi.org/10.48550/arXiv.2303.16899

[49] Margot Hanley, Solon Barocas, Karen Levy, Shiri Azenkot, and Helen Nissenbaum. 2021. Computer vision and conflicting values: Describing people with automated alt text. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 543–554. https://doi.org/10.1145/3461702.3462620

[50] Elisa L Hill-Yardin, Mark R Hutchinson, Robin Laycock, and Sarah J Spencer. 2023. A Chat (GPT) about the future of scientific publishing. *Brain Behav Immun* 110 (2023), 152–154. https://doi.org/10.1016/j.bbi.2023.02.022

[51] Leona Holloway, Swamy Ananthanarayan, Matthew Butler, Madhuka Thisuri De Silva, Kirsten Ellis, Cagatay Goncu, Kate Stephens, and Kim Marriott. 2022. Animations at Your Fingertips: Using a Refreshable Tactile Display to Convey Motion Graphics for People who are Blind or have Low Vision. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16. https://doi.org/10.1145/3517428.3544797

[52] Shelley Hughes. 2024. *York academics collaborate on soundtrack of BAFTA-nominated film.* https://www.york.ac.uk/news-and-events/news/2024/research/academics-bafta-film/

[53] Dhruv Jain, Rachel Franz, Leah Findlater, Jackson Cannon, Raja Kushalnagar, and Jon Froehlich. 2018. Towards accessible conversations in a mobile context for people who are deaf and hard of hearing. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 81–92. https://doi.org/10.1145/3234695.3236362

[54] Gaurav Jain, Basel Hindi, Connor Courtien, Conrad Wyrick, Xin Yi Therese Xu, Michael C Malcolm, and Brian A Smith. 2023. Towards Accessible Sports Broadcasts for Blind and Low-Vision Viewers. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7. https://doi.org/10.1145/3544549.3585610

[55] Gaurav Jain, Basel Hindi, Connor Courtien, Xin Yi Therese Xu, Conrad Wyrick, Michael Malcolm, and Brian A Smith. 2023. Front Row: Automatically Generating Immersive Audio Representations of Tennis Broadcasts for Blind Viewers. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3586183.3606830

[56] Maham Javaid. 2023. *How oral storytelling helped a blind man see the Montgomery brawl.* https://www.washingtonpost.com/nation/2023/08/12/montgomery-riverfront-brawl-blind-tiktok-andy-slater/

[57] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38. https://doi.org/10.1145/3571730

[58] 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*. 1–3. https://doi.org/10.1145/3517428.3550394

[59] 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*. 1–17. https://doi.org/10.1145/3597638.3608381

[60] Ju Yeon Jung, Tom Steinberger, Junbeom Kim, and Mark S Ackerman. 2022. “So What? What’s That to Do With Me?” Expectations of People With Visual Impairments for Image Descriptions in Their Personal Photo Activities. In *Designing Interactive Systems Conference*. 1893–1906. https://doi.org/10.1145/3532106.3533522

[61] Daniel Killough and Amy Pavel. 2023. Exploring Community-Driven Descriptions for Making Livestreams Accessible. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13. https://doi.org/10.1145/3597638.3608425

[62] Georgina Kleege. 2023. Fiction Podcasts Model Description by Design. In *Crip Authorship*. New York University Press, 318–325. https://doi.org/10.18574/nyu/9781479819386.003.0033

[63] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. *arXiv preprint arXiv:2205.10646* (2022). https://doi.org/10.48550/arXiv.2205.10646

[64] Ernst Kruijff, Alexander Marquardt, Christina Trepkowski, Jonas Schild, and André Hinkenjann. 2015. Enhancing user engagement in immersive games through multisensory cues. In *2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games)*. IEEE, 1–8. https://doi.org/10.1109/VS-GAMES.2015.7295773

[65] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. ImageExplorer: Multi-Layered Touch Exploration to Encourage Skepticism Towards Imperfect AI-Generated Image Captions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15. https://doi.org/10.1145/3491102.3501966

[66] 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*. 1–14. https://doi.org/10.1145/3411764.3445233

[67] Mariana Lopez, Gavin Kearney, and Krisztián Hofstädter. 2022. Seeing films through sound: Sound design, spatial audio, and accessibility for visually impaired audiences. *British Journal of Visual Impairment* 40, 2 (2022), 117–144. https://doi.org/10.1177/0264619620935935

[68] Lucy.q. 2023. *i <3 my fans!! #paris #sacrecoeur #dailyvlogs*. https://www.instagram.com/p/CuuHpRmvVqN/

