OmniScribe: Authoring Immersive Audio Descriptions for 360° Videos

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

Chao-Hsien Ting
chaohsien1222@gmail.com
National Taiwan University
Taipei, Taiwan

Chia-Sheng Hung
yoyung0809@gmail.com
National Taiwan University
Taipei, Taiwan

Wan-Chen Lee
wanchen30@gmail.com
National Taiwan University
Taipei, Taiwan

Liang-Jin Chen
liangjin.lj.chen@gmail.com
National Taiwan University
Taipei, Taiwan

Yu-Tzu Chao
ytchao54@gmail.com
Audio Description Development Association
Taipei, Taiwan

Bing-Yu Chen
robin@ntu.edu.tw
National Taiwan University
Taipei, Taiwan

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

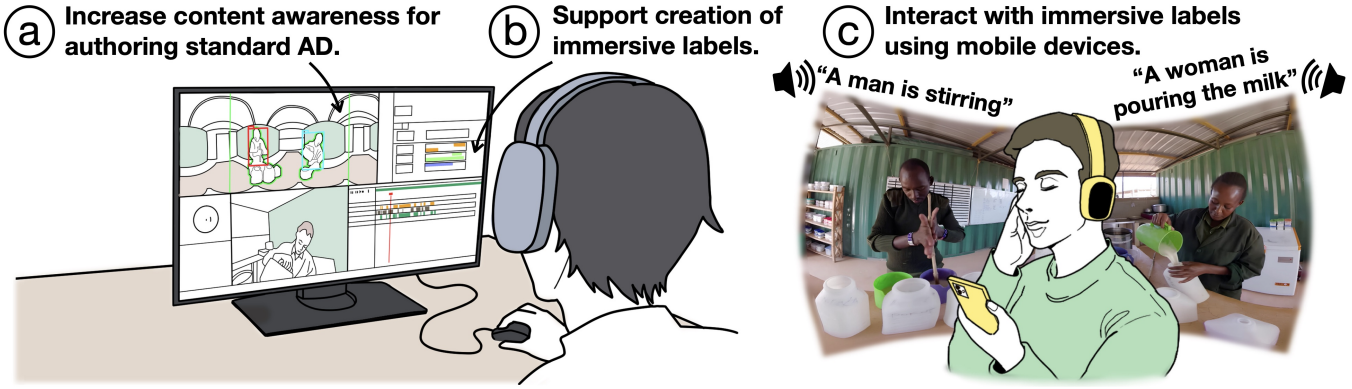

Figure 1: OmniScribe to make 360° videos accessible. A describer is using the OmniScribe web authoring interface, which supports them to (a) better understand the 360° content to author standard audio descriptions and (b) create immersive labels, in order to enable (c) BVI people to interact with 360° content immersively using smartphones and headphones. Video source (Reteti Elephant Sanctuary 360 VR): https://youtu.be/1ox36lZjG0A

ABSTRACT

Blind people typically access videos via audio descriptions (AD) crafted by sighted describers who comprehend, select, and describe crucial visual content in the videos. 360° video is an emerging storytelling medium that enables immersive experiences that people may not possibly reach in everyday life. However, the omnidirectional nature of 360° videos makes it challenging for describers to perceive the holistic visual content and interpret spatial information that is essential to create immersive ADs for blind people. Through a formative study with a professional describer, we identified key challenges in describing 360° videos and iteratively designed OmniScribe, a system that supports the authoring of immersive ADs for 360° videos. OmniScribe uses AI-generated content-awareness overlays for describers to better grasp 360° video content. Furthermore, OmniScribe enables describers to author spatial AD and immersive labels for blind users to consume the videos immersively with our mobile prototype. In a study with 11 professional and novice describers, we demonstrated the value of OmniScribe in the authoring workflow; and a study with 8 blind participants revealed the promise of immersive AD over standard AD for 360° videos. Finally, we discuss the implications of promoting 360° video accessibility.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility technologies**.

KEYWORDS

360° video, audio description, virtual reality, multimedia, accessibility, Blind, visual impairment, sonification, computer vision, mobile

ACM Reference Format:
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 *The 35th Annual ACM Symposium on User Interface Software and Technology (UIST '22), October 29-November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3526113.3545613

UIST '22, October 29-November 2, 2022, Bend, OR, USA

ACM ISBN 978-1-4503-9320-1/22/10…$15.00
https://doi.org/10.1145/3526113.3545613

1 INTRODUCTION

Blind or visually impaired (BVI) people typically access videos via audio descriptions (AD) as established by the Web Content Accessibility Guidelines (WCAG) [45] and the American Council of the Blind [36]. Describing videos is a time-consuming process, which requires a full understanding of the video, selecting crucial visual elements to describe, and several iterations to write the descriptions. 360° video is an emerging storytelling medium adopted on several media platforms (e.g., YouTube, New York Times, BBC News), enabling people to be immersed in scenarios they may not possibly reach in everyday life. Generally, people access 360° video by wearing a head-mounted display (HMD) or panning to rotate the field of view on the screen. However, either method takes viewers much time to find the focus of the video and absorb the content, which affects the pace of video understanding [23, 24]. Previously, researchers developed visual cues to increase viewers' awareness on out-of-sight content (e.g., picture-in-picture view [24] or arrows [23]), or automatically guide the viewer to the machine-recognized salient content in 360° videos [15, 23, 37, 49].

To make 360° videos accessible, the omnidirectional nature of 360° videos further introduces several challenges and opportunities for both video describers and BVI people. To describers, for instance, observing omnidirectional content and interpreting spatial information are challenging on HMDs, and they expressed a need to have a flat interface with a global view to better fit the workflow of describing videos [7]. However, it is challenging to understand extensive information, their direction and movement at a time in the distorted equirectangular view or limited normal field of view. To BVI people, it was noted that standard ADs made the experience of consuming 360° videos no different from that of 2D videos [6], which resulted in BVI viewers missing the immersive experience afforded by 360° videos otherwise available for sighted viewers. Previous studies highlighted the interactivity and control agency as crucial factors to distinguish the experience of 360° videos from 2D videos [6–8]. For instance, allowing BVI people to decide their own path to explore in 360° videos would be valuable in creating similarly immersive and interactive experiences sighted people have. However, the proposed methods are still under discussion, and it is unclear if they can promote accessible and immersive experiences for BVI people. Thus, in this work, we aim to understand *how to make 360° videos accessible and immersive to BVI people?* and *how an authoring tool can support video describers to better understand 360° video content to achieve that creation?*

To answer these questions, we first identified the challenges and needs of authoring ADs for 360° videos by interviewing a professional describer and synthesizing the insights from prior works [6–8]. We then concluded several design goals, including facilitating holistic 360° content understanding, providing conceivable information for mental construction, enabling the authoring of immersive labels, and setting low entry for the broader population.

Based on the design goals, we introduce OmniScribe, a system for authoring immersive ADs for 360° videos. OmniScribe consists of a web authoring interface for describers to create immersive ADs and a mobile prototype for BVI people to consume them. The web authoring interface aims to facilitate the understanding of holistic 360° content for describers. It is powered by the metadata retrieved from a video preprocessing pipeline, which enables shot/audio segmentation, content-awareness components (Figure 6) such as section division, saliency, and object tracking overlays, and the content map (Figure 3b). OmniScribe also enables describers to author spatial ADs and create immersive labels for BVI people. Then, our mobile prototype enables BVI users to access the immersive ADs generated by the web interface, including listening to spatial ADs, feeling vibrations to be aware of the scene transition and descriptions, and exploring objects by turning around (Figure 7).

We conducted a user evaluation with 8 novice and 3 professional describers to understand the effectiveness of OmniScribe in their authoring workflow, how the AI features helped with content awareness, and how the features are perceived from novice and professional describers' perspectives. Through another user evaluation with 8 BVI people, we demonstrated that OmniScribe-generated immersive ADs were preferred by BVI people, discovered insights such as the cognitive tradeoffs between standard and OmniScribe ADs, and further revealed implications for richer interactions that would be valuable for improving 360° video accessibility. OmniScribe represents an essential step towards making 360° videos accessible, and its technical approach may find applications broadly for increasing accessibility and immersion for 3D immersive multimedia.

2 RELATED WORK

Our work was inspired by the development and the lively discussion on promoting immersive media accessibility. We summarize the types of media and their accessible alternatives along the dimensions of *time* and *space* (Figure 2), and highlight the focus of OmniScribe.

