

Making Short-Form Videos Accessible with Hierarchical Video Summaries

Tess Van Daele

Department of Computer Science
The University of Texas at Austin
tessvandaele@utexas.edu

Jalyn C Derry

Department of Computer Science
The University of Texas at Austin
jalyn.derry@utexas.edu

Akhil Iyer

Department of Computer Science
The University of Texas at Austin
akhil.iyer@utexas.edu

Yuning Zhang

Department of Information Science
Cornell University
yz3227@cornell.edu

Amy Pavel

Department of Computer Science
The University of Texas at Austin
apavel@cs.utexas.edu

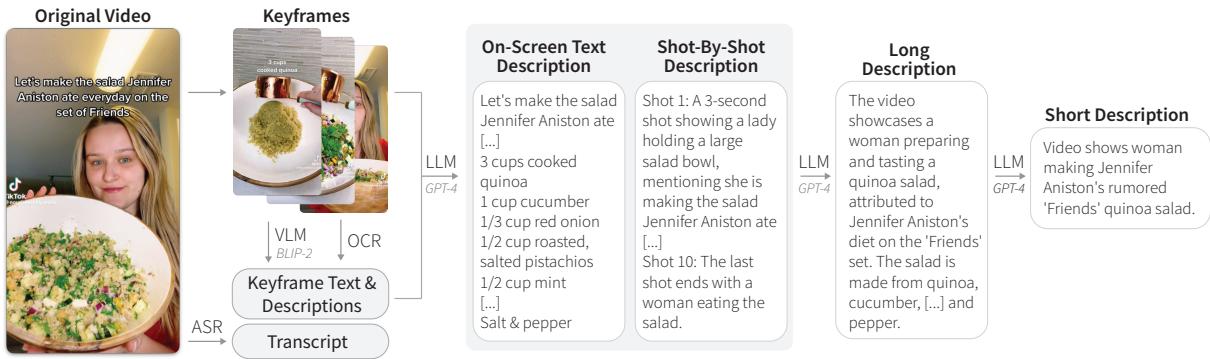


Figure 1: ShortScribe makes short-form videos accessible with hierarchical video descriptions. ShortScribe extracts video data by identifying key frames then applying automatic speech recognition (ASR), automated description (BLIP-2), and optical character recognition (OCR). A large language model (GPT-4) then generates multiple descriptions. TikTok by @nourished.by.mads [54].

ABSTRACT

Short videos on platforms such as TikTok, Instagram Reels, and YouTube Shorts (i.e. short-form videos) have become a primary source of information and entertainment. Many short-form videos are inaccessible to blind and low vision (BLV) viewers due to their rapid visual changes, on-screen text, and music or meme-audio overlays. In our formative study, 7 BLV viewers who regularly watched short-form videos reported frequently skipping such inaccessible content. We present ShortScribe, a system that provides hierarchical visual summaries of short-form videos at three levels of detail to support BLV viewers in selecting and understanding short-form videos. ShortScribe allows BLV users to navigate between video descriptions based on their level of interest. To evaluate ShortScribe, we assessed description accuracy and conducted a user study with 10 BLV participants comparing ShortScribe to a baseline interface. When using ShortScribe, participants reported higher comprehension and provided more accurate summaries of video content.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

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

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0330-0/24/05.

<https://doi.org/10.1145/3613904.3642839>

CCS CONCEPTS

- Human-centered computing;
- Accessibility systems and tools;

KEYWORDS

Short-Form Video, Accessibility, Video Description, Summaries

ACM Reference Format:

Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making Short-Form Videos Accessible with Hierarchical Video Summaries. 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, 17 pages. <https://doi.org/10.1145/3613904.3642839>

1 INTRODUCTION

Today, more than 1 billion users actively watch short-form videos across platforms such as TikTok, Instagram Reels, and YouTube Shorts [8]. Short-form videos usually range from 30 to 60 seconds in length and are presented in an algorithmically curated stream. To capture viewer attention, creators make densely packed videos featuring rapid scene changes, on-screen text overlays, and fast-paced action. Creators also engage with trends by reusing popular music and meme audio, and by stitching or overlaying their responses to other videos. While short-form videos are now a dominant source

of information, entertainment, and cultural references, they remain inaccessible to millions of blind and low vision (BLV) viewers.

Prior work has explored making videos accessible with manual [36], semi-automated [15, 46, 60, 76] and automated [17, 73] *audio descriptions* – or, narrations of visual content in the video. The duration and density of short-form videos make it challenging to fit audio descriptions within gaps in the audio. For videos without audio gaps, accessibility guidelines recommend adding extended descriptions that pause the video to narrate important visual content [16]. While such extended descriptions can be useful for educational videos [61], they lengthen the content and abruptly interrupt the audio [60]. Gleason et al. explored how to make silent GIFs accessible by manually creating three types of descriptions (alt text, source audio, and source audio with audio descriptions), finding that audience members preferred alt text descriptions as they were efficient to read with their screen reader [25]. Prior research and video accessibility guidelines have yet to explore how to make short-form videos accessible to blind and low vision viewers.

To understand the current practice of BLV viewers, we conducted formative interviews and a co-watching exercise with 7 BLV participants who regularly watched short-form videos. Participants reported frequently encountering videos with audio unrelated to the visual content of the video (e.g., trending songs, meme audio), making it difficult to determine what was happening on screen. Participants also reported that they did not know whether the video would be accessible ahead of time, thus having to watch the majority of a video just to find out it was inaccessible. When encountering inaccessible videos, participants either skipped the video, asked a friend for more information, or posted the video on online communities to request descriptions if they were particularly interested in the content. Still, participants ended up skipping a majority of inaccessible videos and rarely followed up for more information, limiting their access to short-form content.

To make short-form videos accessible, we present ShortScribe, a system that provides BLV short-form video viewers with hierarchical video summaries, or video descriptions at multiple levels of detail that viewers can access depending on their interest. To create these descriptions, our system first segments videos into shots (*i.e.* camera or scene changes) and extracts visual information from each shot using vision language models (BLIP-2, OCR). Next, it uses a large language model (GPT-4) to summarize the extracted visual information, creating descriptions for each shot (shot-by-shot descriptions) and for the entire video (short description, long description, and on-screen text). Using the short description, participants can quickly determine if the content of a video is interesting to them, and then flexibly explore additional details through the long description, shot-by-shot descriptions, and on-screen text.

We evaluated ShortScribe in a within-subjects study with 10 BLV participants who compared ShortScribe to a baseline interface created to simulate a typical short-form video platform. Participants demonstrated improved video comprehension and reported that they found unique uses for each of the descriptions we provided. Participants expressed unanimously that they would use ShortScribe in the future and that it improved their experience watching short-form videos.

Our work contributes:

- A formative study revealing current practices and challenges of watching short-form videos for BLV users
- Design and development of ShortScribe, a system that provides BLV users with hierarchical visual descriptions
- A user study demonstrating that ShortScribe improved the experience of watching short-form videos as well as selecting which videos participants wanted to watch.

2 BACKGROUND & RELATED WORK

Our work on making short-form videos accessible with hierarchical video descriptions relates to prior work in video accessibility, hierarchical summarization, and social media accessibility.

2.1 Video Accessibility

While a long history of work has explored how to make videos accessible for BLV viewers, prior research has not yet explored how to make short-form videos accessible [53]. We review primary approaches for making videos accessible:

2.1.1 Audio Description. Videos are often inaccessible to blind viewers when the visual content cannot be understood from the audio alone [45]. Web Content Accessibility Guidelines (WCAG 2.0) recommend that to make videos accessible, authors can create a summary of the content or add audio descriptions that synchronously narrate the visual content while avoiding overlapping with important audio content [72]. To support authors creating audio descriptions, prior work has proposed manual [36, 40], collaborative [51], and (semi-)automated [15, 17, 24, 46, 52, 60, 73, 76] approaches to create descriptions for a range of videos including long-form traditional films and TV shows [17, 24], user-generated videos [36, 46, 51, 52, 60, 73], livestreams [38, 40], and 360-degree videos [18, 23, 37]. Short-form videos often include continual audio such that clear gaps do not exist for adding audio descriptions.

When audio descriptions do not fit within audio gaps, the Web Accessibility Initiative suggests providing *extended descriptions* that pause the underlying video so there is additional time to add descriptions [16]. While extended descriptions may be well-suited for educational content, they introduce delays and confusing interruptions that are unsuitable for short-form videos. Prior work has explored prompting authors to add descriptions during recording [62] or providing users control over the playback of extended descriptions [60, 61, 66]. Rescribe let users control whether they receive inline, extended, or hybrid descriptions ahead of time [60], whereas other systems provided users on-demand access to extended descriptions [61, 66] or answers to their visual questions [66].

All prior work on inline and extended audio descriptions focused on long-form videos and thus created *time-aligned* descriptions, or descriptions that played back according to the time in the video (e.g., describing the setting at the beginning of a scene). To accommodate short-form videos, we explore text descriptions of the video that are not time-aligned to provide an overview of the video as a whole and detail on-demand.

2.1.2 Text Descriptions and Summaries. Commercial and research tools have explored using text descriptions or summaries of a video support BLV and sighted audiences. Video platforms for long-form videos such as YouTube [4], Vimeo [3], and Netflix [1], display a

text “title” (e.g., *Backyard Squirrel Maze 1.0*) and text “description” (e.g., “*Squirrels were stealing my bird seed so I solved the problem with mechanical engineering :)*”) for each video [64]. Creators use the text title, text description, and video thumbnail to preview the video content, and prior work noted that BLV audience members use text titles to select what video to watch [45]. Short-form videos lack such text titles that may provide a useful summary to BLV audiences. While short-form video platform (e.g., TikTok, Instagram Reels) creators can use the “caption” field to add text descriptions, this usually contains extra context about the video (e.g., the products displayed, the backstory behind the video, or an opinion on the topic) rather than describing the video content which might benefit BLV users but be repetitive to sighted audiences. Discussions about adding text descriptions for short-form videos exist [42], but the practice is not widespread. We explore using automatically generated text descriptions to support BLV users gaining a quick overview of short-form videos.