[69] Yalan Luo, Weiyue Lin, Yuhan Liu, Xiaomei Nie, Xiang Qian, and Hanyu Guo. 2023. Wesee: Digital Cultural Heritage Interpretation for Blind and Low Vision People. In *IFIP Conference on Human-Computer Interaction*. Springer, 123–131. https://doi.org/10.1007/978-3-031-42280-5_8

[70] María Jesús Machuca and Anna Matamala. 2022. Neutral voices in audio descriptions: What does it mean? *Babel* 68, 5 (2022), 668–696. https://doi.org/10.1075/babel.00287.mac

[71] Kelly Mack, Edward Cutrell, Bongshin Lee, and Meredith Ringel Morris. 2021. Designing Tools for High-Quality Alt Text Authoring. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14. https://doi.org/10.1145/3441852.3471207

[72] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. 2017. Understanding blind people’s experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 5988–5999. https://doi.org/10.1145/3025453.3025814

[73] 3Play Media. 2022. *The Ultimate Guide to Audio Description*. https://www.3playmedia.com/learn/popular-topics/audio-description/

[74] Lisa Montenegro. 2022. *In 2022, Video Is Where We All Need To Be*. https://www.forbes.com/sites/forbesagencycouncil/2022/01/28/in-2022-video-is-where-we-all-need-to-be/

[75] Valerie S Morash, Yue-Ting Siu, Joshua A Miele, Lucia Hasty, and Steven Landau. 2015. Guiding novice web workers in making image descriptions using templates. *ACM Transactions on Accessible Computing (TACCESS)* 7, 4 (2015), 1–21. https://doi.org/10.1145/2764916

[76] Meredith Ringel Morris, Jazette Johnson, Cynthia L Bennett, and Edward Cutrell. 2018. Rich representations of visual content for screen reader users. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–11. https://doi.org/10.1145/3173574.3173633

[77] Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P Bigham, and Shaun K Kane. 2016. With most of it being pictures now, I rarely use it Understanding Twitter’s Evolving Accessibility to Blind Users. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5506–5516. https://doi.org/10.1145/2858036.2858116

[78] Martez E Mott, John Tang, and Edward Cutrell. 2023. Accessibility of Profile Pictures: Alt Text and Beyond to Express Identity Online. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3544548.3580710

[79] Annika Muehlbradt and Shaun K Kane. 2022. What’s in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Transactions on Accessible Computing (TACCESS)* 15, 1 (2022), 1–32. https://doi.org/10.1145/3507659

[80] Mukhriddin Mukhiddinov and Soon-Young Kim. 2021. A systematic literature review on the automatic creation of tactile graphics for the blind and visually impaired. *Processes* 9, 10, 1726. https://doi.org/10.3390/pr9101726

[81] Vishnu Nair, Hanxiu’Hazel’ Zhu, and Brian A Smith. 2023. ImageAssist: Tools for Enhancing Touchscreen-Based Image Exploration Systems for Blind and Low Vision Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17. https://doi.org/10.1145/3544548.3581302

[82] Rosiana Natalie. 2022. Cost-effective and Collaborative Methods to Author Video’s Scene Description for Blind People. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–5. https://doi.org/10.1145/3491101.3503814

[83] 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*. 1–4. https://doi.org/10.1145/3373625.3418030

[84] 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*. 1–15. https://doi.org/10.1145/3441852.3471201

[85] Netflix. 2020. *Our Planet | Forests | FULL EPISODE | Netflix*. https://www.youtube.com/watch?v=JkaxUblCGz0&t=83s

[86] Netflix Inc. 2023. *Audio Description Style Guide v2.5*. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-5

[87] Alexandre Nevsky, Radu-Daniel Vatavu, Timothy Neate, and Elena Simperl. 2023. Accessibility Research in Digital Audiovisual Media: What Has Been Achieved and What Should Be Done Next? (2023), 94–114. https://doi.org/10.1145/3573381.3596159

[88] WSFA News. 2023. *Full Video: Viewer records as Montgomery riverfront brawl begins*. https://www.wsfa.com/video/2023/08/07/full-video-viewer-records-montgomery-riverfront-brawl-begins/

[89] NotWildlin. 2023. *@Andy Slater i hope this kinda helps*. https://www.tiktok.com/@notwildlin/video/7265363866069093678

[90] OpenAI. 2023. *GPT-4*. https://openai.com/product/gpt-4

[91] OpenAI. 2023. GPT-4V(ision) System Card. (2023). https://cdn.openai.com/papers/GPTV_System_Card.pdf