2.1 Video Accessibility and Audio Descriptions

World Wide Web Consortium (W3C) has established Web Content Accessibility Guidelines (WCAG) as a reference for image creators to add proper captions as an accessible alternative [47] for BVI people to receive equal information as sighted people. Guidelines for describing videos [36, 45, 46] have also been made for content creators to add a description layer that makes video content non-visually accessible. However, describing videos is time-consuming in that it requires a full understanding of the video and takes several iterations to write and fit the description into the available time in-between dialogues [17, 38]. Further, videos with domain knowledge would require describers to collaborate with content creators or

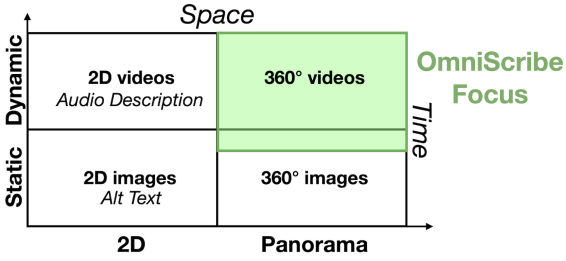

Figure 2: The scope of OmniScribe. OmniScribe mainly contributes to supporting AD authoring of 360° videos, and its techniques for interactive scene and object descriptions could potentially be used for 360° images as well.

domain experts to deliver the correct information for BVI people [8]. Such costs and obstacles impeded the participation of video creators to describe their videos, leaving the responsibility of describing videos to the third party as a post-hoc solution. However, the limited number of professional describers cannot keep up with the rapid pace of video creation and the huge volume of requests from BVI people, which further spurred the research interest in developing tools to support the authoring of audio descriptions for a broader population [16, 35, 50]. In our work, OmniScribe also takes this into account by setting low entry for allowing people with a wide range of expertise to create AD for 360° videos.

2.2 Tools to Support AD Authoring

One challenging part of description authoring is to fit the description into the time in-between dialogues (known as *inline ADs*), which requires describers to iteratively *(i)* locate the available gaps in the timeline and *(ii)* estimate the playback time of their description. To address this, YouDescribe [16], LiveDescribe [2], and Gagnon et al. [9] proposed timeline-based visualization techniques to help description writers better grasp the audio structure. YouDescribe [16] is a web platform nurtured by amateur volunteers to describe the videos requested by BVI people, while ViScene [35], also intended as a nonprofessional-sourced platform, provided additional expert feedback to improve description quality. In contrast to inline ADs, a video full of dialogues or crucial visual elements would require a video pause to ensure all key information is fully described (known as *extended ADs*). Rescribe [38], to combine the benefits of inline and extended ADs, introduced extended-inline descriptions that can automatically extend the background audio to accommodate out-of-bound descriptions to make them more acoustically-natural, and also be able to fit the descriptions to the limited time gaps by removing less-essential words. In contrast to prior works focusing on addressing problems in the *time* domain, OmniScribe focuses on processing *rich visual and spatial* information of 360° videos to facilitate information understanding and enable novel workflows for immersive description authoring.

2.3 Accessibility of Mixed-Reality Content

Immersive media is becoming increasingly popular and portable in our everyday lives that can create hyper-realistic and engaging experiences for users anywhere and anytime. However, different from well-established guidelines for images and videos, immersive media accessibility [48] is still under discussion and construction that necessitates input from the research community. Previously, researchers attempted to render immersive media BVI-accessible in different domains such as hardware accessories [3, 18, 32, 40, 53], applications [44, 54] and input/navigation techniques [43]. Among the hardware accessories, white canes [3, 18, 40, 53] were designed and crafted to allow BVI users to intuitively access and engage with virtual content through haptic and audio feedback. Other non-visual navigation techniques were also introduced to empower BVI people to explore the virtual world on their own, such as raycasting with joysticks [32] or acoustic maps for mental map construction [12, 27, 33, 43]. To make the visual content accessible, Zhao et al. [54] presented SeeingVR, a set of 14 tools to enhance visual awareness for low-vision users, and Herskovitz et al. [14] converged a task-oriented design space for making AR spatial content accessible. To democratize immersive media, the ImAc project [29, 30] presented a framework of collaboration between content creators and service providers on offering personalized and accessible experiences through ImAc players [31] in different usage scenarios (e.g., watching with different accessories). These attempts infused valuable contributions to the development of immersive media accessibility.

2.4 Accessibility of 360° Videos

Enabled by commercially-available devices, content creators can easily film 360° videos and use them as new storytelling techniques that offer immersive watching experiences. However, when filming or editing videos, accessibility is often neglected [25]. For 360° videos, it is still unclear whether and how the standard inline or extended ADs can be applied to describe the intricate omnidirectional visual information [23, 24, 37, 49]. To investigate this, Fidyka et al. [7, 8] conducted focus group studies with professional describers and BVI users to understand the challenges of describing and consuming 360° videos. They revealed several challenges for describing 360° videos, such as dividing 360° content into multiple sections, selecting and prioritizing content to describe, as well as the apparatus and workflow of describing 360° videos. At the same time, potential ways to consume AD were also explored, such as using spatial audio to signal the place of virtual elements [8] and other interactions (e.g., volume, head orientation) for accessing virtual content [7]. However, it is still unclear how specifically an interface should satisfy for describing 360° videos. In response, we focus on these questions in this work: *(i) how to make 360° videos accessible and immersive to BVI people?* and *(ii) how an authoring tool can support video describers to better understand 360° video content to achieve that creation?* Next, we detail our formative study that informed our design and development of OmniScribe.

3 FORMATIVE STUDY

To answer the above-mentioned questions, we conducted a series of interviews and conversations with a professional describer, who is also our co-author (referred to as CYT) and currently the head of the Association of Audio Description in Taiwan. CYT had fostered several professional describers and dedicated herself to promoting AD in the past 20 years. CYT had exceptional experiences in describing different visual media with over one hundred works, including movies, TV or stage shows, etc. Although she had not described 360° videos before, she had designed immersive audio instructions and haptic proxies for guiding BVI people in museums.

3.1 Method and Limitation

We collected data by interviewing a professional describer due to the scarcity of relevant research, especially the lack of perspectives on software necessities and development. Moreover, we co-designed closely with CYT due to her outstanding expertise and that most describers in Taiwan are closely cultivated and supervised by her, which limited our recruitment of describers with diverse backgrounds. Hence, we also considered and synthesized the insights from prior research [6–8] with our study as a whole in order to identify, supplement, and confirm the needs and key challenges.

3.2 Interview Procedure

The interviews were mainly structured as two sessions: First, we inquired about her experiences in AD, and then we introduced her to 360° videos, including the playback techniques on smartphones and websites, media formats (e.g., cube and equirectangular maps), popular 360° videos, and the types of region of interest (ROI) [24]. Prior to the first session, CYT was also informed of the prior research attempts [6–8]. Having basic understandings in mind, two weeks later in the second session, CYT shared her experiences in designing audio guidance in museums and instructions on cultural relics, which might be akin to the experience of exploring virtual 3D space [7]. Finally, CYT was asked to comment on the challenges and needs to describe 360° videos and the potential ways BVI people could consume them. We also had several offline conversations with CYT throughout the design process of OmniScribe. The dynamics between CYT and the other co-authors were similar to prior works on co-designing assistive technology [19] or community-based participatory design [4] to engage experts and stakeholders in the design process and recognize their contributions as co-authors.

3.3 Challenges to Make 360° Videos Accessible

We report the identified challenges of describing 360° videos and potentially-essential components to create an immersive experience for BVI people when consuming 360° videos.

3.3.1 Hard to perceive holistic 360° content for describers. The common remarked challenge of describing 360° videos was to understand the holistic content [7, 8]. Describers expressed a need to have a flat general view instead of using HMDs to look around, which makes multiple ROIs hard to be observed technically at a time [15, 23, 24, 37, 49], which was also confirmed by CYT: *"describer is supposed to be an omniscient who can know the whole story and details to better prioritize information to describe ... the 360° videos made me overwhelmed as I was not able to be fully aware of each object or event, and their temporal changes ... I needed to keep turning around to track something. It's extremely demanding."*

3.3.2 Unclear section division conceals the direction information. An equirectangular map encodes six sections (e.g., front, back, left, right, top, bottom) into a 2D image that is visually-distorted and the section boundaries are not clearly divided, making the direction of objects hard to interpret. Signaling the place where the described object or event is taking place could create a sense of involvement in the virtual scene. CYT highlighted this issue from the describers' perspective: *"These sections should be distinct so that the describers are able to describe the concrete direction of the objects and the whole spatial compositions of the scene which are both essential to construct mental map."* Headline [7] or clock order [39] are frequently-cited methods to mark the direction for BVI people, echoing with CYT's prior experience: *"I typically used clock position to inform the direction of the work I was going to describe and a human guide would help correct their orientation."*

3.3.3 Constructing mental model with conceivable information. Besides directions, conceivable information such as the number or size of objects is also essential to construct BVI people's mental model towards the described content, as confirmed by CYT: *"Life-related representation for describing size would facilitate blind people to construct their mental model more easily, you can say the animal is aligned with your knee instead of saying it's 50cm tall ... or typically when I describe a pearl necklace, I would count the number of pearls to make blind people imagine the fineness."* Providing key information of surroundings is also necessary for mental model construction, as CYT remarked *"You need to help blind people construct mental model for each new scene by providing information, like how you jump into the scene, what's inside the scene, what causes visual impacts to you. All should be equally presented to blind people in an organized way ... imagine when you (sighted people) reach a new place, you would look around to get visual components, and assemble them together as a complete scene."*

3.3.4 Interactivity and agency. Describers considered 360° videos an interactive medium as sighted people could choose where to explore [6–8]. They emphasized the importance of agency for BVI people to control which story world and path to explore that allows them to craft their own experiences as sighted people can. The lecturer in [6] also highlighted the importance of interactivity as an imperative role to make the consumption experience of 360° videos different from that of videos with standard ADs. CYT also underlined that agency and interactivity is crucial for BVI people when accessing new objects or media, and that the prepopulated description would be tedious [6]: *"BVI people would be forced to receive information you prepare for them unlike sighted people's self-exploring experiences in 360° videos."*

In addition, CYT also raised concerns regarding high-tech support as most describers are not tech-savvy, and typically used off-the-shelf video editors to manage timelines for ADs, as well as a simple document editor to script in their specialized format.