Prior work has also explored text descriptions for GIFs (short, silent videos) and as a replacement for audio descriptions in long-form videos. Gleason et al. explored how to make GIFs accessible using three types of manually authored descriptions: alternative text description (most preferred by users), adding original source audio, and adding audio descriptions to original source audio [25]. Similar to images, GIFs are silent and often convey a brief concept or scene (e.g., “Oprah shrugs”). We explore short text descriptions for short-form videos, but as short-form videos convey more information than GIFs we use multi-modal summarization to take into account audio content and provide flexible access to additional descriptions. At the other end of the spectrum, for long-form videos, accessibility guides advocate for including *descriptive transcripts* that describe visuals in the video and transcribe the audio into captions in the same document to support people who are Deaf-blind and others [16, 35]. Prior work has explored authoring descriptions and captions [15, 46, 60], but little prior work has explored creating descriptive transcripts. Similar to a descriptive transcript, AVScript [34] supports BLV video editors with an “audio-visual script” that included transcribed speech and descriptions of visual scenes. Building on such detailed descriptions, we explore enabling users to optionally access shot-by-shot descriptions that summarize the audio and visual details in each part of the video.

2.1.3 Beyond Descriptions. Videos can also be made more accessible using alternative modalities such as haptics or sound. Prior work has demonstrated the benefit of using haptics to conveying spatial information (e.g., location of actors on screen and facial expressions) in movies [71] and 360-degree videos [37]. Foley, or the reproduction of everyday sound effects added to videos to enhance audio quality, can artificially add rich information to the audio track of videos [13]. Such approaches enable rich spatial and sonic experiences, but we explore text descriptions in this project as they are broadly usable with existing devices and screen readers.

2.2 Heirarchical Summaries and Descriptions

Prior work has used hierarchical summaries to support searching, browsing, and skimming of long text [70, 77], images [33, 41, 68], audio recordings [43, 44], and videos [58, 59]. Wikum and Context Trees both explore hierarchical summarization as an approach to

crowdsource summarizations of text [70] and help people skim discussions [77]. Li et al. use bottom up hierarchical summarization to enable listeners to explore long-form dialog effectively, providing an overview and the ability to navigate to points of interest [43, 44]. SceneSkim provides video summaries at three levels of detail (plot summary, script, and captions), helping film professionals in searching and browsing for specific moments within a film [58]. Video Digests provides chapter and section summaries for learners skimming and browsing lecture videos [59]. While this prior work provides hierarchical summarization for text [70, 77] and speech-based content [43, 44, 58, 59], we consider providing hierarchical summarization of *visual descriptions* of videos to facilitate gaining a high-level overview or low-level details about visual content.

While “a picture is worth 1000 words” and vision language models can produce long descriptions [9], most BLV users do not want to read a 1,000 word description for every image. Accessibility research has started to explore using hierarchical image descriptions as an approach to support users in selectively navigating image details. Prior work uses vision language models to provide similarities and differences between images [33] or focus on specific parts of an image [41]. As efficiency is important for short-form videos, we explore hierarchical summaries as an approach for users to flexibly explore video details.

2.3 Social Media Accessibility

The rise of social media (e.g., TikTok [7], Instagram [5], YouTube [4]) has brought a rise in inaccessible images and videos. Prior work has investigated ways to support BLV social media users as authors [14, 34, 40] and consumers [25–27, 40, 45, 74] of new types of images and videos used on social media. Researchers, practitioners, and advocates have iteratively released guidance on effective descriptions for images and video on social media [25, 26, 40, 42]. We build on such prior work by exploring current practices of BLV short-form video audiences and exploring text descriptions for making such videos accessible. Prior research has studied short-form videos to assess their impact on viewers [19, 50, 69] and recent research explores accessibility issues of short-form videos to uncover opportunities to support neurodiverse viewers [65] and viewers with physical disabilities [22]. We explore the accessibility of short-form videos to uncover opportunities to support BLV viewers.

3 FORMATIVE STUDY

Prior work has studied viewing practices, accessibility challenges, and solutions of BLV viewers for both long videos like TV shows [57], livestreams [40], and YouTube videos [45], and short videos like silent GIFs [25] commonly found on social media. We conducted a study to explore these topics for short-form videos, a new form of social media content not yet explored by prior work.

3.1 Method

The formative study included 7 BLV participants who accessed their mobile device using a screen reader and had experience watching short-form videos (Table 2). We recruited participants using social media and mailing lists. Participants were 20–57 years old and described their visual impairment as blind (5 participants) or some light perception (2 participants).

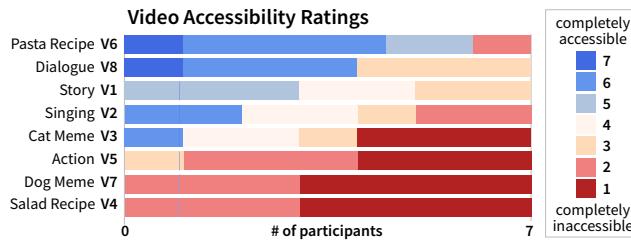


Figure 2: Participant ratings of video accessibility for pre-selected videos.

We asked participants a series of demographic and background questions including what types of short-form videos they watched, how they found those videos, the accessibility barriers that they encountered, and how they navigated those barriers. We then asked participants to watch 8 pre-selected short-form videos (Table 3). To ensure the 8 pre-selected videos covered a wide range of accessibility levels, we first created a 4-point scale to categorize the level of audio-visual match (as audio-visual mismatches indicate inaccessible video segments [46]): Unrelated Audio (e.g., a trending song), Somewhat Related Audio (e.g., a meme-style audio), Partially Informative Audio (e.g., recipe narration that covers some but not all details), and Mostly Informative Audio (e.g., a talking head). Videos were selected by scrolling through a newly created TikTok account, eliminating videos not in English or containing inappropriate content, until we found 4 videos per category. Finally, 2 videos were selected for each category such that the entire sample of videos covered a wide range of lengths and topics. Participants watched the pre-selected videos in a random order, and then provided accessibility ratings from 1 to 7 (completely inaccessible to completely accessible), and explained what factors impacted their rating, similar to Liu et al. [45]. Finally, we invited participants to think aloud as they watched videos on their own short-form video feed with their preferred platform for 10 minutes (2 participants used TikTok, 1 Instagram Reel, 3 YouTube Shorts, and 1 Twitter). We conducted this 1.5 hour long study via 1:1 remote Zoom interviews and compensated participants \$37.50 via Amazon Gift Card. Participants were asked to screen-share during the interview. This study was approved by our institution's IRB.

To collect data, one of the authors took notes during the interviews. Another author re-watched the entire set of zoom recordings, adding to the notes and constructing an affinity diagram [47]. The two authors later discussed themes that emerged in the affinity diagramming process to align on the results.

3.2 Findings

3.2.1 Current Practice. Participants watched short-form videos daily or weekly for entertainment (P1, P2, P4), staying up to date on popular trends (P4, P6), following creators (P3, P7), engaging with content shared by friends (P1, P2, P3, P4, P5, P7), and seeking information about specific interests (P1, P4, P5) like music, politics or education. P6 highlighted the importance of short-form videos being brief in length such that you quickly capture key information: “*when I first started, I watched longer videos. But I find now I move*

to shorter videos, videos that get to the point quickly.” All participants faced accessibility barriers when accessing short-form videos. When encountering these barriers, most participants skip the video (P1, P2, P3, P6, P7), but others seek help from close friends (P2, P4, P5) or social media groups that provide video descriptions for the BLV community (P2, P3). Participants also mentioned using the author-written video caption (P1, P6) or searching online (P6) to gain more information. P4, P5, and P7 expressed frustration when they failed to find more information to better understand the video, feeling that they might miss crucial content.

3.2.2 Short Form Video Accessibility. Overall, participants rated the accessibility of the pre-selected videos as a 3.21 ($\sigma = 1.94$) on a scale of 1 (completely inaccessible) to 7 (completely accessible) (Figure 2). Similar to Liu et al. [45], participants noted that videos with more speech such as stand-up comedy or podcast excerpts (P2, P4, P7), educational videos (P1), and singing or music related videos (Figure 2, V2) were more accessible than those with less speech. Short-form videos also presented unique accessibility challenges: **Repurposed Audio:** Participants reported that short-form videos that reused audio from other sources were challenging to understand. In some cases, the audio was somewhat related to the visual content, but still did not contain enough information for participants to understand the video. V3 and V7 both used audio memes. In V3, the video depicts a cat yawning, the audio features a woman saying “*Mm mm mm (disapproval), today drained me*”, the text on screen says “*My cat sleeping for 18 hours, stomping on my head at 4am..., and beating up his brother for no reason*” and the caption states “*He works so hard*”. In V7, the video features a dog cuddling chipmunk stuffed animals and the audio is a woman singing “*If you're not real, how come I feel this way? Little babies*.” With access to only the audio and author written captions, participants commonly misinterpreted the subject of these two videos. For instance, P4 interpreted V3 as a fatigued woman shaking her head, and V7 as a person with a baby. For a video depicting a recipe montage set to entirely unrelated music (V4), participants recognized the video as being inaccessible (Figure 2), and were not misled by the audio. As P1 shared: “*when it's just like music and images I just skip past that because that's not accessible*”. Due to the presence of repurposed audio on the platform, participants occasionally guessed audio was repurposed when it was not. For example, for a video with a woman singing a song to the camera (V2), P6 asked if the video was of a person lipsyncing to a song.

Micro Videos: Participants indicated that extremely short videos (5 seconds or less) often had uninformative or ambiguous audio that interfered with screen reader audio, making these videos inaccessible. Participants thus rated videos 5 seconds or less (V3, V5, V7) with accessibility scores of 3 or less (Figure 2). While V3 and V7 featured meme-style repurposed audio, V5 included original audio of a man saying “no!” then a splash. Participants were unsure what happened in the video (a dog jumped into a pool onto a man). All videos play in a continuous loop (e.g., V5 had a “no” then splash loop) and participants reported that the looping audio was overwhelming and overlapped with the audio of their screen reader such that they needed to pause the video to look for more information, disrupting their browsing.