[92] Athina Panotopoulou, Xiaoting Zhang, Tammy Qiu, Xing-Dong Yang, and Emily Whiting. 2020. Tactile line drawings for improved shape understanding in blind and visually impaired users. *ACM Transactions on Graphics (TOG)* 39, 4, 89–1. https://doi.org/10.1145/3386569.3392388

[93] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 747–759. https://doi.org/10.1145/3379337.3415864

[94] Helen Petrie, Chandra Harrison, and Sundeep Dev. 2005. Describing images on the web: a survey of current practice and prospects for the future. *Proceedings of Human Computer Interaction International (HCII)* 71, 2 (2005). https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ad28153c8ee2a3fa6fc90075b8643ce51eb6d59f

[95] Wong Fu Productions. 2015. *How Old Is She?!* https://www.youtube.com/watch?v=91lYBbBkftA

[96] Kyle Rector, Keith Salmon, Dan Thornton, Neel Joshi, and Meredith Ringel Morris. 2017. Eyes-free art: Exploring proxemic audio interfaces for blind and low vision art engagement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21. https://doi.org/10.1145/3130958

[97] Jacob M Rigby, Duncan P Brumby, Anna L Cox, and Sandy JJ Gould. 2016. Watching movies on Netflix: investigating the effect of screen size on viewer immersion. In *Proceedings of the 18th international conference on human-computer interaction with mobile devices and services adjunct*. 714–721. https://doi.org/10.1145/2957265.2961843

[98] Pablo Romero-Fresco and Louise Fryer. 2013. Could audio-described films benefit from audio introductions? An audience response study. *Journal of*

Visual Impairment & Blindness 107, 4 (2013), 287–295. https://doi.org/10.1177/0145482X1310700405

[99] Andreas Sackl, Franziska Graf, Raimund Schatz, and Manfred Tscheligi. 2020. Ensuring accessibility: Individual video playback enhancements for low vision users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–4. https://doi.org/10.1145/3373625.3417997

[100] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5. 147–156. https://doi.org/10.1609/hcomp.v5i1.13301

[101] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2018. Evaluating and Complementing Vision-to-Language Technology for People who are Blind with Conversational Crowdsourcing.. In *IJCAI*. 5349–5353. https://www.ijcai.org/Proceedings/2018/0751.pdf

[102] Marco Salsiccia. 2023. *SVG Artwork*. https://marconius.com/svg/

[103] Ather Sharif, Olivia H Wang, Alida T Muongchan, Katharina Reinecke, and Jacob O Wobbrock. 2022. VoxLens: Making Online Data Visualizations Accessible with an Interactive JavaScript Plug-In. In *CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3491102.3517431

[104] Kurt Shuster, Samuel Humeau, Hexiang Hu, Antoine Bordes, and Jason Weston. 2019. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12516–12526. https://doi.org/10.1109/CVPR.2019.01280

[105] Rachel N Simons, Danna Gurari, and Kenneth R Fleischmann. 2020. I Hope This Is Helpful Understanding Crowdworkers' Challenges and Motivations for an Image Description Task. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26. https://doi.org/10.1145/3415176

[106] Alexa Siu, Gene SH Kim, Sile O'Modhrain, and Sean Follmer. 2022. Supporting accessible data visualization through audio data narratives. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3491102.3517678

[107] Andy Slater. 2023. *#Inclusivity #Alabama #ancientanbiens*. https://www.tiktok.com/@thisisandyslater/video/7264770242721697070

[108] Joel Snyder. 2005. Audio description: The visual made verbal. In *International Congress Series*, Vol. 1282. Elsevier, 935–939. https://doi.org/10.1016/j.ics.2005.05.215

[109] Abigale Stangl, Ann Cunningham, Lou Ann Blake, and Tom Yeh. 2019. Defining problems of practices to advance inclusive tactile media consumption and production. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 329–341. https://doi.org/10.1145/3308561.3353778

[110] Abigale Stangl, Shasta Ihorn, Yue-Ting Siu, Aditya Bodi, Mar Castanon, Lothar Narins, and Ilmi Yoon. 2023. The Potential of a Visual Dialogue Agent In a Tandem Automated Audio Description System for Videos. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–16. https://doi.org/10.1145/3597638.3608402

[111] 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*. 1–13. https://doi.org/10.1145/3313831.3376404

[112] Abigale Stangl, Emma Sadjo, Pardis Emami-Naeini, Yang Wang, Danna Gurari, and Leah Findlater. 2023. "Dump it, Destroy it, Send it to Data Heaven": Blind People's Expectations for Visual Privacy in Visual Assistance Technologies. In *Proceedings of the 20th International Web for All Conference*. 134–147. https://doi.org/10.1145/3587281.3587296