3.4 Design Goals

Driven by the results from our formative study, we present our design goals of OmniScribe as follows:

G1 - Facilitating holistic 360° content understanding. As noted that 360° content comprises in-parallel focuses and unclear division (Section 3.3.1 & 3.3.2) that are hard to be fully and correctly perceived, one of our goals is to facilitate the understanding of 360° content for describers through the description authoring interface.

G2 - Providing conceivable information for mental model construction. To help BVI people construct the mental map of the new scene in 360° and promote a sense of immersion, one of our goals is to conveniently provide conceivable information such as direction or number of objects for the describers to use.

G3 - Enabling immersive labels. As mentioned in Section 3.3.4, the interactivity and control agency would be a crucial factor to make the experience of 360° video different from that of videos with standard ADs. Thus, one of our aims is to enable and streamline the creation of immersive labels for describers and BVI people.

G4 - Low floor for all population. Aside from considering the non-tech-savvy backgrounds of professional describers, we also anticipate more non-professionals could participate in AD production, same as the vision of prior works [2, 16, 34, 38] that sourced from the non-professional population to address the considerable requests of AD. Hence, OmniScribe should have a low entry barrier for the broader population to easily learn and use.

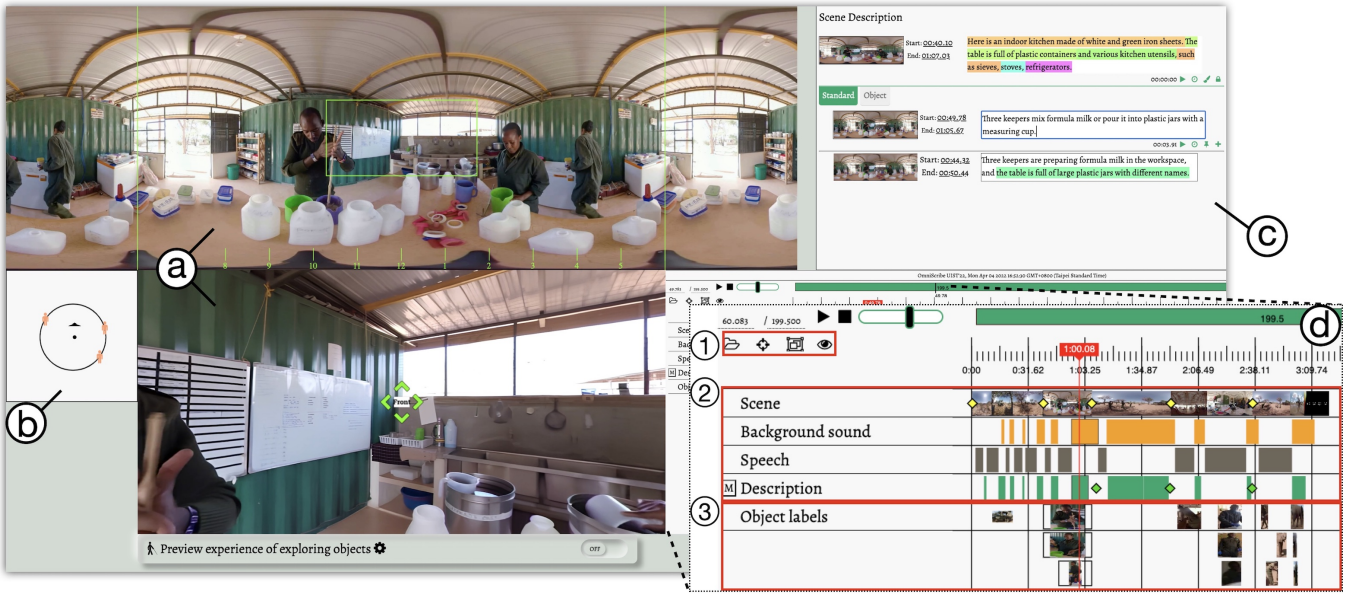

Figure 3: Overview of the OmniScribe description authoring interface. (a) The augmented equirectangular view with clock meter (top) and NFOV (bottom). (b) Content map presenting dynamic objects and viewing angle. (c) Description authoring panel for authoring standard AD, as well as scene and object descriptions. (d) Timeline panel (zoomed-in view): 1) toggles for selecting video files, section division, object tracking, and saliency overlays (left-right); 2) timelines for the scene, background sound, speech, and description. The yellow diamonds marked in the scene layer represent the created scene descriptions. The green diamonds marked in the description layer represent the extended descriptions; and 3) the created object descriptions are visualized by object thumbnails.

4 OMNISCRIBE

OmniScribe includes a web authoring interface for general users to create immersive ADs, and a mobile prototype for BVI people to access them for 360 videos.

The front-end web authoring interface includes: *(i)* a timeline-based navigation panel (Figure 3d) implemented based on Timeliner[1] that visualizes the layers of shot boundary by video thumbnails, video background sound, video speech, description, and labeled objects, *(ii)* an authoring panel for authoring scene descriptions, standard ADs, and object descriptions (Figure 3c), and *(iii)* an Equirectangular view and a Normal Field of View (NFOV) of video with overlay content (Figure 3a). The front-end interface was powered by several Javascript libraries, including fabric.js,[2] panolens.js,[3] and rangy-highlighter.js.[4] OmniScribe also employs a preprocessing pipeline to retrieve visual data from many state-of-the-art models (Section 4.4). We also used the Google text-to-speech (TTS)[5] service to generate .MP3 files of ADs and estimate their duration. The back-end server was implemented using Python Web Flask, which communicates all the obtained data to the front-end interface through socket-io. Details and more information can be found at OmniScribe.org.

[1]https://github.com/zz85/timeliner
[2]http://fabricjs.com/
[3]https://github.com/pchen66/panolens.js
[4]https://github.com/timdown/rangy
[5]https://cloud.google.com/text-to-speech

Next, we detail the main interface components and their functionalities in the following order: *(i)* components for authoring immersive labels, *(ii)* content-awareness components, *(iii)* proof-of-concept mobile prototype that renders OmniScribe-generated ADs for BVI people, and *(iv)* a video preprocessing pipeline enabling the above features.

4.1 Components for Creating Immersive Labels

In this section, we introduce the three immersive labels that aim to increase the sense of immersion (**G3**) for BVI people beyond standard AD, including spatial AD, scene descriptions, and object descriptions.

4.1.1 Authoring spatial AD. The playback of standard AD is generally monophonic. In contrast, spatial audio can signal directions, which is widely-adopted in immersive media to simulate natural hearing in the real world. Spatial audio was also used in commercial apps [1, 28] to provide spatial information as navigation hints for BVI people. Therefore, we introduce spatial AD to further enhance the sense of immersion by positioning the narrator in the place of the described content to signal its directions (**G2,G3**). To author spatial ADs, the describer can use the brushing tool to paint the sound paths of the selected description on the equirectangular view as shown in Figure 4d. OmniScribe then transforms the 2D painted path into 3D spherical coordinates to be visualized on the video view during future playbacks, and for rendering immersive sound in OmniScribe's mobile prototype (Section 4.3).

Figure 4: Workflow for authoring spatial AD. (a) In the description container, the user can paint the sentence by first highlighting it and then clicking the "brush" icon. (b) The user can click the painted sentence, and (c) OmniScribe will prompt actions required: label path, delete or cancel. (d) After clicking "label path", the user can draw the sound path for the painted sentence on the equirectangular view.

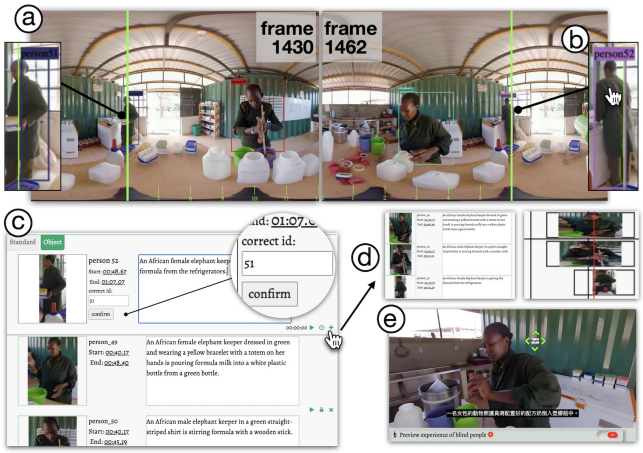

Figure 5: Workflow for authoring object descriptions. (a) The same person moving cross the left and right boundaries in different timestamps makes the object tracker fail (with ID=51 and 52). (b) The user can select the person from the tracking overlay. (c) To amend the false-tracking, the user can input the previous tracking number (ID=51) to confirm their same identity. (d) After completing the description and clicking the "plus" icon, the new object label is appended to the gallery and timeline panel. (e) The user can enter the preview mode by clicking the toggle to simulate the object exploration BVI people will have on the mobile prototype.