Reaction Videos: P6 reported that reaction videos in which one creator stitched the video of another creator to react to it, were difficult to understand from audio alone. P6 shared that it would be challenging to add descriptions for such videos: “*Multi-layers of audio description would be needed, you know. What’s the original person doing? And how does that match up with you?*” (P6).

Complex Actions: P7 reported that short videos with excessive animations and movements were not accessible, as these videos are unlikely to be adequately described in limited time. For example, P5 said that “*Dance is always inaccessible*”, as the complex, fast-paced movements and changing facial expressions in dance make it difficult to fully describe. P6 explained that even a narration referencing primarily visual content, such as “*birds flying half a mile away*,” signifies inaccessibility. While the narration references the visual focus for sighted viewers, it does not provide sufficient descriptions of visual details for BLV viewers to also appreciate the visuals (e.g., movements of the birds, appearance of half a mile).

3.2.3 Platform Accessibility. Participants rated platform accessibility as 3.36 ($\sigma = 1.49$) on a scale of 1 to 7 (completely inaccessible or accessible, respectively) and highlighted key accessibility barriers:

Video Controls: All participants reported that play/pause video controls were challenging to use on TikTok as they required tapping the screen, a gesture that was not supported when using VoiceOver. Several participants also noted difficulties with other playback controls, such as skipping, fast-forwarding, and rewinding. Most short-form video platforms loop video playback by default, posing accessibility challenges. When a loop occurred, it was challenging for participants to tell from the audio alone if the application was still playing the video, had switched to the next video, or if the video had looped. Participants expressed a preference for proactive video control, rather than automatic video playback, and noted that video controls on YouTube Shorts were the most accessible.

Button labels: All participants reported that clutter and a lack of clarity in button labels led to the platform being inaccessible. As P6 described: “*The comment buttons and the share buttons, I don’t know which video even they are connected to. I may find a play button, but it’s not necessarily the one for the video that I’m trying [to play].*”

Platform Updates: Five participants noted that updates in the platform layout, particularly changes in the button positions, incurred a steep learning curve, leading to moments of inaccessibility.

3.2.4 Participant Suggested Accessibility Improvements. Participants suggested adding descriptions for short videos (6 participants) and providing access to the text on-screen (5 participants) to improve understanding of the short-form video content. Three participants suggested making it easier to access the author-written video caption and user comments, as these helped participants decide whether or not to watch a video. Two participants suggested that developing additional VoiceOver gestures for fast navigation.

3.3 Design Implications

Our formative study revealed design opportunities for making short-form video viewing and browsing accessible:

- D1.** Provide efficient access to on-screen visuals and text.
- D2.** Support users in recognizing audio and visual mismatches.
- D3.** Enable screen reader control of video playback and browsing.

- D4.** Support users in deciding on a video to watch.
- D5.** Maintain fast-paced video viewing and browsing.
- D6.** Provide access to complex visual content (e.g., dance).

We support BLV short-form video viewers with short, long and shot-by-shot descriptions that we crafted according to these design goals. First, to enable screen reader users to efficiently decide whether or not to watch a video (**D4**), we provided **short descriptions** that share a brief summary of video content (**D1**). Such short descriptions provide a similar function to long-form video titles, as prior work indicated that video titles support BLV users in selecting videos to watch [45]. While short descriptions can support users quickly selecting videos (**D4**) or noticing audio-visual mismatches (**D2**), short descriptions optimize for efficiency (**D5**) over completeness and thus may leave out important information. We enable BLV users to flexibly gain detailed access to audio-visual content and on-screen text (**D1**) with **long descriptions, shot-by-shot descriptions**, and a transcript of **on-screen text**. We prioritize displaying the short description first in the interface, and enable flexible description access to longer descriptions, such that users can decide whether or not they want to spend additional time to read longer descriptions (**D5**). Reading longer descriptions can support users in clarifying misunderstandings from mismatched audio, or even recognizing more subtle audio-visual mismatches such as an ingredient that a recipe narration left out (**D2**). We create our interface with explicit video playback control such that the interface is accessible and efficient to use for screen reader users (**D3, D5**). Our work does not address **D6**, but it offers a rich opportunity for future work.

4 SYSTEM

We present ShortScribe, a system that makes short-form videos accessible to BLV users by providing hierarchical descriptions of short-form videos.

4.1 Viewing Videos with ShortScribe

Mari is blind and recently started cooking more often. She likes to find ideas for new recipes by watching short-form videos on TikTok. While scrolling, Mari comes across a video set to a popular song (Figure 3, left). To find out what the video is about, Mari uses her screen reader to read the short description that states the video features a woman making a quinoa salad (Figure 3A). As Mari is looking for easy and filling lunch options, she checks out the long description to read a more detailed description that shares a high-level overview of the ingredients (e.g., “pistachios, mint, chickpeas”). She decides she wants to make the salad later, so she reads the on-screen text, which specifies detailed ingredient amounts (e.g., “3 cups cooked quinoa”). She copies the ingredients into her shopping list, and reads the shot-by-shot description to find how the creator prepares the salad (adds ingredients one by one to a large bowl). Mari continues to scroll to another video and reads the short description to notice this video is about skincare products (a topic she is not interested in), so she immediately scrolls to the next video in her feed.

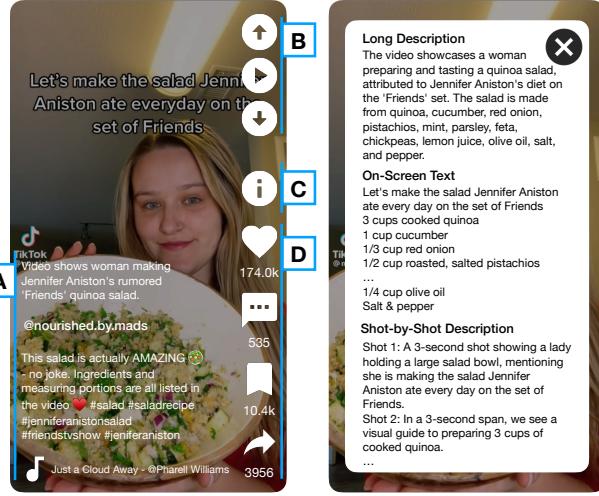


Figure 3: The ShortScribe interface consists of (a) front screen video information including the short description, username, caption, and audio title, (b) video controls, (c) a button to open the description pane which includes the long description, on-screen text, and shot-by-shot descriptions, and (d) video statistics. Video Credit: TikTok used with permission from @nourished.bymads [54].

4.2 Interface

We created a mobile interface to provide access to ShortScribe’s hierarchical video summaries (Figure 3). ShortScribe features a *video pane* that mimics existing short-form video platforms. We designed the video pane to be screen reader accessible by modeling the design off of YouTube Shorts as formative study participants highlighted it as the most accessible platform, and further modifying the design according to suggestions in the formative study (e.g., do not loop the video). We added ShortScribe’s short descriptions to this video pane. ShortScribe’s *description pane* is unique to our interface and lets users flexibly gain access to additional information via long descriptions, shot-by-shot descriptions, and on-screen text transcripts. Our interface serves to demonstrate how ShortScribe’s hierarchical video summaries may be integrated into existing short-form video platforms.

4.2.1 Video Pane. The video pane features the current video, video controls, and additional information about the video (e.g. short description, username, caption, number of likes). To switch between videos, users navigate to the previous or next buttons with their screen reader and double-tap the buttons. To play and pause a video, users double-tap the play button. The video pane displays information about the video in the following read order: short description, username, caption, source audio title, description pane button, number of likes, number of comments, number of bookmarks, and number of shares. The video pane directly displays the short description. The long descriptions, on-screen text, and shot-by-shot descriptions are included in the description pane.

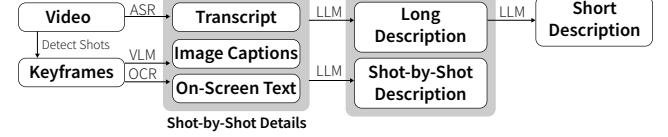


Figure 4: ShortScribe takes a video as input, transcribes the audio using automatic speech recognition (ASR), segments the video into shots, and selects the middle frame of each shot as a keyframe. It then processes the transcript, generated image captions (BLIP-2), and on-screen text (OCR) to produce video data for each keyframe. We use a large language model (GPT-4) to summarize this data into a short, long, and shot-by-shot description.

4.2.2 Description Pane. The description pane enables users to flexibly gain access to additional in-depth information about the video. In the description pane, the user can access the video’s long description, on-screen text, and shot-by-shot descriptions. The long video description provides a paragraph summarizing the video based on data such as on-screen text, audio transcription, and vision-to-language model-generated image captions. Users can also directly access on-screen text extracted from the video which is particularly useful for videos where creators included this type of text(e.g., recipes that list the ingredients on-screen). Finally, users can access the shot-by-shot descriptions which include descriptions of each shot in the video and provides more granular information on the video. To go back to the video pane after reading this information, users can double tap the close button.

4.2.3 Implementation. The interface (Figure 3) for ShortScribe was implemented using React.js (front-end) and real-time Firebase database (back-end) that logged user interactions for button clicks. All descriptions created from the pipeline were pre-loaded into JSON files and integrated into the React.js code base to filter out inappropriate words in the descriptions. We tested to ensure the interface was compatible with popular screen readers such as Apple VoiceOver and NVDA.

