Title

"Better Than Nothing" Or Not Enough? User-Centered Reflections On AI-Generated Audio Descriptions Across Media Formats

ISBN

9798273395756

Author

Yee, Diana

Publication Date

2025-12-09

UNIVERSITY OF CALIFORNIA

SANTA CRUZ

"BETTER THAN NOTHING" OR NOT ENOUGH? USER-CENTERED REFLECTIONS ON AI-GENERATED AUDIO DESCRIPTIONS ACROSS MEDIA FORMATS

A thesis submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE

in

COMPUTATIONAL MEDIA

by

Diana Yee

December 2025

The Thesis of Diana Yee is approved:

____

Professor Sri Kurniawan, Chair

____

Professor Kate Ringland

____

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

Abstract

"Better Than Nothing" Or Not Enough? User-Centered Reflections On AI-Generated Audio Descriptions Across Media Formats

By

Diana Yee

AI-generated audio descriptions (AI ADs) offer scalable solutions for making visual media accessible to blind and low-vision (BLV) audiences. Nevertheless, little is known about how BLV users experience and evaluate these descriptions across emerging platforms. In this qualitative study, we conducted semi-structured interviews with ten (N=10) BLV participants, recruited based on divergences in prior survey ratings, to explore their perceptions of both human- and AI-generated ADs in contexts ranging from traditional film to short-form video and livestreams. Thematic analysis revealed four key themes: (1) information prioritization and genre-sensitive details, (2) the social dynamics of shared viewing, (3) the "better-than-nothing" consensus tempered by emotional contextual gaps in new media accessibility deserts, and (4) the artistry-precision dilemma. Our findings highlight the need for adaptive, transparent, and user-informed AD systems that balance narrative resonance with efficiency. We concluded the design recommendations for co-designing AI-assisted accessibility tools in partnership with BLV communities.

Fig.1. Visualization of **high-frequency terms** from participant interviews, highlighting central concepts such as narrative flow, independence, context, and precision in audio description.

Author's Note

This thesis incorporates data that were also used in an ACM conference poster authored by the research team. While the dataset is the same, all analyses, interpretations, and written content presented in this thesis are original and extend beyond the scope of the prior publication.

1. Introduction

Audio descriptions (ADs) serve as a crucial accessibility channel for blind and low-vision (BLV) individuals by providing verbal accounts of visual media (Bitter, 2012). AD fills in essential narrative elements, supports comprehension, and enables participation in the cultural and social practice of shared media viewing. Nevertheless, AD remains inconsistent in quality, availability, and relevance across platforms. While traditional AD guidelines were formulated with television and cinema in mind, the current media ecosystem, with livestreaming, video clips, and short-form content, introduces new accessibility challenges.

Newer media platforms such as YouTube, TikTok, Instagram Reels, and short-form livestreaming represent what many BLV viewers experience as “accessibility deserts.” These environments rarely include human-narrated AD, leaving substantial gaps in access for content that is now central to contemporary media consumption. Besides, the rise of generative AI offers scalable solutions for AD. Still, these systems often fail to capture the narrative cohesion, emotional resonance, and cultural adaptability that BLV audiences value (Gonzalez et al., 2025; Cheema et al., 2024). These cases reveal a deeper challenge: automated systems tend to view AD as a captioning task rather than a user-centered narrative practice.

Across both traditional and emerging media contexts, a fundamental question persists: **what information matters most for BLV viewers?** Prior work has examined guideline structure and automated pipelines, yet relatively little is known about how BLV viewers interpret AD in real-life situations. Even less is known about how AD acts in social dynamics during shared viewing, or how expectations shift across genres.

To address this gap, semi-structured interviews were conducted. The interviews were designed to uncover the reasoning behind BLV audiences’ preferences and their evaluations of AD designs. The goal of this qualitative approach is to capture the interpretive, contextual, and socially situated nature of AD use.

Specially, we investigate the following research question:

- **RQ:** How do BLV users experience and evaluate audio descriptions, including AI-generated content, across different media formats and viewing scenarios?

2. Related Work

2.1 Foundations of Audio Description and Guideline Development

AD initially emerged as an accessibility practice designed to provide verbal access to visual media for BLV audiences in the United States (Andressend, 2023). Its subsequent expansion across Europe, reflecting region-specific practices rather than unified standards during the early 2000s (Bitter, 2012). Early AD guidelines targeted broadcasting institutions, translation organizations, and national accessibility bodies that sought to standardize production workflows for describers rather than respond to the experiential needs of end-users.

Foundational AD guidelines are largely procedural, established conventions around neutrality, objectivity, and consistency, instructing describers to follow a “saying what you see” philosophy that discourages inference or emotional interpretation (Starr & Braun, 2024). Jiang et al. (2024) similarly note that authorship guidelines determine “what content to include or which tone of voice to use,” reinforcing AD as a procedural, non-interpretive task. Timing conventions also play a central role. Guidelines require that AD avoid overlapping with dialogue or other critical audio elements, meaning that describers must fit information into narrow and often unpredictable gaps in the soundtrack (Liu, 2021; Natalie, 2024). These constraints reflect the assumptions of traditional film and broadcast environments, where consistent pacing and discrete pauses formed when and how description could be delivered.

Contemporary systems research often synthesizes these established standards to support tool development or evaluate automated workflows. Pavel et al. (2020) compiled representative operationalized AD guidelines. Their synthesis makes explicit several principles that recur across early standards: describers should “describe important visual content essential to comprehending the video” while “avoiding describing content that can be inferred from the audio and following a general-to-specific pattern. The guidelines further mandate that describers “match the tone and vocabulary of the source video,” use concise language, and “being objective and avoid[ing] editoralizing.” These synthesized rules demonstrate that AD conventions historically prioritized informational economy, tonal neutrality, and precise timing-bound placement.

As Starr and Braun (2024) argued, describers inevitably engage in interpretive decision-making even when guidelines caution against it, since narrative cues, character motivations, and emotional context cannot be conveyed without some degree of inference. User-centered research similarly demonstrates that BLV comprehension relies on relational details, atmospheric cues, and contextual framing that extend beyond early guideline conventions (Liu et al., 2021). Moreover, guidelines developed for traditional film do not map cleanly onto today's diverse media landscape, which includes short-form video, livestreams, participatory content, and increasingly AI-mediated workflows. As media formats evolve, the gap between guidelines and user needs widens. These foundational standards provide structure but are limited in their ability to account for the interpretive work, pacing demands, and genre variations that shape viewers' experiences of contemporary media.