[113] Abigale Stangl, Kristina Shiroma, Nathan Davis, Bo Xie, Kenneth R Fleischmann, Leah Findlater, and Danna Gurari. 2022. Privacy concerns for visual assistance technologies. *ACM Transactions on Accessible Computing (TACCESS)* 15, 2, 1–43. https://doi.org/10.1145/3517384

[114] 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 *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–15. https://doi.org/10.1145/3441852.3471233

[115] Amara Tariq and Hassan Foroosh. 2015. Feature-independent context estimation for automatic image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1958–1965. https://doi.org/10.1109/CVPR.2015.7298806

[116] Google Gemini Team. 2023. Gemini: A Family of Highly Capable Multimodal Models. (2023). https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf

[117] The American Council of the Blind. 2003. *Guidelines for Audio Describers*. https://adp.acb.org/guidelines.html

[118] The American Council of the Blind. 2023. *All About Audio Description*. https://adp.acb.org/ad.html

[119] The American Council of the Blind. 2023. *The Audio Description Project*. https://adp.acb.org/

[120] Twenty Thousand Hertz. 2020. *Tudum! It's Netflix*. https://www.20k.org/episodes/netflix

[121] John Patrick Udo, Bertha Acevedo, and Deborah I Fels. 2010. Horatio audio-describes Shakespeare's Hamlet: Blind and low-vision theatre-goers evaluate an unconventional audio description strategy. *British Journal of Visual Impairment* 28, 2 (2010), 139–156. https://doi.org/10.1177/0264619609359753

[122] John-Patrick Udo and Deborah I Fels. 2010. Enhancing the entertainment experience of blind and low-vision theatregoers through touch tours. *Disability & Society* 25, 2 (2010), 231–240. https://doi.org/10.1080/09687590903537497

[123] Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 319–326. https://doi.org/10.1145/985692.985733

[124] Violeta Voykinska, Shiri Azenkot, Shaomei Wu, and Gilly Leshed. 2016. How blind people interact with visual content on social networking services. In *Proceedings of the 19th acm conference on computer-supported cooperative work & social computing*. 1584–1595. https://doi.org/10.1145/2818048.2820013

[125] Agnieszka Walczak and Louise Fryer. 2017. Creative description: The impact of audio description style on presence in visually impaired audiences. *British Journal of Visual Impairment* 35, 1 (2017), 6–17. https://doi.org/10.1177/0264619616661603

[126] 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*. 1–12. https://doi.org/10.1145/3411764.3445347

[127] Kara Warner. 2023. *Discover the Art of Audio Description with 'All the Light We Cannot See'*. https://www.netflix.com/tudum/articles/all-the-light-we-cannot-see-aria-mia-lorbeti-audio-introduction

[128] Margot Whitfield, Raza Mir Ali, and Deborah Fels. 2016. LiveDescribe web redefining what and how entertainment content can be accessible to blind and low vision audiences. In *Computers Helping People with Special Needs: 15th International Conference, ICCHP 2016, Linz, Austria, July 13-15, 2016, Proceedings, Part I 15*. Springer, 224–230. https://doi.org/10.1177/0145482X1210600304

[129] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, et al. 2019. Disability, bias, and AI. *AI Now Institute* 8 (2019). https://ainowinstitute.org/wp-content/uploads/2023/04/disabilitybiasai-2019.pdf

[130] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1180–1192. https://doi.org/10.1145/2998181.2998364

[131] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. 4507–4515. https://doi.org/10.1109/ICCV.2015.512

[132] YouDescribe. 2023. *YouDescribe - Audio Description for YouTube Videos*. https://youdescribe.org/

[133] YouTube. 2016. *The latest YouTube stats on when, where, and what people watch*. https://www.thinkwithgoogle.com/data-collections/youtube-stats-video-consumption-trends/

[134] Beste F Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A Miele, and Ilmi Yoon. 2020. Human-in-the-Loop Machine Learning to Increase Video Accessibility for Visually Impaired and Blind Users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 47–60. http://doi.org/10.1145/3357236.3395433

[135] Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning. *arXiv preprint arXiv:2311.17435* (2023). https://arxiv.org/pdf/2311.17435.pdf

[136] Lotus Zhang, Jingyao Shao, Augustina Ao Liu, Lucy Jiang, Abigale Stangl, Adam Fourney, Meredith Ringel Morris, and Leah Findlater. 2022. Exploring Interactive Sound Design for Auditory Websites. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16. https://doi.org/10.1145/3491102.3517695