4.3 Pipeline

4.3.1 Extracting audio and visual information. Given a video, we first transcribe the audio in the video using Google Cloud’s Speech Transcription API /criteoglespeech. We then segment the video into shots, or continuous visual content segments using FFMPEG’s SceneDetect [2] to recognize sudden luminance changes. For each shot, we selected the middle frame as the representative keyframe. For example, if a shot is 30 frames long, we select the 15th frame as the keyframe. For each keyframe, we detected any on-screen text embedded in the video using Optical Character Recognition (OCR) from Google’s Video Intelligence API [6]. We filtered out any extracted text with confidence less than 0.95 to capture added on-screen text but avoid extraneous text (e.g., signs in the background of a video). We also filtered out content captured from the TikTok watermark (e.g., usernames and the word “TikTok”). We then used the BLIP-2 XXL image captioning model [39] to generate 5 candidate descriptions for each keyframe. The model generated the candidate descriptions with nucleus sampling, a required length of

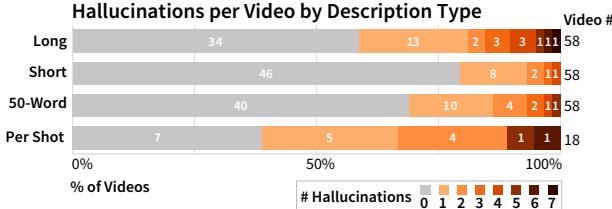


Figure 5: We analyzed hallucinations in descriptions for 58 videos (long, short, 50-word descriptions) and for a subsample of 18 videos (per shot descriptions). Descriptions for each video contained 0-7 hallucinations. Short descriptions had the lowest percentage of videos with hallucinations, while shot-by-shot descriptions had the highest percentage of videos with hallucinations.

5 to 20 words, a top-p value of 0.9, and a temperature of 1. To select a final description from the 5 generated candidate descriptions, we computed the CLIP score [12] between each candidate description and the shot keyframe, selecting the visual description with the highest score.

4.3.2 Generating Descriptions. For each shot, we gather the visual description of the shot’s keyframe, extracted on-screen text, and the corresponding audio transcript. To create the **shot-by-shot descriptions**, we prompt GPT-4 [10] to write a summary of each shot. We provide the audio and visual information for all of the shots at once to GPT-4 in order to provide context for each individual shot summary (full prompt in Appendix A.2). To generate the **long descriptions**, we similarly prompt GPT-4 to summarize the audio and visual information in all of the shots into a single summary paragraph (full prompt in Appendix A.1). If the long description exceeds 50 words, we condense the long description into a 50-word description and use the 50-word description as the long description for the video (full prompt in Appendix A.3). Finally, we create the **short descriptions** by prompting GPT-4 to condense the long description with the prompt: “*Condense the summary below such that the response adheres to a 10 word limit.*”

5 PIPELINE EVALUATION

We measured the coverage and accuracy of the short descriptions, long descriptions, shot-by-shot descriptions, and BLIP-2 generated image descriptions from ShortScribe.

5.1 Dataset

We first selected a set of 58 short-form videos. We selected the videos by creating a new TikTok account to remove any viewing history, then scrolling through the suggested videos to select videos spanning a variety of general audience genres: pets, memes, dance, music, and recipes. The selected short-form videos were 5 to 90 seconds in length and we collected 5 to 15 videos per genre. We selected a subset of 8 of these videos (Table 5) to represent a variety of accessibility levels (e.g., somewhat related audio, fully narrated) and wrote detailed summaries of the visual content for each video (i.e. human summaries). We used these summaries to evaluate coverage in the pipeline evaluation, and participant comprehension in

Description Type	Hallucinations			Coverage		Words	
	μ	σ	#	μ	σ	μ	σ
Short Description	0.33	0.78	19	75%	22%	10	1
50-Word Description	0.57	1.08	33	90%	20%	43	5
Long Description	0.97	1.63	56	100%	0%	136	37
Shot-by-Shot Description	1.44	2.01	26	100%	0%	190	186

Table 1: The mean (μ) and standard deviation (σ) of hallucinations and percent summary coverage for each description type. We also report the total number of hallucinations and summary points covered per description type. Hallucinations consider 58 videos (short, 50-word, and long descriptions) or a subset of 18 videos (shot-by-shot). Coverage considers the 8 pre-selected videos used in user study. Words considers all 58 videos.

the user study of the system. We ran our pipeline on all 58 videos, and included input data and pipeline results for all 58 videos in the Supplemental Material.

5.2 Analysis

We evaluated the accuracy and coverage of ShortScribe descriptions. For accuracy, we assessed short, 50-word, and long descriptions for all 58 videos, and shot-by-shot descriptions for a subset of 18 videos (8 used for the user study, plus 10 randomly sampled). For coverage, we assessed short, 50-word, long, and shot-by-shot descriptions for the 8 videos for which we had human-written summaries.

To evaluate the accuracy of descriptions, we examined the descriptions and videos to tally the number of inaccurate statements in each video (e.g., the description contained “guinea pig” when no evidence of a guinea pig was present in the visuals). We consider statements in the descriptions incorrect if they could not be justified by any visual content, audio content, or on-screen text content. For shot-by-shot descriptions, we consider statements within each shot incorrect if they could not be justified by any content within their corresponding timestamps. To evaluate coverage of the descriptions, we compared the descriptions to the human-written summaries by tallying the number of important details mentioned by both the description and summary. One researcher labeled all data reported in the pipeline evaluation. To verify inter-rate reliability of our accuracy and coverage codes, a second researcher also labelled the data and we computed weighted Cohen’s kappa for each description type. Agreement was “moderate” to “almost perfect” ($\kappa = 0.50 - 1.0$) for accuracy and coverage across all description types except shot-by-shot description accuracy that had “fair” agreement ($\kappa = 0.24$) [49]. In examining differences for shot-by-shot descriptions, researcher counts of inaccurate statements differed by 1.16 on average ($\sigma=1.38$). As shot-by-shot descriptions were much longer than other descriptions (averaging 190 words, Table 1) disagreements primarily occurred due to one researcher noticing an error the other missed and vice versa, rather than disagreement in the established code.

Video: A toddler and her mom sing a song together.

BLIP-2 Image Description

SHOT 1: “a child with an **expression of anger** on her face and mom”

On-Screen Text

“...maybe this thing was a masterpiece beauty piece til you tore it all up ... running scared I...”

Long Description

“The video ... primarily featuring a **melancholic narrative** ... portrays a toddler who **seems to be in emotional distress** ...”

Short Description

“Video presents **distressed toddler**, on-screen broken relationship text, mother **consoles**”

Figure 6: An analysis of the errors in one of the 2 of 58 videos that had more than three errors in the short description. The video depicts a lighthearted singalong. BLIP-2 mistakenly recognizes a toddler concentrating on singing as angry, and the on-screen text shows a quiz with the lyrics to a sad song (*All Too Well* by Taylor Swift). The long description and then short description incorrectly infer that the video is sad.

5.3 Accuracy

Overall, a majority of short, 50-word and long descriptions did not contain incorrect statements (Figure 5).

For long descriptions, 34 of the 58 videos contained no incorrect statements. Additionally, 13 of 58 long descriptions contained only one error, and 11 of 58 long descriptions contained more than one error (Figure 5). We investigated the long descriptions with more than three errors and found that the videos corresponding to these descriptions have no clear, informative audio or on-screen text content. For example, a video showcasing an unconventional apple peeler gadget without any useful description or on-screen text captions. The BLIP-2 model could not correctly identify what the gadget was doing based on the key frames. For instance, in one of the generated descriptions, BLIP-2 interpreted the gadget as a “*mint toothbrush holder*,” resulting in the long description mentioning an “*electric toothbrush device*.”

For short descriptions, there were no incorrect statements for 44 of the 58 videos. Additionally, only 3 videos had more than one incorrect statement (Table 1). The video with the greatest number of incorrect statements was a parent and their child performing a sing-along to a popular Taylor Swift song. BLIP-2 interpreted the child’s emotions as distressed due to their facial expressions. Additionally, the audio transcript captures the lyrics of the song about a broken romantic relationship. As a result, the long description and short description completely misunderstood the tone of the video as somber rather than humorous (Figure 6).

For the condensed 50-word descriptions, there were no incorrect statements for 40 of the 58 videos (Table 1). We found that the video with the greatest number of incorrect statements for this description type was also the child sing-along video due to the inaccurate shot descriptions and misleading audio. We also found that the shot-by-shot descriptions struggle to accurately describe videos with intricate visual details. For example, in a soup recipe video, the image captions incorrectly identify several ingredients, resulting in 6 incorrect statements throughout the shot-by-shot descriptions.

5.4 Coverage

Overall, long descriptions and shot-by-shot descriptions captured all of the important details that we identified for the eight videos (Table 1). The short and 50-word descriptions capture 75% and 90% of the important details respectively. ShortScribe descriptions are comparable to human video descriptions in terms of coverage, but ShortScribe’s are generally more verbose than human descriptions. For example, one video is a person in a car zooming the camera in on a street sign reading “*Drury Ln.*” and yelling out “*The Muffin Man*” as a reference to the children’s song. ShortScribe generated a long description that included the following segment: “*The video begins with a car driving through a scenic route surrounded by trees and hills, and the on-screen text introduces the location as DRURY LN. The audio features a voice expressing surprise amongst the ambient noise of the driving car. The second shot, lasting four seconds, focuses on a road sign clearly showing the name 'DRURY LN', reiterating the location. However, there's no accompanying text or audio in this particular shot.*” The description mentions the street name twice and includes an unnecessary final sentence.

6 USER EVALUATION

We conducted a user study with 10 BLV participants to examine how our descriptions impact video comprehension, selection, and preference compared to a baseline interface.

6.1 Method

In a within-subjects study, participants used both ShortScribe and a baseline interface to watch short-form videos. The study was 1.5 hours long, conducted in a 1:1 session via Zoom, and approved by our institution’s IRB. We compensated participants \$37.50 USD.

Participants. We recruited 10 BLV participants (8 female and 2 male) who use a screen reader to access their mobile device by using mailing lists and Facebook groups (Table 4). Participants ranged from 27 to 68 in age and described their visual impairment as blind (6 participants) or some light perception (4 participants). 9 out of 10 participants had experience watching short-form videos in the past although this was not a requirement for the study. P2, P3, P7, and P9 participated in the formative study.

Materials. The study consisted of two tasks: a video comprehension task and a video selection task. For the comprehension task, we selected 8 videos from our 58-video dataset (Section 5.1) to represent a range of accessibility levels and split the videos into two groups (VG1 and VG2) such that each group had videos with similar levels of accessibility (Table 5). Each group had a video with reused meme audio, a video with audio original to the creator, a recipe video with limited auditory description, and a talking head of a person listing out items within a theme. For the video selection task, we used the remaining 50 videos in the dataset and randomly divided them into two groups (VG3, VG4). For each task, participants viewed one video group with the ShortScribe interface and one video group with the baseline interface. We randomized and counterbalanced the videos and interface pairs within each task. We also randomized the order in which they viewed the interfaces.