4.1.2 Authoring scene descriptions. As mentioned in our formative study, the mental construction of the new environment is a potential key factor for BVI people to immerse themselves in the virtual environment. However, the limited available time gaps in a video make it challenging for users to detail the scene entirely in time. In this regard, OmniScribe preempts an AD slot for each scene (**G4**), allowing the describer to provide more details about them (**G2,G3**). Scenes are automatically detected and segmented once the video is loaded. Scene descriptions can then be manually played by BVI users in the mobile prototype.

4.1.3 Authoring object descriptions. To allow BVI people to immersively interact with the video content beyond only hearing AD of the video (**G3**), OmniScribe enables the describer to select the crucial objects and describe them, which we call object descriptions (Figure 5). These object descriptions can then be rendered through the mobile prototype for BVI users to explore. The moving path of the object was prepopulated (**G4**) using object tracking in the preprocessing stage. The audio path of the object description is automatically mapped to the moving path of the object. Thus, users do not need to spatialize the object descriptions manually with the brushing tool as mentioned above.

4.2 Content-Awareness Components

For video presentation, we included both the equirectangular view and NFOV, as they provide different aspects for users such as *global vs. partial*, and *distorted vs. normal* content. To amend the separation of the left and right boundaries of the equirectangular view, we augmented the video with 25% of the left and right content [42] with a salient green line indicating the augmentation (Figure 3a top). We further added the clock meter model at the bottom of the equirectangular view for users to better interpret the direction of content, which is a common approach to inform BVI users of directions in the real world. Next, we detail the main components for enhancing content-awareness, including view control widgets, section division overlays, saliency and object tracking overlays, and the content map of dynamic objects.

4.2.1 View control widgets. To help users position themselves in the 360° world (**G1,G2**), OmniScribe uses a rectangular view indicator in the equirectangular view (Figure 3a top) to roughly indicate what is presented in NFOV. The view indicator can be panned in either the equirectangular view or NFOV and is synchronized across the two. In NFOV, we also added section control widgets (Figure 3a bottom) for the six sections: top, bottom, left, right, front and back views, which allow users to focus on the desired section by clicking the section tag or shifting using arrows, rather than panning back and forth to position themselves as in off-the-shelf 360° video players.

4.2.2 Section division overlay. As mentioned above, the divisions of the six visual sections are unclear to perceive and interpret. To address this, OmniScribe includes an overlay to the equirectangular view to distinguish all sections by outlining the border of each section (**G1,G2**). The section names are shown on mouse hovering, and each section can be clicked to focus (Figure 6a).

4.2.3 Saliency overlay. The equirectangular image encodes all 360° information in a 2D format that is hard to observe at one time. Therefore, we aimed to increase visual awareness by enhancing the contour of salient objects (**G1**). OmniScribe outlines salient objects with green strokes (Figure 6) generated in Section 4.4.2.

Figure 6: OmniScribe visual overlays: (left) section division outlines the top, bottom, front, back, left, and right. (middle) The visual saliency overlay. (right) The object recognition and tracking overlay.

4.2.4 Object tracking overlay. The visualization of bounding boxes for detected objects can serve as another cue for users to observe the visual flow and follow specific content (**G1**), or infer the number of objects (**G2**). Therefore, OmniScribe presents the object bounding boxes in another visual overlay (Figure 6). The object bounding boxes also allow users to easily author object descriptions (**G4**) as shown in Figure 5 and Section 4.1.3.

4.2.5 Content map of dynamic objects. To enhance the direction-awareness of surrounding objects (**G2**), OmniScribe further visualizes the detected objects into a circular map by centering the viewer and placing the iconic representations around (Figure 3b). Once an icon is clicked, the user will be automatically guided to the clicked object in the other video views. A viewing compass is rendered to indicate the direction of facing and the corresponding field of view.

4.3 Mobile Prototype for Rendering OmniScribe-Generated Descriptions

To consume the OmniScribe-generated immersive ADs (**G3**), we developed a proof-of-concept mobile prototype on iOS, using iPhone and AirPods as the hardware setup (Figure 1d). Using our app, BVI people can listen to spatial ADs during the video playback. The smartphone will vibrate to notify users of scene transitions, and users can then proactively access and listen to the scene descriptions by tapping the screen to pause the video. After the playback of a scene description is finished, users can explore the spatially-anchored object descriptions by turning around (Figure 7).

The OmniScribe mobile prototype can automatically render the immersive ADs for the video using the data (.MP3 and .JSON) generated from the web authoring interface. Our app uses GvrAudioEngine[6] to render the spatial audio for users to hear using headphones. The headphones with gyroscope provide data of the users' head orientation to the smartphone app. OmniScribe-generated ADs can thus be updated in real-time based on the users' head orientation to ensure the immersive labels are placed in correct global 3D coordinates.

4.4 360° Video Preprocessing Pipeline

The functionalities mentioned above are powered by our custom 360° video preprocessing pipeline, as detailed below.

4.4.1 Shot boundary and audio segmentation. In the timeline panel, OmniScribe automatically visualizes the *Scene*, *Background Sound* and *Speech* layers once the video is loaded. To achieve this, we utilized Doukhan et al.'s CNN-based sound segmentation model [5] to segment regions in audio as noise, music, and speech, then visualized the speech and non-speech audio as separated layers (Figure 3d.2). To detect shot boundaries, we applied TransNet V2 [41] to obtain the time frame of each boundary and visualized the first frame of each identified shot to the *Scene* layer (Figure 3d.2).

Figure 7: Demonstration of the mobile prototype. Using a smartphone and a headphone, BVI users can acquire the immersive ADs such as object descriptions by orienting themselves to the anchored objects in the 3D space.

4.4.2 Saliency detection. In OmniScribe, the main purpose of salient object detection (SOD) is to help users localize crucial objects under different circumstances, such as dark scenes, scenes with multiple foci, etc. To achieve this, we used the TRACER [20] model that features SOD with its incorporation of attention-guided tracing modules and adaptive pixel intensity loss. However, objects in the flattened equirectangular images might be distorted or split, making the saliency detection challenging. To address this, we converted an equirectangular image into a cubemap, which contains six faces: Up, Down, Front, Back, Right, and Left. We concatenated all combinations of two adjacent faces (i.e., BD, BL, FD, FR, LD, LF, RB, RD, UB, UF, UL, UR) as twelve input images for TRACER. After inferencing, the saliency maps are stitched back to one complete cubemap and then converted to the equirectangular format. The contours of the detected salient area is then used for visualization.

4.4.3 Object tracking. For object tracking, we used YOLOX [10] as the object detector trained on the COCO dataset [22], and ByteTrack [52] as the object tracker. However, the objects distorted or split in an equirectangular image are challenging to track. To address this, we slightly decreased the detector confidence threshold and increased the aspect ratio limit of the tracker to reduce the loss of tracking caused by distortion. Additionally, we augmented 25% of the left and right content of the equirectangular map to alleviate circumstances where the tracked target is split across two

[6]Link to Google GvrAudioEngine documentation

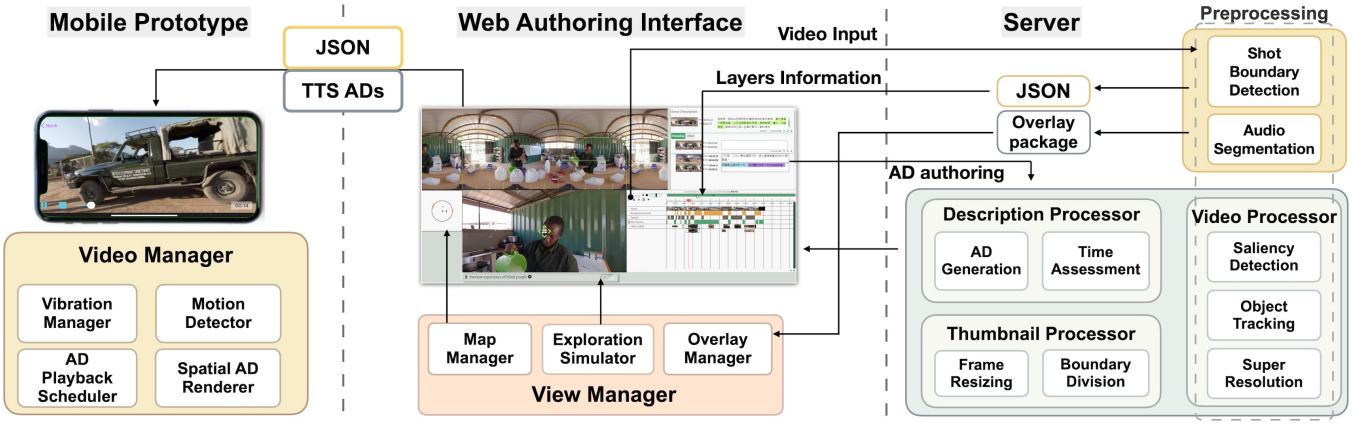

Figure 8: OmniScribe system diagram, which consists of three parts: server, web client, and mobile app. The server is responsible for video preprocessing, and rendering the layer information, overlays, TTS ADs, and thumbnails to the web client. The web client uses a view manager to render the above information based on the video timeline. The output of the web client is then used in the mobile app to render immersive labels and vibrations based on users' controls in real-time.

boundaries and recognized as two unique objects. We also utilized the shot boundary detection module [41] to separate the video into individual clips. For each clip, we computed the complexity score based on the number of detected objects. If the complexity score is higher than our predefined threshold, we reiterated ByteTrack to re-associate detected objects to improve the tracking precision.