2.2 BLV User Preferences, Narrative Needs, and Media Context

Empirical research consistently shows that BLV audiences rely on contextual, relational, and narrative cues that extend beyond the prescriptive boundaries of traditional AD guidelines. Liu. et al. (2021) find that participants repeatedly sought additional information when key visual or auditory cues, such as visual references, text, or sounds, were missing. These omissions make it harder to follow who or what is involved and to understand how events connect across a scene, leading audiences to seek more contextual information during viewing (Gonzales, 2024). Accessibility was diminished when contextual anchors or linking cues were omitted, suggesting that understanding relies heavily on narrative coherence rather than literal visual transcription. Jiang et al. (2024) similarly noted that existing guidelines do not account for the variability in BLV users' needs across media types, and that BLV viewers rely on cues to understand connections among actions, narrative shifts, and the broader context of a film. Across these findings, it is clear that BLV audiences prioritize narrative grounding rather than strictly literal descriptions.

BLV information needs also vary substantially by viewing scenarios. Jiang et al. (2024) report that preferences shift across instructional, entertainment, and narrative contexts: procedural videos require step-by-step detail and descriptions of object manipulation; expressive or social content elicits requests for relational or emotional cues; and fast-paced or visually dense formats demand orientation markers, transitions, and spatial cues. These findings echo those of Liu et al. (2021), whose participants sought additional context when encountering unfamiliar settings, implied relationships, or rapid visual changes. Such scenario-dependent preferences challenge the assumption that a single descriptive strategy can serve across genres and platforms, and highlight how guideline-bound AD fails to reflect the diversity of BLV viewing practices.

Differences in visual history further complicate BLV information needs. Chmiel and Mazur (2021) show that congenitally blind participants often prefer explicit labeling of emotions, facial expressions, and visually encoded meaning. In contrast, participants with acquired blindness rely more on residual visual memory and may choose less explicit detail. This variation demonstrates that BLV audiences do not form a homogenous group and that standardized approaches may not accommodate the full range of perceptual experiences.

Social context also plays a vital role in how BLV audiences engage with video content. Liu et al. (2021) describe how participants make strategic decisions about when to rely on AD, when to focus on dialogue or environmental sound, and when to ask for clarification from sighted companions. In shared-viewing settings, participants sometimes avoid interrupting others or drawing attention to their access needs, even when they miss key visual information. These accounts show that AD supports both informational access and social participation, underscoring the need for flexible, context-aware descriptive practices.

2.3 AI-Generated Audio Description and Human AI Collaboration

Early automated AD systems relied heavily on object detection and captioning pipelines, limiting their ability to support BLV users' narrative and relational needs. Bodi et al. (2021) evaluate two state-of-the-art video captioning AI tools: NarrationBot and InfoBot, combining YOLOv3 keyframe detection with captions generated via the Pythia model to produce baseline and on-demand descriptions without human intervention. Their system "takes a step toward removing humans from the loop," illustrating how early work framed AD as a computational pipeline rather than a narrative reasoning task. This engineering-centric approach reappears in Chuang and Fazli's 2023 CLearViD framework, which applies curriculum learning to video description models to progressively handle increasingly visually complex content.

With the rise of multimodal LLMs and VLMs, automated AD has become more fluent and scalable, but these systems still require human post-editing and struggle with temporal coherence. Gao et al. (2025) describe modern AD generation as a three-stage pipeline: dense video captioning, automatic post-editing, and evaluation, noting recent multimodal models bring the field "a step closer" to automation. At the same time, they argue that the post-editing stage "is not yet dispensable," as high-quality AD depends on integrating both "local context" and "global context," especially in long-form media. Xie et al. (2024) similarly demonstrate the limits of training-free AD generation: The AutoAD-Zero framework uses a VLM to produce dense video captions and an LLM to summarize them, performing well on static visual details but struggling with "temporal reasoning" and cross-frame relationships. Chu et al. (2024) echo these challenges in LLM-AD, which leverages GPT-4V for AD generation but may introduce hallucinations, reinforcing the need for human oversight.

BLV users increasingly rely on AI AD in daily settings, yet these tools often lack relational detail, and human-AI collaborative systems offer a more reliable and customizable path forward. Gonzalez et al. (2025) show that BLV participants use scene-description apps to gain quick “scene overviews” and identify people or objects, but frequently report issues with “accuracy,” missing relational cues, and difficulty interpreting scene changes.

Systems that incorporate user control reveal persistent gaps in the adaptability and relevance of AI-generated AD, demonstrating that existing generative model pipelines do not yet meet the diverse needs of BLV. In DescribeNow, users must actively choose between concise or detailed descriptions via keyboard shortcuts (Cheema et al., 2024), highlighting that fixed-style automated AD cannot satisfy varying genre- and context-dependent needs. DescribePro (Cheema et al., 2025) shows that describers regularly refine GPT-4o drafted through NLP and rely on silence and scene-change detection to correct model output, indicating that AI alone cannot reliably manage timing or content selection. Similarly, CustomAD (Natalie et al., 2024) finds that BLV audiences require over “length, emphasis, speed, voice, tone, and language,” and that a single AD track cannot accommodate this diversity. Across these systems, human involvement and customization are not optional improvements, but evidence of the limitations of current AI-generated AD. We extend this by grounding AI AD prompts in empirically ranked user preferences, enabling more adaptive and inclusive descriptions.

3. Methods

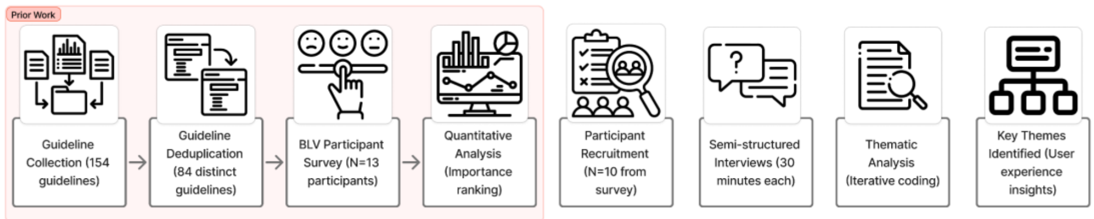

Fig.2. **High-level process overview of the study,** beginning with the collection and deduplication of 154 audio description guidelines into a set of 84 distinct items. This was followed by the iterative development, discussion, and refinement of a coding codebook to systematically analyze the guidelines. The process culminated in a survey distributed to blind and low-vision participants, with the resulting analysis forming the basis for this investigation into AI-generated audio description.

We conducted semi-structured interviews with ten BLV participants (N=10) to address critical gaps in our prior mixed-methods research on AD guidelines. The interview study builds on a previous survey (N=13) that examined BLV users' experiences and preferences for AD through 5-point Likert scale ratings of public AD design guidelines collected from *Netflix Accessibility Guidelines, Ofcom Guidelines, Media Access Canada Guidelines* (MAC)*,* and the *Described and Captioned Media Program Guidelines* (DCMP). Where the previous survey identified divergent user ratings of guideline importance, this study deployed interviews to investigate the underlying rationale for such discrepancies. All steps involved in our research have been approved by the Institutional Review Board (IRB).