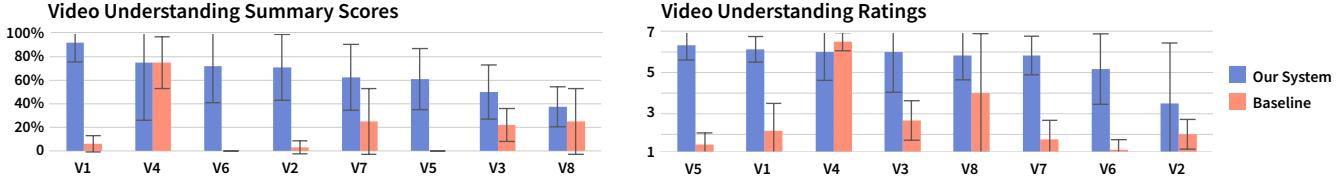


Figure 7: Video comprehension for videos V1-V8 using our system (left, blue) and a baseline interface (right, orange) measured by scoring participant written video summaries (Video Summary Scores) and participant’s ratings of their video understanding (Video Understanding Ratings). Ratings of the video understanding ranged from 1, did not understand, to 7, completely understood. Error bars depict the 95% confidence interval.

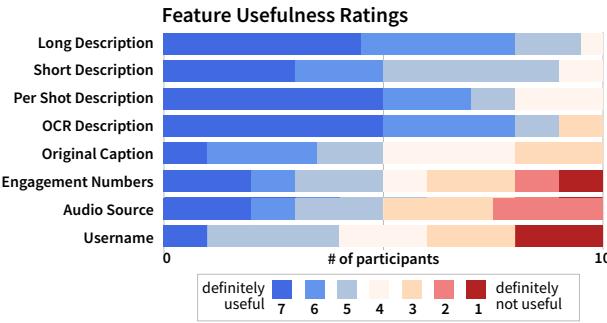


Figure 8: Participants rated the usefulness of each feature for understanding the video. The description features (first four features) are provided by ShortScribe only, and the remaining features (last four features) were originally available on the short-form video platform. Engagement numbers refers to the number of likes, comments, and bookmarks.

within each task. For both of the interfaces, we included information originally available on YouTube Shorts (e.g., author-written captions, author username, audio title, # likes, comments, and bookmarks) along with the videos. For the ShortScribe interface, we added the descriptions generated in the pipeline (short, long, and shot-by-shot) as well as on-screen text.

Procedure. We asked participants demographic and background questions about their experience with short-form videos (if any), and then provided a 10-minute tutorial on how to use both the baseline and our system’s interface. The rest of the study was split into two tasks:

Video Comprehension Task. Participants watched 4 pre-selected short-form videos with the baseline interface and 4 videos with the system interface. For each video, we allowed participants to investigate as much information as they wanted and watch the video as many times as they wanted. We then asked participants to provide a short summary of the video and share any questions they had about the video. We also asked participants to rank their understanding of the video on a scale from 1 (did not understand) to 7 (completely understood).

Video Selection Task. Participants were given 25 pre-selected short-form videos with the baseline interface and 25 videos with the system interface. Participants were asked to freely engage with

the sample feed, similar to how they would typically engage with short-form video platforms. We asked them to think-aloud as they scrolled through the feed and optionally watched the videos. This task lasted until participants reached the end of the 25 videos or 15 minutes passed.

At the end of the study, we conducted a final interview to understand participants’ experience with our system and asked participants to rate the usefulness of system features.

Analysis. We recorded and transcribed the interviews and recorded all interactions with both interfaces. To examine participant summaries, one researcher first prepared the summaries by randomly sorting them for each video. Another researcher, unaware of the participant author or interface condition of each summary, graded the participant summaries using the same human-generated summaries used to analyze coverage (Section 5.2). Another researcher read interview transcripts to derive themes.

6.2 Results

Overall, all participants reported that they would prefer to use ShortScribe to watch short-form videos over the baseline interface. Their average willingness to use ShortScribe in the future was 6.7 ($\sigma = 0.67$) on a scale from 1 (not likely to use in the future) to 7 (very likely to use in the future). Participants expressed that ShortScribe would broaden their access to a wider range of videos and improve the viewing experience: “*this makes me feel like I could view more videos*” (P3), “*it would give me a whole new avenue [...] I would even pay for it*” (P6). Participants also rated ShortScribe on a scale from 1 (not useful) to 7 (very useful) on how useful they found it for both understanding and selecting videos (Figure 8). For both of these questions, participants rated ShortScribe as significantly more useful compared to the baseline interface: video comprehension ($\mu = 6.5$, $\sigma = 0.71$ vs. $\mu = 2.4$, $\sigma = 0.97$; $Z = 2.78$, $p < 0.01$) and video selection ($\mu = 5.1$, $\sigma = 1.73$ vs. $\mu = 2.3$, $\sigma = 1.42$; $Z = 2.63$, $p < 0.01$). Significance was measured with the Wilcoxon Signed Rank test.

Improved Video Comprehension. The accuracy of participant-written summaries significantly improved when using ShortScribe compared to the baseline interface ($\mu = 73\%$, $\sigma = 26\%$ vs. $\mu = 20\%$, $\sigma = 30\%$; $Z = 4.61$, $p < 0.01$) (Figure 7, left). Participant’s self-reported video understanding also significantly improved when using ShortScribe compared to the baseline interface ($\mu = 5.89$, $\sigma = 1.53$ vs. $\mu = 2.53$, $\sigma = 1.93$; $Z = 4.99$, $p < 0.01$) (Figure 7, right).

When watching videos with the baseline, participants relied heavily on the audio and original video caption for information in the video. For example, participants watching V6 (a video of a cat with a SpongeBob audio asking a best pet friend to come back) with the baseline interface summarized the video using the topic of the audio (e.g., P4 summarized “*looking for a person or a pet who is lost*”). With ShortScribe, however, participants accurately summarized the content of the video despite the audio mismatch (e.g., P9 summarized “*there is a cat and he stole a straw...*”) (Figure 7, V6). When using the baseline, participants also tried to use the original video caption for information on the video but reported that “[captions] are just so variable” (P3) as “*sometimes [captions are] really useless, other times [they are] good.*” (P10). P3 explained how creators are often more concerned with creating an intriguing hook than a descriptive summary when writing the caption: “*usually they are not as helpful because people are trying to draw attention, and put in hashtags.*” As a result, the caption can also mislead participants. For example, the caption for video 1 mentioned funny animals even though no animals were present in the video, causing 4 out of the 5 participants who saw this video with the baseline interface to falsely include an animal in their summary. With ShortScribe, participants spent more time reading the descriptions and less time reading the caption.

Using ShortScribe’s Descriptions. All of the descriptions provided by ShortScribe (long, short, shot-by-shot and on-screen text) were ranked higher in terms of usefulness when compared to the baseline information (original caption, engagement numbers, audio source, username) (Figure 8). Participants reported that each type of description has a unique purpose.

The short description scored an average of 5.7 ($\sigma = 1.06$) on a 7-point scale on how useful they found it with 3 participants (P2, P6, P9) giving it a perfect 7 (Figure 8). Participants reported that the short description was helpful for gaining a concise overview of the video to assess whether they were interested in watching the video and/or exploring additional descriptions. P9 commented that it offered a “*brief glance about what it’s gonna be about,*” which she could use to determine if it is worth investigating the video further. P2 reported that she used it as a way to “*measure her interest*” and would use it to “*decide whether [she is] going to look into more detail.*” During the *Video Selection Task*, we observed how participants used the descriptions flexibly, only accessing the descriptions when they thought it was worthwhile. P9, for example, would often only read the short description so as to keep the browsing experience low effort. When watching a video of a haircut with audio from a popular song, she read the short description to find out the topic of the video and commented that “*I’m not really interested in hair videos so I’ll just skip this.*”

Participants rated the long descriptions as 6.2 ($\sigma = 0.89$) on the 7-point scale on how useful they found it (Figure 8). Participants who rated the long descriptions a 7 (P1, P2, P4, P6, and P8) reported that they liked how the long description “*seemed to cover more of what was done in the video*” than the short description (P8) and provided a good balance between detail and conciseness.

Participants reported that the on-screen text was particularly useful when videos contained important text information, such as ingredients in a recipe (V3 and V7) or a key joke (V1 and V5). When participants watched a video with important text using the baseline

interface, they were unaware of the information they were missing and responded with frustration once we told them what they had missed in the video. P10 found the on-screen text particularly useful and saw it as “*a personal touch*” from the creator of the video.

Participants rated the shot-by-shot descriptions a 6 ($\sigma = 1.25$) on the 7-point usefulness scale (Figure 8). Participants reported that these descriptions offered them the most detailed and sequential narrative of the video. P6 commented that she used shot-by-shot descriptions to “*get the sequence and the whole message*” of the video. Additionally, P4 mentioned that the shot descriptions were a “*broken down*” version of the video and made her realize that “*a lot of stuff was packed into that video that I wouldn’t have known.*”

Overall, participants found value in the descriptions provided by ShortScribe and reported benefits beyond improved video comprehension and selection. P3 shared that ShortScribe “*helps with the frustration of saving a video to show someone later*” as the video descriptions often answered questions she had about the video that she might typically save to ask a sighted person to answer later. P4 commented that even for videos considered more accessible, she could use the descriptions to quickly confirm that she was not missing any information. For example, when watching a video of a woman singing in task 2, she reported that the descriptions were helpful in confirming her assumption that there was a person singing and the audio was not from an outside source.

Unknown Errors. While using ShortScribe, participants encountered some errors in the descriptions, often without being aware of them. This happened when descriptions mentioned objects that were not present in the video but resembled objects that were. For example, during task 2, one video featured a kitchen gadget for cutting watermelon, but one of the shot-by-shot descriptions incorrectly identified the watermelon as meat. This led P4 to express surprise, saying “*I never would’ve guessed there would be meat at the end.*” P1 had a similar experience while watching Video 6 (Table 5). In this video, the creator addresses the camera directly, but one of the descriptions stated that she is “*interacting with TikTok on her phone.*” As a result, P1 and P2 reported that they thought there was a woman watching a TikTok on her phone rather than speaking to the camera.