4.4.4 Super resolution. The NFOV retrieves a partial region of the equirectangular view, which degrades the resolution when being scaled to the normal viewing size on the screen. To address this, we used SwinIR [21], a state-of-the-art image restoration method based on the Swin Transformer [26], to restore high-quality images from low-quality ones with its feature extraction and image reconstruction modules.

5 EVALUATION: DESCRIBERS USING OMNISCRIBE TO CREATE DESCRIPTIONS

In this evaluation, we aim to understand *(i) How is the usability of OmniScribe? (ii) How novice and professional describers use the OmniScribe functionalities to describe 360° videos?* and *(iii) How describers assess the quality of ADs authored using OmniScribe?*

5.1 Video Materials

We selected two videos around three minutes long from different source providers and domains (In 360: The teenager learning combat tactics at school - BBC News[7] and Reteti Elephant Sanctuary 360 VR - San Diego Zoo Safari Park[8]). The two videos comprise scenes with different types of ROIs [24], and have a regular and low portion of narration (46.2% and 2.97% time of video respectively).

5.2 Participants

We recruited both novice and professional describers for our study in order to explore the diverse usage of OmniScribe. Eight participants (5 M and 3 F) were recruited through public recruitment in our university, aged 22 to 25 (mean=22.6), who did not have any experience using or authoring AD in the past. Three professional describers (3 F) were recruited, including our co-author CYT (age 52), and two other professional describers (age 42 and 49), who had experience in making AD for over three years. In the following sections, we refer to the eight novice describers as N1-N8, CYT as P1, and the other two professional describers as P2 and P3.

5.3 Tasks

To understand if OmniScribe can better support AD authoring for 360° videos, we created a modified version of OmniScribe as a Baseline (Figure 9) which includes both the equirectangular view and NFOV that a commercial timeline-based video editing software (e.g., Final Cut Pro) typically provides. Additionally, we included the text editing capabilities of the description panel in the Baseline condition. This represents a strong baseline; as to our knowledge, there is no existing tool available to specifically support the authoring of ADs for 360° videos. In summary, the Baseline interface included *(i)* an equirectangular view and the corresponding NFOV, *(ii)* a panel for description authoring, and *(iii)* a timeline navigator displaying the tracks of background sound and speech.

The tasks were then counterbalanced among the two software conditions (Baseline vs. OmniScribe) and the two videos, where each participant was assigned to use one software to describe one video in the first task and use the second software to describe the second video in the second task. For the three professional describers, we ensured that two different videos were described at least once using different software.

5.4 Procedure

To familiarize novices with AD and 360° video, we prepared a brochure that documented the current AD guidelines, sample videos with standard AD, and the introduction of 360° videos. We sent it to our participants a week before the study to review in detail, and encouraged them to think about how to describe 360° videos. For professional describers, we sent out another brochure with only introductions of 360° videos and had informal conversations to familiarize them with 360° videos and the purpose of our study.

[7] In 360: The teenager learning combat tactics at school - BBC News
[8] Reteti Elephant Sanctuary 360 VR - San Diego Zoo Safari

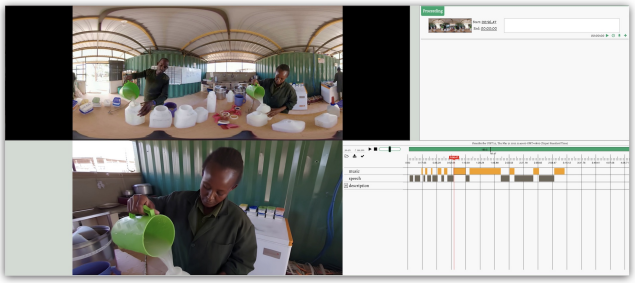

Figure 9: Interface for our Baseline condition, which includes the equirectangular view and NFOV (similar to Final Cut Pro), the description panel, and the timeline panel with visualized audio profiles of the video.

At the beginning of the study, we briefly reviewed the brochure with our participants to ensure their understanding of AD and 360° video. Participants were then asked to perform the two tasks, followed by a semi-structured interview after each task to rate (using 7 Point Likert scales) and comment on each interface feature, and self-assess the quality of their authored ADs. On average, novice describers took around 4 hours to complete the study, while professional describers took about 8 hours to meet their personal standards of AD quality. Our study was approved by our institution's IRB, and participants consented to participate in the study through email or verbal consent before the study started. Participants were compensated with a rate of $10/hour (novice) and $20/hour (professional) for their participation.

5.5 Results: How Describers Used OmniScribe?

In this section, we report the usability of OmniScribe and the usage scenarios of each functionality observed in our study from both novice and professional describers' perspectives.

5.5.1 Usability. Overall, both novice (M=5.38, SD=0.74) and professional (M=5.33, SD=0.58) describers agreed that OmniScribe is intuitive and easy-to-use. The four novice describers who used OmniScribe in their first task reported that OmniScribe was a bit overwhelming to learn and use at first due to its various functionalities. For professional describers, as they had developed their own practice of authoring AD, they expressed the needs of more advanced features, such as note-taking areas, annotating tools for marking timestamps, speed and volume controls of generated AD.

5.5.2 View control widgets. Both novice (M=5.88, SD=0.83) and professional (M=6.33, SD=0.58) describers found the view control widgets highly useful. The view indicator was especially frequently used to observe the details in the scene by all participants. Section control widgets, however, were found to be more useful by professional describers to quickly position themselves, as P2 commented: *"As I needed to keep panning the NFOV to clarify the details in the scene, the widgets helped me quickly return to the section I want."*

5.5.3 Section division overlay. Although both novice (M=5.00, SD=1.85) and professional (M=5.00, SD=2.65) describers generally agreed that the section division overlay could help them easily position the video content, there was some variations in their opinions and usage. Novice describers mostly used the overlay ad-hoc, such as

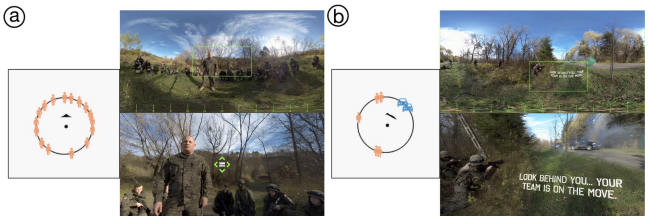

Figure 10: Example video shots where describers found the content map to be helpful. (a) The viewer is surround by 16 people. (b) There are people behind the viewer that are hard to observe, but are obvious from the content map.

when describing content with specific positions, or when creating spatial AD with precise paths. In contrast, two professional describers (P1, P2) regarded the overlay as a highly useful tool that they would always keep on when watching the 360° video to construct their spatial understanding of the scenes. Professionals took NFOV as the primary information source and the equirectangular view as a supplement, as commented by P1: *"Other than watching the video the first time, I would like to always turn on this overlay to position myself ... The occlusion is not a problem at all as the (equirectangular) view is supplementary to me, which gave me the entire story ... I described content mainly based on NFOV, which is clear and detailed."* However, one professional describer (P3) who rated 2 was still concerned about occlusion issues.

5.5.4 Object tracking overlay. Both novice (M=5.63, SD=0.92) and professional (M=5.33, SD=1.15) describers found object tracking overlay to be useful. Besides labeling object descriptions, six participants commented that the overlay allowed them to quickly grasp what they can describe, as illustrated by N7 *"This overlay was one I always turned on ... it largely saved my time to find focus"* and N1 *"This could be a reminder for me to not miss the detail of the video."* Professional describers also shared the same opinions. They valued this function when wanting to understand precise and conceivable information, as P2 noted: *"It was like a notification for me to observe something important like there are people here or something is happening out there ... I also find this overlay useful when I counted the number of people."* P1 also considered this overlay to be powerful in understanding the video and counting objects: *"This object overlay made me aware of the visual flow and cluster of dynamic content, which in turn told me where to look at."*

5.5.5 Saliency overlay. Both novice (M=2.88, SD=1.36) and professional (M=4.00, SD=1.73) describers found saliency overlay to be less useful, likely due to its overlapping utility with the object tracking overlay. Both overlays increased content awareness toward individual objects, but the saliency overlay is more distracting due to prediction noises. However, some participants (N4, P2) still found it promising for special circumstances such as foggy or dark scenes. P1 also noted that the saliency overlay could be potentially useful to highlight content they may missed due to small size or low resolution, while P3 deemed the saliency overlay as a black box *"The system should explain the definition of saliency so that I can decide when and how to use it."*

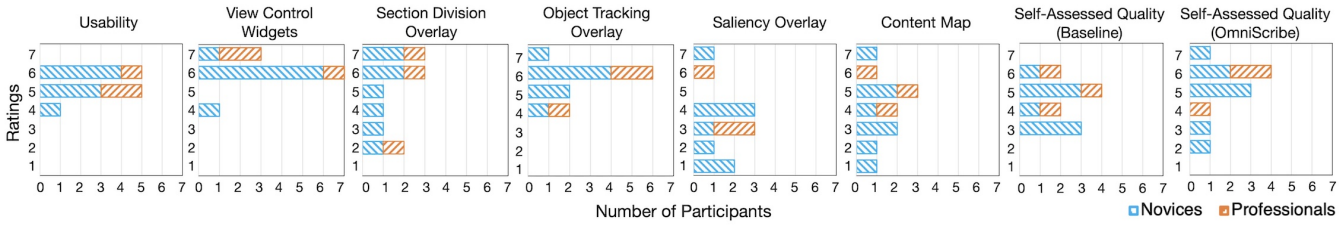

Figure 11: Subjective ratings on each functionality, usability and self-assessed quality of ADs. Novice describers are marked in blue (N=8), and professional describers (N=3) are marked in orange.