3.1 Study Design

The study used a qualitative design to explore how BLV viewers experience and evaluate AD across different media contexts. The goal was to understand how participants interpret the strengths and shortcomings of both human-narrated and AI AD, and how these interpretations form their expectations for future tools.

A semi-structured interview format was selected to guide discussion while allowing participants to introduce their own priorities and examples. The interview protocol was built around two focus areas:

(1) **Guideline-specific probing.** Interview questions were customized to each participant's divergence profile, targeting 3-5 guidelines to distinguish perceptual reactions from reflective judgments. For example:

> "You rated [Guideline] X for enjoyment but Y for comprehension. Could you explain what influenced your rating?"
>
> "You rated the following guideline Z while others averaged A. How do you interpret the differences in guideline ratings among the overall survey average?"

(2) **Open contextual exploration** focusing on AI-generated AD experiences, social tension, and platform AD experiences. The interview questions were designed to be open-ended and flexible, allowing participants to elaborate on their experiences and perspectives. For example:

“How is your experience with an AI-generated AD compared to a human-narrated AD?”

“When it comes to traditional TV shows and movies versus newer media like YouTube videos or TikTok, how does your experience with AD differ, if at all?”

The specific guideline items used for each participant’s customized probing set are listed in **Appendix A**.

3.2 Participant Recruitment

Participants were recruited through a BLV-specific community center maintained by our research group. This network includes individuals who had previously taken part in an earlier study examining AD guideline preferences and expressed interest in participating in accessibility research.

For this follow-up interview study, participants were selected based on quantitative divergence thresholds derived from earlier survey data within the volunteer pool:

(1) Intra-participant divergence between the enjoyment-comprehension rating of the AD guidelines, and

(2) Inter-participant rating deviation, identifying individuals whose ratings differed significantly from those of others in the pool.

These thresholds ensured that the interview sample included participants with expectations, preferences, and points of disagreement regarding AD design.

Eligible participants were contacted via email and invited to participate in a remote interview. Each interview session lasted about 30 minutes via Zoom, depending on participant availability, and participants were compensated with a $20 Amazon e-gift card upon completing the study.

3.3 Participants Characteristics

Participants represented a diverse range of visual experiences, viewing habits, and media preferences. The levels of vision ranged from total blindness to limited light perception, as well as progressive conditions affecting central or peripheral vision. This range enabled the study to capture diverse strategies for interpreting visual content and varying levels of reliance on AD or AI tools.

The frequency with which participants consumed video content also varies. Most participants reported watching media multiple times a day, including a mix of long-form films and television, educational content, instructional or hobby-based videos, and short-form online material. About one-third of the participants consumed media several times per week. Only two participants reported barely consuming media weekly. These differences reflected a broad spectrum of reliance on visual media in daily life.

Genre exposure in the group was similarly varied. The majority of participants reported consuming entertainment content, making it the most common category across the sample. Educational or work-related content was also prevalent, with eleven participants incorporating it into their regular viewing. Instructional and "how-to" content formed another critical category, used by nine participants to support tasks such as cooking, assembling items, or learning new skills. A smaller subset of two participants watched other miscellaneous content, such as News and personal videos on YouTube.

These characteristics varied in vision levels, media frequency, and genre diversity, creating a heterogeneous sample well-suited for examining how AD and AI AD operate across a range of real-world viewing contexts.

3.4 Data Analysis

Data were analyzed using a reflexive thematic analysis approach. This method was selected to capture the nuances of how BLV viewers interpret AD, respond to missing or inaccurate information, and navigate AI AD across different platforms and genres. Analysis began with repeated readings of the interview transcripts to build familiarity with participants' descriptions, examples, and interpretive reasoning. Open coding was then conducted line by line, identifying meaningful units such as reactions to timing issues, preferences for character identification, responses to emotional inference, and strategies for handling missing information. Codes were kept close to participant phrasing to maintain specificity and avoid premature categorization.

The initial codes were then organized into broader topical categories that reflected shared concerns across the interviews. These categories included priorities for essential information, challenges specific to AI AD, expectations regarding genre, preferences for neutral language, and social dynamics surrounding shared viewing. Categories also capture strategies participants used to supplement or replace missing AD with AI tools to interpret visual content.

Relationships among categories were examined through iterative refinement and memo writing. This process produced four central themes: Information Prioritization, Social Dynamics, Better-Than-Nothing Accessibility in New Media Deserts, and the Artistry-Precision Dilemma. These themes represented the most consistent patterns in how participants evaluated description quality, interpreted gaps in understanding, and adapted accessibility constraints across different media environments.

As sighted researchers, the authors acknowledge differing lived experiences among BLV participants. Thus, the authors approached this study with a focus on listening and amplifying participants' voices rather than normative design standards, and worked to minimize bias and ensure that the interpretations reflected participants' language, priorities, and lived realities.

4. Findings

Analysis of the interview data revealed four central themes that describe how BLV audiences interpret AD across different media contexts. These themes highlight what information matters most for comprehension, how AD operates within social dynamics, how AI AD fills and complicates access gaps, and how genre conventions influence descriptive expectations.

4.1 Information Prioritization

Participants emphasized the importance of context-driven filtering of AD content that meaningfully contributes to understanding what is happening on screen. While narrative-relevant details were deemed essential, excessive information generated cognitive overload. Participants described several dimensions of information prioritization that shaped whether a description supported or disrupted their comprehension.

4.1.1 Character Tracking and Narrative Continuity

Effective AD depends on providing the information that enables viewers to follow the story as it unfolds. This includes highlighting actions, gestures, and changes in setting that move the narrative forward, as well as offering cues that help listeners track narrative progress.