[137] Zhuohao Jerry Zhang, Smirity Kaushik, JooYoung Seo, Haolin Yuan, Sauvik Das, Leah Findlater, Danna Gurari, Abigale Stangl, and Yang Wang. 2023. {ImageAlly}: A {Human-AI} Hybrid Approach to Support Blind People in Detecting and Redacting Private Image Content. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)*. 417–436. https://www.usenix.org/conference/soups2023/presentation/zhang

[138] Luowei Zhou, Chenliang Xu, and Jason Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32. https://doi.org/10.1609/aaai.v32i1.12342

A ALL PARTICIPANT-SPECIFIC SCENARIOS

Table 6: All 17 of our participant-specific video probes. Titles are hyperlinked to the video on YouTube. Some videos are linked at specific timestamps to capture the video start time used during our studies.

Pseudonym	Type	Platform	Goal	Description	Title / Link
Alice	Vlog	TVS	Engage with friends / entertainment	A day in the life of a software engineer	day in my life as a software engineer in NYC * in-office edition *
Blair	Sports	TVS	Entertainment	A highlight reel of a professional women's tennis match	Beatriz Haddad Maia vs. Leylah Fernandez | 2023 Montreal Round 2 | WTA Match Highlights
Colin	Music video	TVS	Entertainment	Music video for *Take On Me* by a-ha	a-ha - Take On Me (Official Video) [Remastered in 4K]
Diana	Video game / how-to	TVS	Learn how to do something	A playthrough of the video game, *The Last of Us*	What Have I Gotten Myself Into... * The Last of Us First Playthrough * Part 1
	Comedic	SNS	Entertainment	A short video of a dog	Dogs funny reaction to entering optical illusion rug! #shorts
Emily	How-to / DIY	TVS	Learn how to do something	Tutorial for DIY desk upgrades	DIY Desk Upgrades
Felix	Action / foreign language	SS	Entertainment	A fight scene from a Korean drama, *Vincenzo*	Vincenzo Cassano – Tailor Fight Scene
Grace	Cooking / how-to	TVS	Learn how to do something	A video explaining how to cook four different meals	4 Meals Anyone Can Make
	Informational	TVS	Learn about a person, event, or idea	A video about Manhattan's grid plan	Where Manhattan's grid plan came from
Haley	Music video	TVS	Entertainment	A montage of animated characters engaging in adventure	Theme Song | Elena of Avalor | @disneyjunior
Isaac	Music video	TVS	Entertainment	Music video for *Bad Blood* by Taylor Swift ft. Kendrick Lamar	Taylor Swift - Bad Blood ft. Kendrick Lamar
Julia	Live theater	TVS	Entertainment	A musical number from the Broadway show, *Moulin Rouge*	Moulin Rouge! The Musical on Good Morning America
Karla	Exercise / how-to	TVS	Learn how to do something	A fitness instructor goes through a warm-up routine	Pumped Up Cardio Warmup! (Easy, fun, at home workout)
Layne	Video game / how-to	TVS	Learn how to do something	A playthrough of two minigames in *Mario Party*	Mario Party Superstars ALL MINIGAMES!!
Mason	Cooking / how-to	TVS	Learn how to do something	An instructional video of five minute meals	7 Recipes You Can Make In 5 Minutes
Nicki	Music video	TVS	Entertainment	Music video for *You Belong With Me* by Taylor Swift	Taylor Swift - You Belong With Me
Oscar	Action	SS	Entertainment	A fight scene from *The Avengers*	Thor vs Hulk - Fight Scene - The Avengers (2012) Movie Clip HD

B SUMMARY OF FINDINGS

Table 7: Summary of desired details and output modalities for nine different scenarios, represented here by their video types.
Fantastical = Science Fiction, Fantasy, and Animation
Drama = Historical, Romance, Reality, and Drama

		How-To	Info / Edu	Short-Form	Music	Live	Personal	Fantastical	Comedy	Drama
Desired Details	Actions	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Text on Screen	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Subjects		✓	✓	✓		✓	✓	✓	✓
	Settings		✓	✓	✓		✓			
	Clothing			✓	✓		✓	✓	✓	✓
	Visual Aids	✓	✓			✓				
	Scene Changes	✓	✓			✓				
	Facial Expressions							✓	✓	✓
	Visual Effects			✓	✓			✓		
	Equipment	✓								
Output Modalities	Ambient Sound	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Additional Resources	✓	✓	✓	✓			✓		✓
	Audio Cues	✓	✓	✓		✓			✓	
	Tactile Graphics	✓	✓	✓						
	Text to Braille	✓			✓					
	3D Models		✓					✓		
	Background Music			✓	✓					
	Vibration				✓					