Selecting Descriptions to Read. Participants appreciated the flexibility of choosing which descriptions to read, allowing them to establish a balance between the effort they are willing to invest and their desire for more information. P9 commented that “*if there’s too much information, its overwhelming*”. They further explain how the short description helped with this issue, saying that “*I felt like I didn’t have to go look in more detailed descriptions because [the short description] was so detailed*”. Some participants reported that they preferred more detailed descriptions even though they were longer and more time consuming. P6 ranked the most detailed description the highest and explained how it “*used details from the videos very well and didn’t go into too much detail*”. Participants would not only choose what description to read, but would also occasionally leave a description if they were not interested in it (e.g., noticed a shot-by-shot description became repetitive). We observed that some participants appreciated the process of thoroughly exploring the descriptions and found that reading all the supplemental information, regardless of how accessible the video is, contributed to their understanding of the content. Others were less inclined to

read the descriptions and used the flexibility of ShortScribe to only access descriptions based on their needs (e.g., reading descriptions until their questions were answered).

Interaction Improvements. Participants reported potential improvements for ShortScribe, one of which was too much repetition of the same content across the 4 descriptions. Participants also requested: access to the length of the video (P7), separating shot-by-shot descriptions into distinct text components so that they can be read separately by a screen reader (P7 and P10), reformatting the original caption hashtags to be compatible with screen reader (P10), and enabling auto-pause instead of auto-play while scrolling (P8 and P9).

7 DISCUSSION

Our formative study revealed that BLV viewers wanted access to visual information in short-form videos (**D1**, provide visual access). While BLV viewers occasionally use video captions, comments, asked friends or posted on meme pages to gain additional information, existing approaches do not address the accessibility gap. Author-written captions often do not describe visuals, and asking for external help conflicts with the fast pace of short-form videos (**D5**, maintain fast-paced viewing and browsing). Our hierarchical video summaries supported user study participants in gaining access to on screen visuals, improving self-reported video understanding and video summary scores in our *Video Comprehension Task* (**D1**). Our formative study revealed that audio-visual mismatches led to misconceptions about the visual content (**D2**, recognizing audio and visual mismatches). While participants used ShortScribe to clarify their understanding for audio-visual mismatches (**D2**), they also used ShortScribe to recognize mismatches between the author-written captions and the visuals. ShortScribe presents the short descriptions on the video pane (whereas users need to click to the description pane for access to additional descriptions) and thus the short descriptions supported participants in making quick decisions about whether to spend more time on the video or quickly move past it (**D5**). While participants in the user study *Video Selection Task* occasionally used the short description to decide whether to watch the video (**D4**, support users in deciding on a video to watch), they also used the short description to decide whether or not to access more visual information about the video. While the short video descriptions supported efficiency in decision-making and clarification tasks (**D5**), participants reported that longer descriptions (long, shot-by-shot, and on-screen text) were more useful to fully understanding the visual content (Figure 8) (**D1**). Based on the findings of our formative study, we enabled screen reader users to have explicit control of the video in our interface and user study participants were able to control videos with ease (**D3**, screen reader control of video playback and browsing). The high coverage (Table 1) and low number of errors for descriptions (Figure 5) make them immediately useful such that all users reported wanting to use the descriptions in the future. We reflect on our findings to discuss opportunities for future work:

7.1 Impact of Generative Model Performance

Descriptions generated by ShortScribe achieved high accuracy, but still contained errors (Figure 5). While these errors could lead to

misconceptions that BLV viewers may not be able to verify, our studies demonstrated that current short-form videos themselves often lead to misconceptions due to mismatched audio and video as well as misleading author-written captions. Short-form videos may represent a relatively low risk avenue for adopting generated descriptions. Our summarization-based approach that integrated information across multiple modalities (e.g., transcribed speech and visual descriptions) often removed errors that occurred in earlier stages of the pipeline (e.g., transcription errors). Occasionally, the model would amplify the errors (Figure 6). While participants did not directly express concerns about such errors in the user study, care should still be given to reduce errors in the future as large models can err in the direction of bias [29], and misunderstandings of video content have the potential to leave BLV audience members out of the loop (e.g., reacting incorrectly to a video sent by a friend due to a misinterpretation). In the future, we will explore how to reduce such errors by substituting the vision to language model in our pipeline (BLIP-2) with recently released vision to language models (e.g., GPT-V, Google Bard). We will also explore how to mitigate the impact of such errors by providing transparent access to detailed intermediate model outputs, similar to prior work [33, 41]. For example, future work could enable users to click on a description that seems potentially erroneous to reveal the original 5 BLIP-2 generated image descriptions and their CLIP scores. A high variation of the initial descriptions may prompt skepticism. Future work could alternatively use the CLIP score directly to reveal confidence [48].

ShortScribe lets users flexibly access descriptions of different levels of detail (e.g., by reading only the short description, or reading all of the descriptions). However, our one-size-fits-all pipeline does not allow customization of the description content. As a result, descriptions lack content that might be valuable and enjoyable to some users and trivial to others (e.g., detailed description of outfits). Future work may explore generating multiple different types of descriptions (e.g., fashion, background descriptions) that users could navigate between or toggle on and off. Alternatively, future work may explore letting users directly customize summarization prompts (e.g., “leave out any information about color”).

Finally, our pipeline produces description redundancies, both with the original audio of a video and between descriptions ShortScribe. For example, given that both the short and long descriptions provide a summary of the video, the long description often repeats information in the short description. Descriptions also occasionally described the audio of the video unnecessarily (e.g., the description segment “There is no accompanying audio”, or “A voice referencing a car game”). In the future, we could explore filtering out redundancies by further modifying the prompt or comparing the descriptions to the transcript. Our user study revealed, however, that some users prefer being given as much information as possible (e.g., detail about the audio can support quickly remembering context of visual details). Thus, future systems should support users to decide whether to allow repetition or not.

7.2 Extending Support for Short-Form Videos

Short-form videos showcase a wide degree of variability in length, content theme, and audio type. The variability impacts what

descriptions will be useful for what videos. ShortScribe enabled viewers to flexibly scroll on after reading the short description for short videos (e.g., a dog jumps into a pool) and flexibly gain additional access for informative videos (e.g., a recipe). While on-screen text was valuable for jokes and information, it was not useful when directly transcribing the audio. A benefit to our approach of large language model generated summaries, is that the summaries often suppressed redundant information such as a matching transcript and on-screen text making the descriptions work across a range of videos. Similarly, image-based vision to language models (e.g., BLIP-2, GPT-V) perform well across a range of visuals. Still, ShortScribe lack support for specific types of videos such as videos with complex actions (e.g., dance, trick shot challenges) or reaction videos (e.g., a side-by-side view of a video and a person reacting to the video, or a 2x2 view of four musicians playing together) due to limitations of our pipeline and limits of descriptions.

Improving our Generative AI Pipeline: We chose to the vision to language model BLIP-2 in our pipeline due to its high performance across diverse visuals. However, BLIP-2 takes single images as input and thus lacks information to recognize complex movements across time. In the future, we expect the performance of video to language models to improve, and we will also explore specialized pipelines such as accurately reconstructing 3D meshes of humans during dances such that the motions are robust to occlusion [28] then describing the 3D meshes [21, 75]. Vision to language models also struggle to recognize and accurately describe structured videos such as reaction videos. One approach may be to classify and segment reaction videos, create hierarchical descriptions for each individual video, then create a short description that summarizes all of the videos as a whole to provide an overview.

Alternative Description Modalities: Generating audio descriptions for complex actions in a limited time is a well-known challenge. Dance, for example, contains rapid and complex movements as well as facial expressions to convey emotional tone. Describing such movements in real-time without overwhelming the listener is difficult and can often take BLV audiences out of the experience of enjoying the performance [11]. Describing subjective themes such as emotion also violates audio description guidelines [55] which state that descriptions should be made “*without interpretation or personal comment*.” Some suggestions have been made to improve the accessibility of dance through haptic tours that allow BLV audiences to physically touch the set and costumes before the performance [20]. Haptics could also be used to augment descriptions with rich spatial information [71]. When the purpose of the dance video is a tutorial rather than a performance (e.g., a short video on how to do a popular dance move), future work may explore giving detailed auditory-only instructions and feedback as Rector et al. provided for yoga [63]. Sonification has also been used to improve accessibility of spatial navigation [56] and data visualization [30] by representing certain attributes (e.g., distance and data trends) as specific audio queues. However, haptics and sonification are time-based modalities (i.e. they play alongside the video as it plays) and may be challenging to integrate with short-form videos that contain rapid visual changes.

7.3 Platform Recommendations

The increased popularity of short-form videos supports the need for major platforms (i.e. TikTok, Instagram Reels, and YouTube Shorts) to leverage their access to advanced technology to improve accessibility. In the formative study, participants shared their difficulty in accessing both short-form video content and the platforms. We explored ShortScribe as an approach for making short-form videos accessible. Our work reveals the following design implications:

AI Descriptions to Fit Information Needs. While participants found short descriptions beneficial for scrolling (e.g., analogous to viewing the first few frames for sighted viewers), additional information is required for videos of interest. Hierarchical visual description summaries support flexible information access.

Prioritizing Useful Details. In our user study, participants rated the username as the least useful piece of information for understanding a short-form video, yet this is typically the first piece of information presented to them on short-form video platforms. For efficiency, the screen reader order should reflect what will best support BLV users in understanding the content (e.g., the caption for current interfaces).

Screen Reader Access. Platforms also have the ability to make small adjustments that can significantly improve BLV users’ experience watching short-form videos. For example, platforms may consider making on-screen text added using their platform screen reader accessible. Video playback and browsing should be improved by providing easy access to pause to avoid overlapping short videos or enabling “auto-pause” as an alternative to “auto-play.”