Character identification was a central expectation. Participants explained that vague descriptors would not help recognition when the same person appeared later. As one participant put it, *“a man with a mustache walks into the room and has a quizzical look on his face. A man with a mustache means nothing. Saying Mr. Darcy. OK, perfect. I got that character. I'm following the arc of that character throughout the whole storyline.”* (P1) Others emphasized the need for consistent visual markers that make recurring characters identifiable across scenes. As one participant explained, *“there is oftentimes in movies or TV shows where a character isn't, you know, ‘oh, this is Jane,’ unless it's something consistent in what Jane is. Like Jane is always wearing the red shirt.”* (P4) Participants also discussed how AD should select identifiers that are helpful rather intrusive: *“If there's nothing that would kind of trademark her, then I think it's okay to jump to race or ethnicity but if it's not going to be germane to the story, they're not like a central character where you would, I would at least want like a clear picture of, a race or an ethnicity if they're just kind of a secondary thing.”* (P4)

Storyline cues were described as equally important for understanding temporal and spatial shifts. Participants valued brief markers that anchored them in the narrative's structure. One participant explained, *When the audio describer tells his blind audience to cut to a flashback. Oh my God, that is so powerful. Cut to a flash forward. Because now I know how to follow the story. I've just cut forward, I've just cut back in time in the plot. And without those markers, I'm catching up with what just* *happened."* (P1) Location cues served a similar purpose: *"cut to a pasture…I'm instantly clued in to where the scene cut to so I can then resume my analysis of the dialogue and other sound elements of the scene."* (P1) Participants also noted that certain visual expressions functioned as cues connecting one moment to the next. As one participant explained, *"the facial expressions of the character sometimes are significant to the storyline or what's happening, if the person has a sad face or disappointed face. I think those are important details; if you omit those, then you may not be able to. Figure out what's exactly happening."*(P8)

Visual references tied to culture, period, regional terminology, or subtle gestures often carry meaning that is not accessible without sight. They relied on AD to name or contextualize these elements so that culturally dependent visuals did not become confusing or misinterpreted.

4.1.2 Cultural Clarity

Understanding culturally specific references often requires background knowledge that is not visually available to BLV viewers. AD needs to prioritize the elements that carry cultural, regional, or historical meaning so listeners do not miss information. Participants described several situations where AD must briefly name or clarify items, vocabulary, or cultural cues to prevent confusion.

This expectation was reflected in the statement, *"People may not have information or definitions of words being used based on what they're watching, especially if it's something from a different culture or period…There are a lot of things that we won't know what an object looks like or what something is called. Based on this, the thing is a piece that was worn during this time or looks like this. Those details are why we need the audio description."* (P2) **Cross-cultural terminology** also shaped understanding. Participants noted confusion when region-specific terms (e.g., "torch" vs. "flashlight") were not adapted, especially in multilingual content. *"A typical American audience may not understand what specifically those are unless they're knowledgeable of British terminology."* (P7)

Participants described gestures and subtle signals as forms of cultural or contextual meaning that AD must surface, since sighted viewers naturally interpret them. As one participant noted, *"a head nod or a finger gesture or a head turn to the animal when they're referring to Eddie. And it's those sorts of just subtle introductions of additional elements which increase significantly the enjoyment factor for me."* (P1) Language clarity also played a role. Participants explained that comprehension depended more on the clarity of language than on accent-matching: *"It's much more important that what is said is in my language…all the different dialects of how English can be spoken. As long as the narrator is not taking away from the scene by using words or language which is inconsistent with an appropriate description of what's going on, nestled perfectly within the dialogue and the other audio landscape* *of the scene, that to me is the most important."* (P1) Cultural clarity, therefore, acts as a bridge between what the scene shows and what the listener can interpret.

These perspectives show that narrative continuity depends on AD prioritizing the identifiers and cues that allow listeners to recognize characters, follow temporal and spatial shifts, and understand how scenes relate to one another. When these elements are clearly conveyed, participants can follow the storyline more effectively and remain grounded in the unfolding narrative.

4.1.3 Credits and Career Tracking

Information prioritization also extended to how credits were handled. Participants did not view credits as optional; instead, they described them as meaningful when they helped them follow creative contributors or understand who played which character. At the same time, credits to overshadow narrative-relevant details are not preferred.

For example, participants valued **credits** when tied to their personal interest in **career tracking**. As one participant asserted, *"I follow the careers of certain actors, certain producers, certain directors, editors, writers, and I need to hear which actor played which character if that information is in the credits."* (P1) However, an exhaustive listing of technical roles was seen as unnecessary. For some participants, credits held less importance than the film's descriptive content. One participant described this balance: *"I like to know the opening credits and the closing credits. But I would* *rather hear about the scenery if it's important or the appearance of a character…I care more about the actual content than the producer, director, or characters. Those things I could always look up after the fact if I was really curious about them.”* (P4) Another participant echoed this prioritization, valuing only select pieces of credit information: *“I don't need to know the director, the producer, the writers…but if we have time and you can put the actor's name who plays who, that would be interesting.”* (P10)

Participants differed in how they valued credit information, but agreed that credits matter when they support meaningful engagement with the cast or creative contributors. Some viewed credits as necessary for following actors’ and creators’ careers, while others preferred credit information only when it enhanced understanding of the film.

4.2 Social Dynamics of Shared Viewing

Accessing visual information is not only a matter of individual comprehension but also a social experience shaped by how participants watch films with others, when they feel comfortable asking for support, and how they negotiate independence. These dynamics influenced how they used AD when watching with others.

4.2.1 Inclusion, Coordination, and Social Comfort

AD supported, allowing BLV viewers to participate more fully in shared media experiences. Participants emphasized that AD helped them follow the storyline in real time with friends and family, reducing the need for interruptions and enabling smoother coordination in group settings. AD also enriched experiences for sighted viewers, who often found value in the descriptive layer.

One participant explained that sighted companions frequently preferred to keep AD turned on because it offered additional insight: *"We just always have on audio description because they like it. They feel like it helps them, maybe picking up on things they didn't even know."* (P8) In some cases, sighted viewers became enthusiastic adopters of AD. One participant described, *"With my sighted partner, she won't watch a movie now without audio description because she loves, for example, to puzzle. She likes to put together puzzles while she's watching a movie."* (P1) Even subtle descriptors could help sighted viewers follow interactions they might overlook. As one participant observed, *"Sometimes the little descriptors will come up and it'll mention something, it'll refer to something that even the people who don't have visual problems will read, and it will give them some kind of insight into something in the movie. Sometimes I feel like it can enhance it for them if they're actually reading it."* (P5)

AD provided more details in the scene and a small teaser to the social dynamic, where one participant highlighted, *"It's pretty cool. So we get like a kind of a sneak peek at what's about to happen in the next few seconds."* (P7) AD also demonstrates its benefits during musical performances, where sighted viewers gained clarity about movement and staging. One participant recounted: *"The sighted people had never experienced a movie, let alone a musical with audio description. And they thought it was useful because the audio describer described which characters were coming on stage, which ones were leaving, and it was being said aloud."* (P1)

4.2.2 Social Burden and Independence

While AD supported inclusion in shared viewing, participants also described the social burden involved in deciding when and how to use it. Watching with sighted companions introduced concerns about drawing unwanted attention, disrupting others' experience, or relying too heavily on friends and family for support. These tensions shaped how participants balanced social comfort with their own need for access.