5.5.6 Content map. Novice (M=3.75, SD=1.91) and professional describers (M=5.00, SD=1.00) had mixed opinions towards the content map. Some (N2,4,5,7,8) reported that they did not use it during the study, while others (N1,3,6, P1-3) considered the content map useful for specific circumstances in the videos. P2 exemplified her impression when describing the second video: *"At first, I thought this feature was less useful, but when looking into the last scene, I was very impressed and knew I was surrounded by 16 people (Figure 10a)."* P3 also found this feature useful: *"The subtitle said 'look behind you,' so I took a glance on the content map and it turned out that some people were actually behind me (Figure 10b)."* P1, on the other hand, suggested that the content map could be marked with clock positions, and that it could be rendered as a top-down view of the space instead of always placing objects along the circumference.

5.6 Results: How Describers Self-Assessed the Quality of Their Authored AD?

In terms of self-assessed AD quality, ADs authored using OmniScribe were rated slightly higher than the Baseline by both novice (4.88 vs. 4.25) and professional (5.33 vs. 5) describers. We further analyzed the AD scripts generated by our participants, and found that novice describers overall included more directional terms (e.g., front, twelve o'clock, etc.) when using OmniScribe (V1=20, V2=35) compared to Baseline (V1=11, V2=15). However, this difference was not consistently observed across the three professional describers: P1 used seven directional terms in OmniScribe and ten in Baseline, P2 used seven in OmniScribe but none in Baseline, and P3 used two in OmniScribe and none in Baseline.

Overall, all novice participants struggled with selecting crucial content to describe in time when using both OmniScribe and Baseline. Most novice describers could not actually judge if their description met the standard quality as they were neither the professional creators nor the BVI consumers of the ADs, which explains their neutral and conservative self-assessment ratings of their AD quality. Using the Baseline interface, some novices mentioned that their strategies were to focus on either the equirectangular view or NFOV to describe what they see, as stated by N6 who primarily focused on the equirectangular view: *"I did not focus on a specific angle of video, I just kept watching the global one (equirectangular view) and introduced the objects in the view, and thus I used fewer terms related to direction."* In contrast, N4 focused more on NFOV: *"With these two views, I cannot tell where I was in the video world, so I just roughly described what I saw in NFOV."* Interestingly, N1 rated their AD quality a 3 for Baseline and 2 for OmniScribe due to the increased amount of perceived content in OmniScribe: *"The reason I rated lower for OmniScribe is because I found so much information in the 360° video, and I did not know how to describe them all in time. I did not confront this situation when using Baseline first."*

As for professional describers, all of them reported that a day was not enough to describe two three-minute videos; for them, a single three-minute video would typically take many days to complete, and it will be even longer for 360° videos. Hence, they were not quite satisfied with their AD quality. P2, who rated 4 for both Baseline and OmniScribe, said: *"These would be the first draft. Typically we would have multiple rounds to refine the script, which I cannot do them all today."* Overall, all three professional describers commented that OmniScribe provided more information and was helpful for their understanding of the video content. However, they still struggled with the trade-offs between BVI-desired content and directional information which is unique to 360° videos, as P1 pointed out during the study: *"I was trade-offing between directional information and BVI-desired content like the girl with blond hair or blue eyes. It's hard to describe them all in time."* P2 also commented: *"Though OmniScribe made me aware of more content, the decision to describe what is still on me. I got no time to describe directional information, and I think blind people cannot know this is 360° video out of my script."*

6 EVALUATION: BVI PEOPLE CONSUMING 360° VIDEOS WITH IMMERSIVE LABELS

In this evaluation, we aim to understand *whether our proposed immersive labels can improve the sense of immersion for BVI users, including spatial AD, scene descriptions, and object descriptions?*

6.1 Participants

Through word-of-mouth from the authors' connections, we recruited 8 blind people (6 M and 2 F) aged from 20 to 32 (mean=24.88). All of them had prior experiences in consuming AD from different media sources. Three out of eight were blind since birth while the other five lost their vision later in life (Table 1). We refer to our BVI participants as B1-B8 in the following sections.

6.2 Materials and Apparatus

For the two videos, we took the ADs made by CYT in the previous evaluation as the standard version. To control for the description content, we spatialized them into the immersive OmniScribe version; we also collected and synthesized the scene and object descriptions authored in the previous evaluation, and ensured the

Table 1: Demographic information of BVI participants.

ID	Age	Gender	Vision Level	Education
B1	23	Male	Blind, later in life	Undergraduate
B2	21	Female	Blind, since birth	Undergraduate
B3	23	Male	Blind, since birth	Undergraduate
B4	32	Male	Blind, later in life	Doctoral
B5	32	Male	Blind, since birth	Doctoral
B6	20	Female	Blind, later in life	Undergraduate
B7	27	Male	Blind, later in life	Master
B8	21	Male	Blind, later in life	Undergraduate

descriptions are grammatically correct and usable. Each video thus has two versions. In summary, for OmniScribe version, the video *Elephant Sanctuary* (V1) had 21 spatial ADs, 12 object and 9 scene descriptions. The second video *Combat Tactics* (V2) had 20 spatial ADs, 13 object and 7 scene descriptions. Participants were asked to wear AirPods Max and turn on the noise-canceling mode when watching videos, while holding the smartphone to feel the vibrations during scene transitions to access the scene and object descriptions.

6.3 Procedure

Participants were first introduced to the study and asked about their prior experiences in consuming AD as well as how they imagine BVI people could consume 360° videos in a way different from standard AD. Participants were then presented with one of the two videos with its standard version and immersive version in the first session, and then the other video in the second session. For each session, participants were allowed to replay and experience any version they want, and were asked to provide feedback on each feature and their preference between the two versions. Our study was approved by our institution's IRB, and participants provided verbal or email consent. The study took an average of one hour to complete, and each participant was compensated with a rate of $15/hour for their participation.

6.4 Results

For the standard version of both V1 and V2, participants took the same time as the video duration to complete watching the video. For the immersive OmniScribe version, participants spent an additional 236 seconds (V1) and 255s (V2) pausing each video to access the scene and object descriptions. For the discoverability of object descriptions, participants achieved an average of 78.4% in 9.4s/object (V1) and 80.2% in 11.5s/object (V2). However, testing the time and ability to discover labels was not the focus of this study. Instead, we were interested in soliciting in-depth feedback from BVI users about the OmniScribe immersive labels by comparing with a standard baseline version. Overall, all participants preferred OmniScribe to the standard version due to the higher interactivity, sense of immersion, and information details for both videos.

6.4.1 Spatial ADs. For both videos, all participants (mean=6.13 for V1, 6.25 for V2) agreed that spatial ADs rendered a more immersive experience by signaling the direction and dynamics of the described content. However, half of the participants were also concerned that the frequently-jumping ADs would pose a heavier cognitive load

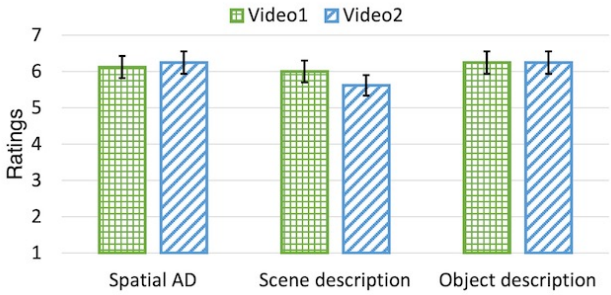

Figure 12: Subjective ratings on the helpfulness of spatial AD, scene and object descriptions in OmniScribe to create a sense of immersion for 360° videos.

for them to decode the directions of spatial ADs, as mentioned by B2: *"It was tiring to interpret the direction of the description, like there is an additional layer of information imposed on me to understand."* On the other hand, some participants (N=3) considered spatial ADs with directional terms a useful combination to help them position the described content faster, as mentioned by B3: *"The describer appeared to my left jointly with his description 'there is a man on my left' made it easy for me to determine the direction of the man."* Besides spatial ADs, most participants (N=7) suggested that the original audio could also be spatialized to create a more immersive experience.

6.4.2 Scene descriptions. Scene descriptions were used to introduce scenes in detail, which required users to pause the video to trigger. Many participants (N=4) commented that the video pause was less favorable, but the acquisition of much detailed information was valuable, as B1 mentioned: *"The scene descriptions helped me construct the video scene in detail without the time constraint. It's worth a pause."* Considering the pauses, two participants mentioned that scene descriptions would be useful for further exploration after having a basic understanding of the video, rather than accessing a new video for the first time as it could disrupt the flow. All participants liked the vibration feedback as a salient notification of the scene transitions.