Short-Form Video “Alt-Text” Field. BLV viewers reported that author-written captions are often used to complement sighted-viewers viewing experience and extend view times rather than describe visual details. ShortScribe’s short descriptions were helpful to clear up misunderstandings. While some content creators use the caption field to add a visual description [42], this is not the norm. An “alt-text” field for authors to write visual descriptions (similar to Instagram and Twitter) could support authors in making their content accessible.

Short-Form Video Accessibility Guidelines Web Content Accessibility Guidelines (WCAG 2.0) recommend that to make videos accessible, authors describe the content of a video or provide synchronous audio descriptions to narrate visual content [72]. Accessibility guidelines, however, have not been established specifically for short-form videos. Our formative study established unique challenges that make short-form videos different from traditional videos, such that new guidelines would be beneficial.

Awareness. Currently, short-form video content creators are not encouraged to consider a BLV audience while creating their content. Creators describing the visual content during production or adding descriptive text could improve short-form video accessibility.

7.4 Beyond Short-Form Videos

Our work demonstrates the potential to use a mix of visual information extraction via vision language models and summarization via large language models to create useful visual descriptions at multiple levels of detail. Prior work explored a similar approach for extracting and summarizing image details [33], but our work

demonstrates the potential of this approach for temporal media. In the future, we can explore extending ShortScribe to long-form videos, livestream recordings, or 360-degree videos. The hierarchical summaries may also support other tasks. For instance, prior work revealed that BLV people want adaptive length descriptions for digital comics [32], shopping images [67], and urban scenes [31]. We hope our work catalyzes future research exploring accessible descriptions that fit diverse user contexts and preferences.

8 CONCLUSION

In this paper, we presented ShortScribe, a system created to make short-form videos accessible to blind and low vision (BLV) viewers. Supported by findings from the formative study, ShortScribe provides on-demand descriptions covering various levels of detail generated by leveraging vision language and large language models. We evaluated the effectiveness of ShortScribe through a technical evaluation of description accuracy and coverage, and a user study with 10 BLV participants. In the user study, participants reported higher comprehension and provided more accurate summaries of video content. Participants flexibly navigated between different descriptions, and found all descriptions to be useful for different purposes. All participants stated they wanted to use ShortScribe in the future. We aim to motivate future work to support people with disabilities in engaging with and creating this new growing form of social media.

REFERENCES

- [1] 1998. Netflix. <https://www.netflix.com>
- [2] 2000. FFMPEG SceneDetect. <https://ffmpeg.org/>
- [3] 2004. Vimeo. <https://vimeo.com>
- [4] 2005. YouTube. <https://www.youtube.com>
- [5] 2010. Instagram. <https://www.instagram.com>
- [6] 2016. Google VideoIntelligence API. <https://cloud.google.com/video-intelligence>
- [7] 2016. TikTok. <https://www.tiktok.com>
- [8] 2021. Thanks 1 Billion! | TikTok Newsroom. <https://newsroom.tiktok.com/en-us/1-billion-people-on-tiktok>. Accessed: 2023-08-25.
- [9] Google 2023. Bard - Chat Based AI Tool from Google. Google. <https://bard.google.com>
- [10] 2023. GPT-4 Technical Report. <https://arxiv.org/abs/2303.08774>
- [11] The New York Times 2023. Hear the Dance: Audio Description Comes of Age. The New York Times. <https://www.nytimes.com/2023/11/11/arts/dance/dance-and-audio-description.html?smid=nytcore-ios-share&referringSource=articleShare&fbclid=IwAR2VHe5olqM9AjsbiHAGwlfx77rzrQK84rYYVs6tXoREMXcEufYGVckdVI>
- [12] Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark Gretchen Krueger Ilya Sutskever Alec Radford, Jong Wook Kim. 2021. Learning Transferable Visual Models From Natural Language Supervision. <https://arxiv.org/abs/2301.12597>
- [13] Laura Almo. [n. d.]. Why is it called “Foley” anyway? <https://web.archive.org/web/20180613090128/http://cinemontage.org/2016/02/called-foley-anyway/>. Accessed: 2023-12-11.
- [14] Cynthia Bennett, Jane E. Martez Mott, Edward Cutrell, and Meredith Morris. 2018. How Teens with Visual Impairments Take, Edit, and Share Photos on Social Media. 1–12.
- [15] Carmen J Branje and Deborah I Fels. 2012. Livedescribe: can amateur describers create high-quality audio description? *Journal of Visual Impairment & Blindness* 106, 3 (2012), 154–165.
- [16] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* 290 (2008), 1–34.
- [17] Virginia P Campos, Tiago MU de Araújo, Guido L de Souza Filho, and Luiz MG Gonçalves. 2020. CineAD: a system for automated audio description script generation for the visually impaired. *Universal Access in the Information Society* 19, 1 (2020), 99–111.
- [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.
- [19] Francesco Chirossi, Luke Haliburton, Changkun Ou, Andreas Martin Butz, and Albrecht Schmidt. 2023. Short-Form Videos Degrade Our Capacity to Retain Intentions: Effect of Context Switching On Prospective Memory. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [20] Jess Curtis, Tiffany Taylor, and Georgina Kleege. [n. d.]. Describing dances: Increasing access for blind and visually impaired audiences. *Dancers Group* ([n. d.]). <https://dancersgroup.org/2019/03/describing-dances-increasing-access-for-blind-and-visually-impaired-audiences/>
- [21] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. 2022. PoseScript: 3D human poses from natural language. In *European Conference on Computer Vision*. Springer, 346–362.
- [22] Jared Duval, Ferran Altarriba Bertran, Siying Chen, Melissa Chu, Divya Subramanian, Austin Wang, Geoffrey Xiang, Sri Kurniawan, and Katherine Isbister. 2021. Chasing play on TikTok from populations with disabilities to inspire playful and inclusive technology design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [23] Anita Fidyka and Anna Matamala. 2018. Audio description in 360° videos: Results from focus groups in Barcelona and Kraków. *Translation Spaces* 7, 2 (2018), 285–303. <https://doi.org/10.1075/ts.18018.fid>
- [24] Langis Gagnon, Samuel Foucher, Maguelonne Heritier, Marc Lalonde, David Byrns, Claude Chapdelaine, James Turner, Suzanne Mathieu, Denis Laurendeau, Nath Tan Nguyen, et al. 2009. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society* 8, 3 (2009), 199–218.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. 2023. Humans in 4D: Reconstructing and Tracking Humans with Transformers. *arXiv preprint arXiv:2305.20091* (2023).
- [29] Lisa Anne Hendricks, Kaylee Burns, Kaito Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*. 771–787.
- [30] Leona M Holloway, Cagatay Goncu, Alon Ilsar, Matthew Butler, and Kim Marriott. 2022. Infosonics: Accessible Infographics for People Who Are Blind Using Sonification and Voice. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery.
- [31] Karst MP Hoogsteen, Sarit Szapiro, Gabriel Kreiman, and Eli Peli. 2022. Beyond the cane: describing urban scenes to blind people for mobility tasks. *ACM Transactions on Accessible Computing (TACCESS)* 15, 3 (2022), 1–29.
- [32] Mina Huh, YunJung Lee, Dasom Choi, Haesoo Kim, Uran Oh, and Juho Kim. 2022. Cocomix: Utilizing Comments to Improve Non-Visual Webtoon Accessibility. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [33] Mina Huh, Yi-Hao Peng, and Amy Pavel. 2023. GenAssist: Making Image Generation Accessible. *arXiv preprint arXiv:2307.07589* (2023).
- [34] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.
- [35] W3 Web Accessibility Initiative. [n. d.]. Transcripts. <https://www.w3.org/WAI/media/av/transcripts/>. Accessed: 2023-12-11.
- [36] The Smith-Kettlewell Eye Research Institute. 2022. YouDescribe. <https://youdescribe.org/>
- [37] 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 (ASETS '23)*.
- [38] Joonyoung Jun, Woosuk Seo, Jiheyeon Park, Subin Park, and Hyunggu Jung. 2021. Exploring the Experiences of Streamers with Visual Impairments. *Proc. ACM Hum.-Comput. Interact.* (2021).
- [39] Silvio Savarese Steven Hoi Junnan Li, Dongxu Li. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. <https://arxiv.org/abs/2301.12597>
- [40] 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.