Some participants worried that AD might be distracting for sighted viewers, especially those unfamiliar with hearing descriptive tracks layered over dialogue. One participant explained, *"most people don't mind that it's on, but I do think some people find it kind of distracting to watch a movie with me, and then it's got the subtitles on and they're trying to kind of watch the action and I'm sort of trying to read the words* *to make sure I'm not missing anything visually."* (P5) Another participant echoed this concern, noting that they often avoided requesting AD in group settings because it felt intrusive: *"when I watch things with them, it's not something they're used to, because it wasn't something that we kind of grew up doing, other groups maybe that does kind of limit their enjoyment and so it's kind of a situation where I don't necessarily if I'm with other people ask for it because I don't want them to be like this is distracting."* (P4)

Participants faced persistent logistical and social barriers. One highlighted synchronization failure is *"impossible to get an audio-description channel to sync with what's on their TV."* (P9) Participants did not want to add labor or discomfort to the group; individualized ways to access AD were preferred. The participant suggested a solution of **personal streams** that would reduce social tension, *"the wireless headset, it would be good if there were more kind of home-based options for things like that, so that audio description could sort of stream to just one person in a group"* (P4) At the same time, participants call out that AD played an essential role in maintaining independence. Instead of interrupting others or feeling apologetic about asking repeated questions, AD allowed them to follow the story on their own. As one participant stated, *"It's honestly so much better for me to enjoy the film without asking the other person, what's happening? What are they doing? I feel like this is especially crucial in films where it's like a moment of silence."* (P8) Participants preferred social independence by avoiding situations where they needed to rely on companions for visual information, describing that dynamic as burdensome for both sides. AD reduced these pressures, letting them enjoy films without negotiating when to ask for help or worrying about taking others out of the experiences.

4.3 "Better-Than-Nothing" Consensus in New Media Deserts

Participants consistently described AI-generated AD as helpful, mainly because it is accessible when no human-narrated description is available. Their accounts show appreciation for the availability of AI descriptions; at the same time, participants also reflect ongoing doubts about their accuracy, relevance, and naturalness. Across interviews, AI was consistently framed as a convenient fallback rather than a reliable or preferred alternative.

New media platforms like TikTok and YouTube were described as **"accessibility deserts."** As one participant put it, the chance of finding usable AD on short-form content is nearly extremely low, *"because I know that the likelihood of it being audio described is low…99.999% of these YouTube instructional videos are not audio described, they're not worth it."* (P1) This characterization of YouTube instructional videos as almost entirely inaccessible highlights the extent of the gap between traditional media and newer platforms. Although only two participants in the prior survey reported engaging with short-form videos, half of the interview sample reported engaging with such content from TikTok and YouTube. Their experiences suggest that AI AD is not a preference, but rather a practical response to the widespread absence of human-narrated AD in these environments.

Participants also described the strategies they use when AD is missing. One participant explained, *"I'll pause the video, take a screenshot of that particular frame, and then I will run that picture through AI and see if I can gather any information about the video. And that usually helps to figure out kind of what the video was about."* (P8) The participants also noted, *"I just get past it because there isn't really a way to figure out what's going on in the video."* (P8) These examples show how some viewers navigate the lack of AD on short-form platforms. In these settings, AI tools are used because they are the only available option, not because participants find them particularly compelling. As a result, AI is valued for enabling some level of access to content that would otherwise be impossible to follow.

Conversely, sports broadcasting surfaced as a counter-model. A participant compared current AD to radio-style sports narration, which he viewed as a much more user-centered form of description. As the participant described, *"the radio person assumes that the user cannot see what's going on, and so gives a fuller description of the action, who's doing what."* (P3) This comparison raised the concern that current AD experiences, including AI-generated descriptions, do not follow the same user-centered design principles. Participants noted that radio narration sets a strong example because it begins from a clear understanding of the listener's position. In contrast, many AD and design guidelines start from the existing visual product and then fit the description into the remaining gaps. Several participants noted that AD can miss key actions or provide details that do not help them follow the scene. The comment about radio narration suggests that participants want AD practices and guidelines that start from a similar assumption to radio's, treating non-visual access as the default rather than an afterthought.

4.3.1 AI is Better Than Nothing, but Still Feels Limited

Participants expressed a pragmatic "better-than-nothing" stance, accepted AI AD functionality as perceptually adequate for accessibility gaps, but still fell short of human narration. Several participants explicitly stated that having AI-generated AD was significantly better than having no description, as one participant explained: *"The difference between having any type of audio description, AI or in person, is worlds apart, better than nothing."* (P3)

Another participant expressed a similar idea: *"I saw one movie that I know had AI, and if I know it's AI, then it wasn't that good; it was actually decent. I may get the job done… I'll tell you, it's a lot better than nothing. If I had to rate it on a scale from 1 to 10, I'd give AI a 6."* (P10) Participants explained that AI becomes especially important in visually dense content where visual information carries the story, as the participant described *"without it, some movies may not need it because the plot isn't real deep, but if it's got a plot where there's a lot of visual stuff going on that fills in* *the gap, then it's real important."* (P10) Others emphasized that AI tools also increase independence in related tasks, such as interpreting images or documents with tools like Be My Eyes. As the participant noted, *"I think that AI is so much better. For one, I can use it whenever I need to, and without having to have another person around. So it definitely provides more independence. And two, I can ask very specific questions, as many as I want, until I have gathered all the information that I need."* (P8) Another participant reinforces this advantage, "*any video I want, I can have it described."* (P6) These comments show that participants can notice when AD is AI-generated and that they do not consider it high quality. At the same time, they are very clear that they would rather have AI AD than no AD at all. However, they also recognized apparent limitations, which are outlined in the following subsections.

4.3.2 Irrelevant Detail Capture In AI AD

A recurring issue raised in the interviews involved AI describing information that did not support comprehension. Several participants observed that AI often highlighted visually striking elements without considering narrative relevance. For example, one participant who regularly used PixieBot to interpret videos explained, *"It's still kind of messy because it focuses in on the wrong things. I feel like AI doesn't understand the whole video. It understands bits of it. So it focuses in on the wrong things."* (P6) Another participant shared a similar frustration where discussing scenes where AI described minor visual features that did not matter. *"For example, there's a scene where the person is wearing shiny shoes. The fact that they're shiny shoes means* *almost nothing to the scene…The artificial describer might capture that and say something about it."* (P1) This tendency to focus on isolated visual elements created confusion and disrupted the flow of the story for some participants. This participant also pointed out the lack of narrative judgment in current AI tools: *"With real estate being so precious, everything being said needs to be really relevant. I think there's a huge risk that audio description produced by artificial intelligence, they're not going to appreciate the relevance of what's being said during the precious moments they have to say it."* (P1) These examples show that while AI can identify objects or visual features, it often lacks an understanding of how those features fit within the overall meaning of a scene. This limitation directly raised concerns that important information might also be left out.