6.4.3 Object descriptions. All participants were very enthusiastic about object descriptions. B5 was impressed during the study: *"It's amazing that I can take actions to interact with the content. This can enhance my impression of the interactive content and increase my engagement with the video."* B6, on the other hand, used object descriptions to confirm her understanding of the video: *"It's interesting that I can turn around to see if the things I comprehended from the ADs are correct, like the direction of objects in the scene. I would be more confident about the video."* B1, who was also passionate about the feature, commented: *"It's promising to turn a video into a game that I can physically interact with, like I can turn around or even walk to see the world."* Most participants (N=7) also mentioned that it may be helpful to first provide a list of actionable items in the scene, so that they are aware of the content to explore.

7 DISCUSSION AND FUTURE WORK

We discuss the implications of our research, including co-designing immersive experiences with BVI people, the tradeoffs of immersion and cognitive load, enabling more interactive mobile experiences, the roles of describers and video creators, the possibilities to power OmniScribe with the crowd, and generalizing OmniScribe to other media formats and stakeholders.

7.1 Co-designing Experiences with BVI People

In our work, we aimed to promote 360° video accessibility for BVI people. As a first step to achieving this, we identified the needs and challenges of describers in authoring AD for 360° videos, and created tools to support it. Future work should further investigate the needs from BVI people's perspective, which would inform techniques for presenting and interacting with our proposed spatial AD along with scene and object descriptions, as well as alternative experiences and hardware accessories for creating a stronger sense of immersion and interactivity. Our user evaluation with BVI participants revealed initial promises of immersive AD over standard AD for 360° videos, and future work could investigate what types of immersive labels are best suited for what types of videos, for a variety of purposes in video consumption by BVI users.

7.2 Roles of Describers and Video Creators

Though selecting content to describe is challenging, professional describers felt more involved and equipped with informative cues when describing 360° videos compared to traditional 2D ones, which allowed them to infer the context and judge the information fidelity more easily, as P2 pointed out: *"360° videos allow me to find more cues in the scene for making correct descriptions, unlike for the general videos sometimes I did not know where the sound comes from and have nowhere to find it."* However, in contrast to the informative 360° cues, P1 regarded 360° videos as an immature media due to the indefinite focus. P3 also noted that describing 360° videos has transformed their role from a describer into a director or a content creator who decides the main storyline presented to BVI people. Future work could explore how content creators could communicate their intent with describers, while leveraging the 360° capacity to provide more context and sufficient cues.

7.3 Tradeoffs of Immersion and Cognitive Load

As 360° videos emphasize their immersive capacity, CYT mentioned that describing from the second-person point of view using *"you are in ..."* could be preferable in creating a sense of immersion to describing from a third-person perspective. This was also mentioned by many BVI participants when asked to imagine the experience of 360° videos prior to the study. However, as mentioned in Section 6.4.1, there were tradeoffs between the sense of immersion and cognitive load when using spatial ADs. Given the small sample size and the subjectiveness of feedback, we cannot conclude when and how the balance can be achieved between immersion and cognitive load. Future work could investigate techniques of describing 360° videos to create a better sense of immersion while minimizing the cognitive load for BVI people through more rigorous and objective measurements (e.g., EEG, or fNIRs).

7.4 Interactive Mobile Systems for 360° Videos

We proposed a mobile prototype to render OmniScribe-generated ADs. One of the functionalities is to enable BVI people to acquire the spatially-anchored object descriptions by turning around. However, the directional terms in the ADs would fail once BVI users turn themselves to different viewing angles. Future work could explore making the content of ADs *responsive* to user states and actions to provide accurate descriptions. Moreover, a variety of interaction techniques of mobile systems could be integrated to access virtual content, such as different in-air pointing techniques, touch gestures, or blind photography [13], etc. Future systems could also provide a diverse range of presentations of video content, such as sonifying the depth or size of virtual objects, using different voice fonts to represent different information layers (e.g., object and scene descriptions) [19], using vibration patterns to represent different objects and textures, or designing haptic overlays or proxies to enrich the non-visual experiences for BVI people.

7.5 Powering OmniScribe with the Crowd

As mentioned in Section 2.1, the limited capacity of professional describers cannot keep up with the large volume of AD requests, which necessitates support from the crowd [11, 16, 35, 51, 55]. Furthermore, professional AD production typically requires further iterations of edits, vocal acting, audio engineering, and proof-listening by BVI people, which could take several days even for a short 2D video, and possibly longer for a 360° video. In Section 5.6, we also found individual differences in the usage of directional terms across our participants. Therefore, we imagine many future opportunities to utilize crowdsourcing to aid the process of 360° video description authoring. For example, future work could investigate how to streamline the authoring process by the crowd, along with quality control, novice training, prompting usage of essential terms for 360° videos, and the use of second or third-person perspective when describing 360° videos. Furthermore, different from 2D videos, 360° videos provide more information and ROIs that could be watched from a variety of perspectives. Crowds could therefore be used to author different storylines, so that BVI people could consume the videos in different ways or judge the information fidelity by comparing multiple versions. Additionally, OmniScribe could retrieve the viewing and fixation data from online audiences to highlight popular ROIs for describers, further augmenting the authoring process.

7.6 Generalizing OmniScribe

As shown in Figure 2, although OmniScribe was scoped to make 360° videos accessible, we envision extending OmniScribe in the future to other media formats and stakeholders. For instance, we see its potential in supporting the accessibility of other media formats such as 2D videos and images, which could also be infused with spatial AD or immersive labels to facilitate rich exploration and interactive information acquisition for BVI people. Furthermore, OmniScribe enables the creation of spatial ADs, which could be used for creating responsive subtitles for people with hearing impairments, or creating responsive visual hints for sighted people to increase the visual awareness on 360° content [23, 24].

8 CONCLUSION

We have presented OmniScribe, a system designed to make 360° videos immersive and accessible. Through a formative study, we identified the challenges and needs of describing 360° videos and how to render more immersive experiences for BVI people. We implemented several content-awareness components for describers to observe the holistic video content, such as section division overlays, saliency overlays, object tracking overlays, content maps, and view control widgets. OmniScribe also enables the user to author immersive labels for BVI people to consume on our developed mobile prototype, including spatial AD, scene descriptions, and object descriptions. Through an evaluation with both novice and professional describers, we demonstrated the usability of OmniScribe and how people perceived and used each functionality when describing 360° videos. In another evaluation with BVI people, we demonstrated the promise of the immersive labels to create a more immersive experience of 360° videos. Finally, we discussed lessons learned, and how OmniScribe can be augmented and generalized to promote immersive media accessibility for everyone.

ACKNOWLEDGMENTS

This research was supported in part by the Ministry of Science and Technology of Taiwan (MOST111-2221-E-002-145-MY3, 111-2218-E-002-028, 110-2634-F-002-051), National Taiwan University, and a Google Research Scholar Award. We thank our anonymous reviewers for their suggestions, all participants of our study, our figure illustrator Yu-Hsuan Kao, and Audio Description Development Association in Taiwan.

REFERENCES

[1] BlindSquare. 2022. BlindSquare. https://www.blindsquare.com/

[2] 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.

[3] Edoardo D'Atri, Carlo Maria Medaglia, Alexandru Serbanati, Ugo Biader Ceipidor, Emanuele Panizzi, and Alessandro D'Atri. 2007. A system to aid blind people in the mobility: A usability test and its results. In *Second International Conference on Systems (ICONS'07)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 35–35.

[4] Tawanna R Dillahunt, Alex Jiahong Lu, Aarti Israni, Ruchita Lodha, Savana Brewer, Tiera S Robinson, Angela Brown Wilson, and Earnest Wheeler. 2022. The Village: Infrastructuring Community-Based Mentoring to Support Adults Experiencing Poverty. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 574, 17 pages. https://doi.org/10.1145/3491102.3501949

[5] David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018. An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 5214–5218.

[6] Equal Entry. 2022. Audio Descriptions for 360 Degree Video: Best Practices. https://www.youtube.com/watch?v=jOX6gxUZq8w

[7] 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.

[8] Anita Fidyka and Anna Matamala. 2021. Retelling narrative in 360° videos: Implications for audio description. *Translation Studies* 14, 3 (2021), 298–312. https://doi.org/10.1080/14781700.2021.1888783 arXiv:https://doi.org/10.1080/14781700.2021.1888783

[9] Langis Gagnon, Claude Chapdelaine, David Byrns, Samuel Foucher, Maguelonne Heritier, and Vishwa Gupta. 2010. A computer-vision-assisted system for videodescription scripting. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 41–48.

[10] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. 2021. YOLOX: Exceeding YOLO Series in 2021. https://doi.org/10.48550/ARXIV.2107.08430

[11] 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*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376728

[12] José Luis González-Mora, A Rodriguez-Hernandez, Enrique Burunat, F Martin, and Miguel A Castellano. 2006. Seeing the world by hearing: Virtual Acoustic Space (VAS) a new space perception system for blind people.. In *2006 2nd International Conference on Information & Communication Technologies*, Vol. 1. IEEE, Institute of Electrical and Electronics Engineers, New York, NY, USA, 837–842.