- [41] Jaewook Lee, Jaylin Herskovitz, Yi-Hao Peng, and Anhong Guo. 2022. Image-Explorer: 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.
- [42] Veronica Lewis. 2023. *How To Write Video Descriptions For TikTok*. Veroniiici. <https://veroniiici.com/how-to-write-video-descriptions-for-tiktok/>
- [43] Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical summarization for longform spoken dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 582–597.
- [44] Daniel Li, Thomas Chen, Alec Zadikian, Albert Tung, and Lydia B Chilton. 2023. Improving Automatic Summarization for Browsing Longform Spoken Dialog. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [45] 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*. ACM, New York, NY, USA, 1–4.
- [46] Xingyu Liu, Ruolin Wang, Dingzeyu Li, Xiang'Anthony' Chen, and Amy Pavel. UIST 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding.
- [47] A. Lucero. 2015. Using affinity diagrams to evaluate interactive prototypes. In *Proceedings of INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14–18*. 231–248.
- [48] 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.
- [49] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 2 (2012), 276–282.
- [50] Ashlee Milton, Leah Ajmani, Michael Ann DeVito, and Stevie Chancellor. 2023. "I See Me Here": Mental Health Content, Community, and Algorithmic Curation on TikTok. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [51] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebriima H Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. The efficacy of collaborative authoring of video scene descriptions. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, New York, NY, USA, 1–15.
- [52] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting Novices Author Audio Descriptions via Automatic Feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [53] Alexandre Nevsky, Timothy Neate, Elena Simperl, and Radu-Daniel Vatavu. 2023. Accessibility Research in Digital Audiovisual Media: What Has Been Achieved and What Should Be Done Next?. In *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences (IMX '23)*. 94–114.
- [54] Maddy Caldwell (nourished.by.mads). 2023. Quinoa Salad. <https://www.tiktok.com/@nourished.by.mads>
- [55] American Council of the Blind. 2020. American Council of the Blind, Audio Description Project, Guidelines for Audio Describers. <https://www.acb.org/adp/guidelines.html>.
- [56] Marius Onofrei, Fabio Castellini, Graziano Pravadelli, Carlo Drioli, and Francesco Setti. 2023. *Video Sonification to Support Visually Impaired People: The ViSAViS Approach*. 503–514.
- [57] Jaclyn Packer and Corinne Kirchner. 1997. *Who's Watching?: A Profile of the Blind and Visually Impaired Audience for Television and Video*. Vol. 11. American Foundation for the Blind New York.
- [58] Amy Pavel, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. Sceneskin: Searching and browsing movies using synchronized captions, scripts and plot summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 181–190.
- [59] Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos.. In *UIST*, Vol. 10. Citeseer, 2642918–2647400.
- [60] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rscribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology (Virtual Event, USA) (UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. <https://doi.org/10.1145/3379337.3415864>
- [61] Yi-Hao Peng, Jeffrey P Bigham, and Amy Pavel. 2021. Slidecho: Flexible non-visual exploration of presentation videos. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–12.
- [62] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. 2021. Say It All: Feedback for Improving Non-Visual Presentation Accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [63] Kyle Rector, Cynthia L. Bennett, and Julie A. Kientz. 2013. Eyes-Free Yoga: An Exergame Using Depth Cameras for Blind & Low Vision Exercise. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '13)*. <https://doi.org/10.1145/2513383.2513392>
- [64] Mark Rober. 2020. *Backyard Squirrel Maze 1.0- Ninja Warrior Course*. Youtube. <https://www.youtube.com/watch?v=hZFjoX2cGg>
- [65] Ellen Simpson, Samantha Dalal, and Bryan Semaan. 2023. " Hey, Can You Add Captions?": The Critical Infrastructuring Practices of Neurodiverse People on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.
- [66] Abigale Stangl, Shasta Ihorn, Yue-Ting Siu, Aditya Bodhi, Mar Castanon, Lothar D 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 (ASSETS '23)*.
- [67] Abigail J Stangl, Esha Kothari, Suyog D Jain, Tom Yeh, Kristen Grauman, and Danna Gurari. 2018. Browsewithme: An online clothes shopping assistant for people with visual impairments. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 107–118.
- [68] Abigail Stangle, Nitin Verma, Kenneth Fleischmann, Meredith 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 (ASSETS '18).
- [69] Ziqi Tan, Shengyu Zhang, Nuoxin Hong, Kun Kuang, Yifan Yu, Jin Yu, Zhou Zhao, Hongxia Yang, Shiyuan Pan, Jingren Zhou, and Fei Wu. 2022. Uncovering Causal Effects of Online Short Videos on Consumer Behaviors. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1–10.
- [70] Vasilis Verroios and Michael Bernstein. 2014. Context trees: Crowdsourcing global understanding from local views. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 2. 210–219.
- [71] Lakshmie Narayan Viswanathan, Troy McDaniel, Sreekar Krishna, and Sethuraman Panchanathan. 2010. Haptics in audio described movies. In *2010 IEEE International Symposium on Haptic Audio Visual Environments and Games*. 1–2. <https://doi.org/10.1109/HAVE.2010.5623958>
- [72] W3C. 2022. *Audio Description (Prerecorded): Understanding SC 1.2.5*. WCAG. <https://www.w3.org/TR/UNDERSTANDING-WCAG20/media-equiv-audio-desc-only.html>
- [73] 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*. Association for Computing Machinery, New York, NY, USA, 1–12.
- [74] 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.
- [75] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. 2018. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters* 3, 4 (2018), 3441–3448.
- [76] 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*. Association for Computing Machinery, New York, NY, USA, 47–60.
- [77] Amy X Zhang, Lea Verou, and David Karger. 2017. Wikum: Bridging discussion forums and wikis using recursive summarization. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2082–2096.

A APPENDIX - GPT-4 PROMPTS

This section of appendix includes the prompts given to GPT-4 which are referenced in the System section of the paper.

A.1 GPT-4 Long Description Summarization Prompt

Your task is to generate a summary paragraph for an entire short-form video based on data extracted from the video. Your summary must be a holistic description of the full video.

The text in quotations defines the format of the data that I will provide you. The video data comprises of data extracted from all shots of the video.

The data is formatted in the structure defined in the quotations:
"SHOT NUMBER

Duration: the number of seconds that the shot lasts

Text on screen: Any text that appears in the shot

Shot audio transcript: Any speech that is in the shot

Shot description: A short visual description of what is happening
in the shot"

A.2 GPT-4 Shot-by-Shot Description Summarization Prompt

Your task is to generate a summary for each shot of a short-form video based on data extracted from the video.

The text in quotations defines the format of the data that I will provide you. The video data comprises of data extracted from all shots of the video.

The data is formatted in the structure defined in the quotations:

"SHOT NUMBER

Duration: the number of seconds that the shot lasts

Text on screen: Any text that appears in the shot

Shot audio transcript: Any speech that is in the shot

Shot description: A short visual description of what is happening
in the shot"

All of the summaries you create must satisfy the following constraints:

1. If the field for text on screen is empty, do not include references to text on screen in the summary.
2. If the field for shot audio transcript is empty, do not include references to shot audio transcript in the summary.
3. If the field for shot description is empty, do not include references to the shot description in the summary.
4. If the field for shot description is empty, do not include references to shot description in the summary.
5. Do not include references to Tiktok logos or Tiktok usernames in the summary.

Provide the summaries in a newline-separated format. There must be exactly one summary for every shot.

You must strictly follow the format inside the quotations.

"Your first summary

Your second summary

Your third summary

More of your summaries...

Your last summary"

A.3 GPT-4 50-word Description Prompt

"Condense the summary below such that the response adheres to a 50 word limit."

B APPENDIX - PARTICIPANT AND VIDEO DATA

The following appendix section contains participant and video data from the formative study and user evaluation.

ID	Age	Gender	Vision Impairment	Onset Age	Assistive Technology	Short-Form Video Platform(s)	Amount of Experience	Frequency	Preference for the type of videos
P1	29	Female	L: Totally Blind; R: Legally Blind	1	VoiceOver, Zoom text	YouTube, Instagram	1 year	4 times per day	Fashion, Informative
P2	33	Male	Totally Blind	11	VoiceOver, Jaws	TikTok, YouTube, Facebook	2 or 3 years	10 hrs per day	Informative
P3	32	Male	Registered Blind	Teen	VoiceOver, Jaws, NVDA	TikTok, Instagram, YouTube, Facebook	3 years	Daily	Informative
P4	29	Female	Legally Blind	12	VoiceOver, Jaws	YouTube, Facebook	< 1 year	1-3 times per week	Comedy
P5	43	Female	Totally Blind	1	VoiceOver	TikTok	2 years	15-20 mins per day	Shorter videos
P6	57	Male	Totally Blind	1.5	NVDA	TikTok	4 years	1-2 times per week	Informative, Live Music
P7	20	Male	Legally Blind	Birth	VoiceOver	TikTok, YouTube, Facebook, Instagram	4 years	1 time per week	Informative

Table 2: Background of Participants in Formative Study. L- left eye, R - right eye

VID	Length	Visual Content	Audio Content
V1	23s	Woman telling her small dog not to bark at larger dogs when at the park	Woman speaking to her dog
V2	1m 17s	Woman singing a song to the camera	Woman singing song
V3	5s	Cat yawning on a bed while laying down	Meme-style audio about how "today drained me."
V4	36s	Woman shows the steps to making a salad often made by a popular actress	Popular upbeat song
V5	5s	Dog jumping in pool even though its owner tries to keep him from doing so	Background noise of dog jumping in pool
V6	40s	Woman giving a detailed step-by-step recipe on how to make a pasta dish	Woman explaining how to do each step
V7	5s	Dog showing off its stuffed toys while laying on the floor	Meme style audio referencing "little babies"
V8	59s	A couple taking a quiz that had been trending	The wife asking the husband questions and him answering

Table 3: Content of 8 Pre-selected Videos from Formative Study

PID	Gender	Age	Visual Impairment	Onset	Preferred Short-From Video Platform (If any)
P1	Female	45	Legally blind	Congenital	YouTube shorts
P2	Female	43	Totally blind	Congenital	TikTok
P3	Female	29	Light perception	Acquired	Instagram Reels, YouTube Shorts
P4	Female	29	Light perception	Congenital	YouTube Shorts
P5	Female	33	Totally blind	Congenital	YouTube, Instagram, Facebook, Tiktok
P6	Female	63	Totally blind	Acquired	YouTube Shorts
P7	Male	32	Light perception	Acquired	YouTube Shorts
P8	Female	68	Light perception	Congenital	None
P9	Female	29	Legally blind	Acquired	Youtube Shorts, Facebook Shorts
P10	Male	27	Totally blind	Acquired	Youtube Shorts

Table 4: Background of Participants in User Study.

VID	Group	Length	Visual Content	Audio Content
V1	1	18s	A dad on a conference call falls in a pool while and continues the call as if nothing happened	Original audio from video
V2	1	10s	Zoom in of a street sign that reads "Drury Ln"	Man exclaims "Oh my gosh, you guys...The Muffin Man"
V3	1	23s	A man preparing vegan mozzarella	The video is introduced followed by a song overtop background noise of preparing the recipe
V4	1	54s	A man gives a list of 5 songs he believes should never be played on the acoustic guitar while playing them on his guitar	The video is introduced followed by the man playing portions all 5 of the songs on the guitar
V5	2	13s	A dog that escaped home rings the front door bell with his nose	Original audio from video
V6	2	8s	A cat is shown stealing a straw from a cup with the sentiment that the creator loves the cat regardless	Meme audio originally from SpongeBob about pets being your best friend
V7	2	33s	A tutorial-style video on how to prepare tato corn chow	The audio is a fantasy style rendition of a popular song
V8	2	25s	A woman gives a list of 5 basic Italian phrases that everyone should know	The video is introduced followed by the woman pronouncing all 5 of the phrases

Table 5: Content of 8 pre-selected videos from User Study.