4.3.3 Missing Information and Context Fragmentation

Participants also described frequent gaps in AI AD. Instead of receiving a continuous, cohesive description, several participants felt that AI provided fragments that did not capture the whole sequence of events. One participant highlighted the central concerns, *"how accurate is this description? What is missing from this description?"* (P2)

Several interviewees explained that these gaps made it difficult to determine whether the AD was accurately capturing the content. Participants explained that missing information could be as simple as omitting an action that triggers a scene change or neglecting to explain how one moment connects to the next. The concern was not only about errors, but also about omissions that affected their confidence in following the story. The earlier comment from P6 in section 4.3.2 points to one reason behind this pattern. In addition to focusing on irrelevant information, the participant noted that *"AI doesn't understand the whole video" and that it* often processes videos *"in bits"* (P6), suggesting the system may not track events across time cohesively.

Due to these gaps, participants occasionally had to work around AI AD's limitations to understand what they were watching. Overall, missing information and fragmented sequences made AI AD feel less dependable, even though it still provided some access in environments where no other options existed.

4.3.4 Disconnected Emotional Meaning With AI AD

Another challenge described by participants was AI's tendency to assign emotional interpretations to scenes, even when the underlying visual cues were ambiguous. Participants reported that AI AD often exaggerated positive emotions, resulting in a tone that did not match the actual context. One participant explained, *"I feel like AI description often kind of paints everything in a really positive light, warm and fuzzy, but it's sort of a situation where not every scene is heartwarming, or not everybody is just plain smiling."* (P4) The participant went on to describe how these moments affected their viewing experience, where AI stated emotional reactions outright instead of describing observable cues: *"I don't like for there to be an inference made* *about what it means or the emotion. I don't like hearing things like their face lit up with excitement. If I can tell by what the character is saying that they're excited, I kind of feel like that almost dumbs it down a little bit. I'd rather as a viewer be totally open and able to make my own discernment of the emotions behind things."*(P4) Furthermore, another participant noted that tone labeling felt unnecessary and sometimes patronizing, as the participant insisted, *"just give me the facts. Don't give me opinionated language about what's on the screen; it doesn't need to say, like, so-and-so said something in a whatever tone of voice. That to me is just very leading and obvious and kind of insulting."* (P5) Together, participants preferred AI AD to remain neutral and descriptive rather than interpretive. The desired factual cues and the felt emotional inference created distance rather than clarity. This concern ties directly into broader expectations for AD to support comprehension without shaping the viewer's interpretation.

Participants stress that trust requires **transparent labeling of AI-generated** content. They wanted clear labeling so that they could adjust their expectations before the film began. As one insisted, *"I would love for there to be a standard that if it is AI-generated and synthetically voiced that be a label right up front that said that the audio description of this movie has been AI-generated."* (P1) Furthermore, the participant extended this point with the demand of ***"human supervision"*** ensure accuracy, explaining that *"the AI suffered no negative consequences of being wrong, or inaccurate or irrelevant."* (P1) These concerns reinforce that even when AI fills access gaps in media deserts, participants still expect accountability and transparent disclosure to maintain trust in what they consume.

4.4 Artistry-Precision Dilemma

A recurring tension in participants' accounts involved how AD should balance neutrality with narrative richness. Participants valued AD that conveys essential visual information with clarity, yet they resisted language that imposed subjective or emotional framing. They described wanting the space to form their own interpretations rather than being guided toward a particular reading of the scene.

Participants rejected descriptive language that relied on personal judgment, such as 'attractiveness' or 'emotional tone'. Subjective labels such as "pretty" or "handsome" were rejected: *"pretty is just a matter of somebody's opinion, not necessarily a universal thing…I'd rather be given non-judgmental descriptors on my kind of make."* (P5) Another participant preferred **autonomy in interpretation**, where AD should avoid asserting emotional meaning, even if mistaken. Since *"even for sighted people…watching something looks exciting based on their gestures…that doesn't mean that everybody interprets that to be the same thing. So I'd rather have the autonomy to be wrong in what I'm thinking than have someone feed me an opinion I may not share."* (P4) These comments demonstrate the desire for AD that preserves interpretive independence. Participants sought descriptions grounded in observable detail rather than inference, enabling them to construct emotional and aesthetic judgements.

At the same time, genre-based adaptation resolved some contradictions. Participants universally endorsed **the dynamic density of descriptions** between artistry and prevision. Descriptions that feel intrusive in one genre may be insufficient in another. For musical films, participants wanted the richness of the audio performance to take precedence, with minimal interruption from AD. The participants describe this balance clearly: *"hear the quality of the singing, then the description cutting into that is not something I'd want."* (P3) In contrast, action-heavy films require more frequent and detailed descriptions to track fast-paced movement and visual stakes. The same participant highlighted this need when recalling a superhero film, *"the big ending fight in Avengers: Endgame. I want a lot of descriptions of what's going on. For me, the action has primacy."* (P3) Participants also emphasized that effective AD depends on timing. Descriptions should fit naturally into available audio space and avoid competing with dialogue or key auditory information. As one participant explained, *"If you take advantage of the moments of silence… or you have to omit certain details because there is not enough time… It's very difficult to hear both things at the same time."* (P8) This genre-based calibration highlights that participants do not view AD as a one-size-fits-all practice. Instead, they expect AD to adapt to the pacing, sensory priority, and conventions of different film types.

5. Discussion

The finding illustrates how BLV viewers interpret AD across both traditional media environments and the rapidly expanding AI AD ecosystem. By situating participants' accounts alongside existing work on AD guidelines, BLV information needs, and automated AD systems, the discussion highlights how current descriptive practices fall short of capturing the complexity of BLV viewing. Each theme contributes to a broader understanding of how viewers negotiate relevance, narrative continuity, emotional neutrality, genre expectations, and social comfort.

5.1 Rethinking Information Prioritization Beyond Traditional Guidelines

BLV viewers prioritize narrative and contextual information that extends beyond what traditional AD guidelines standardize. Early AD guidelines have long foregrounded brevity, neutrality, and a "say what you see" philosophy (Starr & Braun, 2024), which limits descriptive additions to only what can be visually verified. While this approach attempts to minimize interpretation, participants consistently described the need for cues that connect scenes, clarify character relationships, identify recurring characters, and mark shifts in time or setting. These cues are often taken for granted by sighted audiences, yet they are essential to BLV comprehension.

Prior work has similarly recognized that strict literalism may omit critical narrative linking information (Liu et al., 2021; Starr & Braun, 2024). Participants in this study reinforced this concern, noting that missing character names, gestures, or motivations led to confusion or delayed understanding. They also highlighted the importance of culturally or regionally specific terminology, underscoring the need for AD that supports cultural clarity rather than assuming shared visual knowledge. These expectations reflect a shift seen in recent BLV-centered research, in which descriptions must adapt to narrative significance rather than solely to surface visual features.

5.2 Social Dynamics Reveal AD as Both Access and Social Infrastructure

ADs are not only access tools but also a social infrastructure within shared viewing environments. Participants described AD as essential for maintaining independence, avoiding constant requests for clarification, and enabling equal participation. These accounts reflect broader accessibility research, noting that BLV media engagement is formed by social comfort, pacing, and the desire to avoid social burden.

At the same time, participants expressed concerns about distractions or disruptions to the flow of group watching. This tension reveals how AD use is embedded in social relationships, influencing decisions about whether to enable AD and how comfortable viewers feel voicing their needs. While prior AD literature addresses accessibility from an individual perspective, it rarely considers how AD mediates group participation. Participants' suggestions for individualized audio stream and more flexible AD options show that AD design must consider both personal and collective viewing experiences.

5.3 AI AD Expanding Access but Compromising Narrative Cohesion

Across interviews, participants described AI as valuable primarily for its accessibility, "better than nothing" when no human AD is available, especially on short-form platforms that often lack descriptive support. This perspective aligns with recent research noting that AI AD fills accessibility gaps in environments where traditional AD is rare.

However, participants also identified consistent challenges in the current AI AD. These included irrelevant detail selection, fragmented descriptions that fail to connect scenes across time, and emotional overreach that misrepresents tone or character intent. These observations echo ongoing findings in AI AD research that multimodal systems excel at object-level recognition but struggle with narrative coherence, temporal reasoning, and context integration (Xie et al., 2024; Chu et al., 2024).

Participants' comparisons between AI AD and radio-style sports narration further underscore these limitations. Radio narration assumes a listener unable to visualize action and foregrounds user-centered clarity. In contrast, current AI pipelines are typically optimized for generating dense, frame-level captions rather than descriptions tailored to user comprehension. Participants' comments highlight this gap and underscore the need for higher-level reasoning in future AI models that can track characters, motivations, and event cues across an entire scene.

5.4 Artistry-Precision Dilemma and Limits of Literalism

Participants strongly preferred neutral, factual descriptions over interpretation, yet they also expressed a need for expressive cues in specific genres, such as action films, where movement clarity is essential, and musicals, where description should not intrude on vocal performance. These seemingly contradictory expectations highlight the contextual nature of AD preferences.

This duality aligns with a study arguing that AD is inherently interpretive but must be calibrated to the narrative and the user's needs (Starr & Braun, 2024). Participants' accounts expand on this by showing how genre acts as a key structuring factor. Preferences were not universal but shifted depending on pacing, auditory density, and the emotional structure of the media. For example, participants noted that too many descriptions could undermine the experience in dialogue-light or music-centered formats, while insufficient descriptions left gaps during visually complex action sequences.

These findings reinforce the idea that AD cannot be governed by a one-size-fits-all rule set but must adapt its expressive quality, density, and timing to the demands of the genre and scene. The artistry-precision dilemma is therefore less a contradiction than an indication that AD must remain sensitive to narrative context.

6. Limitations

This study offers an in-depth account of BLV viewers' experiences with AD, but several limitations shape how the findings should be interpreted. First, although the participants represented a range of backgrounds and viewing practices, the sample size was relatively small, as is typical for qualitative research. As a result, the findings highlight depth over breadth and should not be generalized to all BLV viewers without caution.

Second, the study relied on retrospective accounts gathered through interviews rather than direct observation of real-time viewing. Participants described what they noticed, valued, or struggled with, but their moment-to-moment decision-making during a film may differ from what they recall afterward. Future work incorporating observational or think-aloud methods may capture interactional nuances that interviews alone cannot reveal.

Third, participants primarily discussed media such as long-form narrative films and streaming services. Although short-form platforms like TikTok and YouTube emerged repeatedly as "accessibility deserts," the study did not systematically analyze those environments. As a result, the findings speak most directly to traditional AD contexts rather than the rapidly evolving landscape of short-form content.

Finally, due to AI AD tools advancing quickly, participants' impressions reflect a specific moment in technological development. As multimodal models improve and platforms refine their integration of AI AD, user expectations and concerns may shift. The findings, therefore, offer a timely, but not static, account of current experiences.

7. Future Work and Design Implications

Future research should build on these findings by examining how AD can support both narrative comprehension and social coordination across a broader range of media environments. One direction involves developing adaptive AD systems that automatically adjust density, timing, and descriptive style based on genre, recognizing that musicals, action films, and dialogue-driven narratives require different descriptive strategies.

Another promising direction is the design of individual or personal AD streams that allow BLV viewers to control how AD is delivered within shared viewing spaces. Participants repeatedly expressed concern about distracting viewers or feeling hesitant to request changes during group watching. Prototyping home-based systems that deliver AD through personal audio channels such as Bluetooth earpieces, smart TV multi-output audio, phone-based companions, or spatial audio modes would allow viewers to access AD without altering the group experience. Physical prototypes could examine how users switch between shared audio and individualized AD, how timing is synchronized across devices, and how such designs affect social comfort and participation.

There is also a need for research on narrative-aware AI models capable of tracking characters, understanding gestures, distinguishing relevant from irrelevant visuals, and recognizing scene transitions such as flashbacks. Participants stressed that current AI tools misunderstand continuity or introduce subjective interpretations, suggesting that next-generation models require training on narrative coherence rather than surface-level object detection.

Finally, future work should examine short-form media ecosystems, where accessibility gaps remain significant and AI-generated tools act as the primary pathway to visual information. Studying how BLV viewers navigate platforms such as TikTok and YouTube, and prototyping lightweight or AI-assisted overlays tailored to rapid content cycles would address an urgent need in new media accessibility.

8. Conclusion

This paper indicates the tacit demands of BLV users in AD across traditional and emerging media environments, including the rapidly growing presence of AI AD. Through semi-structured in-depth interviews with BLV participants, the study provided:

- A refined understanding of information prioritization,

- Recognition of AD as social infrastructure
- A grounded account of AI AD in practice
- A genre-sensitive view of the artistry-precision balance

All together, the findings show that BLV viewers' expectations for AD extend well beyond existing guidelines or current multimodal AI systems designed to support them.

Looking forward, we advocate for human, AI-assisted, or fully automated AD to adopt adaptive, context-aware, and socially sensitive approaches that reflect the diverse information needs of BLV viewers. As multimodal AI continues to advance and platforms experiment with automation, this work provides a timely reminder that access is not only a technical problem but a narrative and relational one. AD must support how people comprehend stories, how they move through media environments, and how they participate socially with others.

9. References

Andreassend, M. (2023). A history of audio description. AI-Media. https://www.ai-media.tv/knowledge-hub/insights/history-audio-description/

Bergin, D., & Oppegaard, B. (2024). Automating Media Accessibility: An Approach for Analyzing Audio Description Across Generative Artificial Intelligence Algorithms. Technical Communication Quarterly, 1–16.

Bittner, H. (2012). Audio description guidelines: A comparison. New Perspectives in Translation, 20, 41–61.

Bodi, A., Fazli, P., Ihorn, S., Siu, Y.-T., Scott, A. T., Narins, L., Kant, Y., Das, A., & Yoon, I. (2021). Automated video description for blind and low vision users. In Proceedings of the ACM SIGCHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI) (pp. 1–7).

Cheema, M., Seifi, H., & Fazli, P. (2024). Describe Now: User-Driven Audio Description for Blind and Low Vision Individuals. arXiv preprint arXiv:2411.11835.

Cheema, M., Seifi, H., & Fazli, P. (2025). DescribePro: An AI-assisted audio description authoring tool for video. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI ’25) (pp. 1–?). ACM. https://doi.org/10.1145/3663547.3746320

Chmiel, A., & Mazur, I. (2021). A homogenous or heterogeneous audience? Audio description preferences of persons with congenital blindness, non-congenital blindness, and low vision. Perspectives, 30, 1–16.

Chu, P., Wang, J., & Abrantes, A. (2024). LLM-AD: Large Language Model-based Audio Description System. arXiv preprint arXiv:2405.00983.

Chuang, C.-Y., & Fazli, P. (2023). CLearViD: Curriculum learning for video description. arXiv. https://arxiv.org/abs/2311.04480

Described and Captioned Media Program. (2022). Description Key - Quality Description. https://dcmp.org/learn/621-description-key---quality-description

Gao, Y., Fischer, L., Lintner, A., & Ebling, S. (2025). Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 471–490). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.29

Gonzalez, R., Collins, J., Azenkot, S., & Bennett, C. (2025). Investigating use cases of AI-powered scene description applications for blind and low vision people. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24). Association for Computing Machinery. Article 3642211, 21 pages. https://doi.org/10.1145/3613904.3642211

Ihorn, S., Siu, Y.-T., Bodi, A., Narins, L. D., Castanon, J. M., Kant, Y., Das, A., & Yoon, I. (2021). NarrationBot and InfoBot: A hybrid system for automated video description. arXiv. https://arxiv.org/abs/2111.03994

Jiang, L., Jung, C., Phutane, M., Stangl, A., & Azenkot, S. (2024, May). "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 (pp. 1–20).

Liu, X., Carrington, P., Chen, X. A., & Pavel, A. (2021). What makes videos accessible to people who are blind or visually impaired? In Proceedings of the Conference on Human Factors in Computing Systems (CHI) (pp. 1–14).

Natalie, R., Chang, R.-C., Sheshadri, S., Guo, A., & Hara, K. (2024). Audio Description Customization. In Proceedings of the 26th ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24) (pp. …). ACM. https://doi.org/10.1145/3663548.3675617

Netflix. (2021). Audio Description Style Guide v2.5. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-5

Media Access Canada. (2012). Media Access Canada (MAC) - Our Projects - Descriptive Video Production and Presentation Best Practices Guide for Digital Environments.
http://www.mediac.ca/DVBPGDE_V2_28Feb2012.asp

Ofcom. (2022). Ofcom's Guidelines on the Provision of Television Access Services. ([n. d.]).

Pavel, A., Chen, X. A., Carrington, P., & Liu, X. (2020).
Rescribe: Authoring and automatically generating audio descriptions.
In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)
(pp. 1–13). ACM. https://doi.org/10.1145/3379337.3415864

Starr, K., & Braun, S. (2024). Omissions and Inferential Meaning-Making in Audio Description, and Implications for Automating Video Content Description. Universal Access in the Information Society, 23(?), 1287–1302. https://doi.org/10.1007/s10209-023-01045-3

Xie, J., Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., & Zisserman, A. (2024). AutoAD-Zero: A training-free framework for zero-shot audio description. In Proceedings of the Asian Conference on Computer Vision (ACCV 2024) (pp. 2265–2281). Springer. https://doi.org/10.1007/978-981-96-0908-6

Appendix A.

Audio Description Guideline Summary Table

G1	When describing actions, determine what is most relevant for the story to flow without negatively impacting the viewers' experience.
G2	When creating Audio Description for a language other than the original language, determine how known/unknown the setting is outside the original audience and describe accordingly (naming vs explicitation. It is best to name and explicate if time allows, e.g., Tower Bridge - a turreted bridge over the river Thames; He wears a barretina – a red Catalan hat.
G3	The vocabulary should reflect the predominant language/accent of the program (for example, American English vs. British English; Castilian Spanish vs. Mexican Spanish, etc.) and should be consistent with the genre and tone of the content while also mindful of the target audience.
G4	Avoid censorship: do not censor any information. Description should be straightforward when addressing nudity, sexual acts, and violence. Language choice should reflect the target audience and rating (be guided by the program content).
G5	The description should include any opening and closing credits with an adjusted tone when not too distracting, but if these interfere with simultaneous dialogue and action, timing adjustments may be made, such as grouping, to introduce the text before or after the actual credit appearance. Credits will be included as time permits. They may be condensed if time is limited. Prioritize credits in order of appearance.
G6	Aim to have the following credits described during opening and/or closing credits: Creator, Writer, Director, Main Cast, Producer, Executive Producer, Director of Photography, Editor, Music & Sound by If unable to cover all credits and if time allows, state that edits have been made with a line such as "other credits follow."
G7	When creating Audio Description for a language other than the original language (i.e., AD that is mixed with a dub), only cover the above credits if they are in the same language as the AD you are creating. Otherwise, state that credits appear in the relevant language with a line such as "opening credits appear in Japanese."
G8	While jargon and technical terms should not be used, film terminology that has entered the common vocabulary can be used when necessary for timing and clarification or when in line with the story and/or genre (for example, "now in close-up").

G9	There must be no errors in word selection, pronunciation, diction, or enunciation.
G10	Describe discernible attributes and expressive gestures, but do not describe what is inferred by them, such as emotional states.
G11	If time permits, describe the series of still images, such as those often used during a documentary interview, to highlight certain subjects being discussed by the person or people being interviewed.
G12	When describing certain passages of time, such as flashbacks or dream sequences, for younger audiences, it is sometimes impractical to use describing conventions that one might use for adults. In some cases, it is necessary to explicitly tell the audience what is happening rather than describing the action.