[13] Anhong Guo, Saige McVea, Xu Wang, Patrick Clary, Ken Goldman, Yang Li, Yu Zhong, and Jeffrey P. Bigham. 2018. Investigating Cursor-Based Interactions to Support Non-Visual Exploration in the Real World. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility* (Galway, Ireland) *(ASSETS '18)*. Association for Computing Machinery, New York, NY, USA, 3–14. https://doi.org/10.1145/3234695.3236339

[14] Jaylin Herskovitz, Jason Wu, Samuel White, Amy Pavel, Gabriel Reyes, Anhong Guo, and Jeffrey P. Bigham. 2020. Making Mobile Augmented Reality Applications Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 3, 14 pages. https://doi.org/10.1145/3373625.3417006

[15] Hou-Ning Hu, Yen-Chen Lin, Ming-Yu Liu, Hsien-Tzu Cheng, Yung-Ju Chang, and Min Sun. 2017. Deep 360 Pilot: Learning a Deep Agent for Piloting through 360° Sports Videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Institute of Electrical and Electronics Engineers, New York, NY, USA, 1396–1405. https://doi.org/10.1109/CVPR.2017.153

[16] The Smith-Kettlewell Eye Research Institute. 2022. YouDescribe. https://youdescribe.org/

[17] Masatomo Kobayashi, Trisha O'Connell, Bryan Gould, Hironobu Takagi, and Chieko Asakawa. 2010. Are Synthesized Video Descriptions Acceptable?. In *Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility* (Orlando, Florida, USA) *(ASSETS '10)*. Association for Computing Machinery, New York, NY, USA, 163–170. https://doi.org/10.1145/1878803.1878833

[18] A. Lecuyer, P. Mobuchon, C. Megard, J. Perret, C. Andriot, and J.-P. Colinot. 2003. HOMERE: a multimodal system for visually impaired people to explore virtual environments. In *IEEE Virtual Reality, 2003. Proceedings.* Institute of Electrical and Electronics Engineers, New York, NY, USA, 251–258. https://doi.org/10.1109/VR.2003.1191147

[19] Cheuk Yin Phipson Lee, Zhuohao Zhang, Jaylin Herskovitz, JooYoung Seo, and Anhong Guo. 2022. CollabAlly: Accessible Collaboration Awareness in Document Editing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 596, 17 pages. https://doi.org/10.1145/3491102.3517635

[20] Min Seok Lee, WooSeok Shin, and Sung Won Han. 2022. TRACER: Extreme Attention Guided Salient Object Tracing Network (Student Abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 11 (Jun. 2022), 12993–12994. https://doi.org/10.1609/aaai.v36i11.21633

[21] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. 2021. SwinIR: Image Restoration Using Swin Transformer. https://doi.org/10.48550/ARXIV.2108.10257

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

[23] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell Me Where to Look: Investigating Ways for Assisting Focus in 360° Video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 2535–2545. https://doi.org/10.1145/3025453.3025757

[24] Yung-Ta Lin, Yi-Chi Liao, Shan-Yuan Teng, Yi-Ju Chung, Liwei Chan, and Bing-Yu Chen. 2017. Outside-In: Visualizing Out-of-Sight Regions-of-Interest in a 360° Video Using Spatial Picture-in-Picture Previews. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 255–265. https://doi.org/10.1145/3126594.3126656

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

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer

using Shifted Windows. https://doi.org/10.48550/ARXIV.2103.14030
[27] Shachar Maidenbaum, Shelly Levy-Tzedek, Daniel-Robert Chebat, and Amir Amedi. 2013. Increasing accessibility to the blind of virtual environments, using a virtual mobility aid based on the" EyeCane": Feasibility study. *PloS one* 8, 8 (2013), e72555.
[28] Microsoft. 2022. Microsoft Soundscape. https://www.microsoft.com/en-us/research/product/soundscape/
[29] Mario Montagud, Issac Fraile, Juan A. Nuñez, and Sergi Fernández. 2018. ImAc: Enabling Immersive, Accessible and Personalized Media Experiences. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video* (SEOUL, Republic of Korea) *(TVX '18)*. Association for Computing Machinery, New York, NY, USA, 245–250. https://doi.org/10.1145/3210825.3213570
[30] Mario Montagud, Pilar Orero, and Sergi Fernández. 2020. Immersive media and accessibility: hand in hand to the future. *ITU* (2020).
[31] Mario Montagud, Pilar Orero, and Anna Matamala. 2020. Culture 4 all: accessibility-enabled cultural experiences through immersive VR360 content. *Personal and Ubiquitous Computing* 24, 6 (2020), 887–905.
[32] Vishnu Nair, Jay L Karp, Samuel Silverman, Mohar Kalra, Hollis Lehv, Faizan Jamil, and Brian A. Smith. 2021. NavStick: Making Video Games Blind-Accessible via the Ability to Look Around. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 538–551. https://doi.org/10.1145/3472749.3474768
[33] Vishnu Nair, Shao-en Ma, Hannah Huddleston, Karen Lin, Mason Hayes, Matthew Donnelly, Ricardo E Gonzalez, Yicheng He, and Brian A. Smith. 2021. Towards a Generalized Acoustic Minimap for Visually Impaired Gamers. In *The Adjunct Publication of the 34th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 89–91. https://doi.org/10.1145/3474349.3480177
[34] Rosiana Natalie, Ebrima Jarjue, Hernisa Kacorri, and Kotaro Hara. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, Greece) *(ASSETS '20)*. Association for Computing Machinery, New York, NY, USA, Article 87, 4 pages. https://doi.org/10.1145/3373625.3418030
[35] Rosiana Natalie, Jolene Loh, Huei Suen Tan, Joshua Tseng, Ian Luke Yi-Ren Chan, Ebrima H Jarjue, Hernisa Kacorri, and Kotaro Hara. 2021. The Efficacy of Collaborative Authoring of Video Scene Descriptions. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 17, 15 pages. https://doi.org/10.1145/3441852.3471201
[36] American Council of the Blind. 2022. The Audio Description Project. https://adp.acb.org/guidelines.html
[37] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology* (Québec City, QC, Canada) *(UIST '17)*. Association for Computing Machinery, New York, NY, USA, 289–297. https://doi.org/10.1145/3126594.3126636
[38] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864
[39] Daisuke Sato, Uran Oh, Kakuya Naito, Hironobu Takagi, Kris Kitani, and Chieko Asakawa. 2017. NavCog3: An Evaluation of a Smartphone-Based Blind Indoor Navigation Assistant with Semantic Features in a Large-Scale Environment. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) *(ASSETS '17)*. Association for Computing Machinery, New York, NY, USA, 270–279. https://doi.org/10.1145/3132525.3132535
[40] Alexa F. Siu, Mike Sinclair, Robert Kovacs, Eyal Ofek, Christian Holz, and Edward Cutrell. 2020. Virtual Reality Without Vision: A Haptic and Auditory White Cane to Navigate Complex Virtual Worlds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376353
[41] Tomáš Souček and Jakub Lokoč. 2020. TransNet V2: An effective deep network architecture for fast shot transition detection. https://doi.org/10.48550/ARXIV.2008.04838
[42] Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman. 2016. Pano2Vid: Automatic Cinematography for Watching 360° Videos. https://doi.org/10.48550/ARXIV.1612.02335
[43] Virgil Tiponut, Zoltan Haraszy, Daniel Ianchis, and Ioan Lie. 2008. Acoustic Virtual Reality Performing Man-Machine Interfacing of the Blind. In *Proceedings of the 12th WSEAS International Conference on Systems* (Heraklion, Greece) *(ICS'08)*. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 345–349.
[44] MA Torres-Gil, O Casanova-Gonzalez, and José Luis González-Mora. 2010. Applications of virtual reality for visually impaired people. *WSEAS transactions on computers* 9, 2 (2010), 184–193.
[45] World Wide Web Consortium (W3C). 2022. Audio Description or Media Alternative. https://www.w3.org/TR/2008/REC-WCAG20-20081211/#media-equiv-audio-desc
[46] World Wide Web Consortium (W3C). 2022. Providing a movie with extended audio descriptions. https://www.w3.org/TR/WCAG20-TECHS/G8.html
[47] World Wide Web Consortium (W3C). 2022. W3C Image Concepts. https://www.w3.org/WAI/tutorials/images/
[48] World Wide Web Consortium (W3C). 2022. XR Accessibility User Requirements. https://www.w3.org/TR/xaur/
[49] Miao Wang, Yi-Jun Li, Wen-Xuan Zhang, Christian Richardt, and Shi-Min Hu. 2020. Transitioning360: Content-aware NFoV Virtual Camera Paths for 360° Video Playback. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. Institute of Electrical and Electronics Engineers, New York, NY, USA, 185–194. https://doi.org/10.1109/ISMAR50242.2020.00040
[50] Beste F. Yuksel, Soo Jung Kim, Seung Jung Jin, Joshua Junhee Lee, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Ilmi Yoon, Yue-Ting Siu, and Joshua A. Miele. 2020. Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382821
[51] Mingrui Ray Zhang, Mingyuan Zhong, and Jacob O. Wobbrock. 2022. Ga11y: An Automated GIF Annotation System for Visually Impaired Users. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 197, 16 pages. https://doi.org/10.1145/3491102.3502092
[52] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. 2021. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. https://doi.org/10.48550/ARXIV.2110.06864
[53] Yuhang Zhao, Cynthia L. Bennett, Hrvoje Benko, Edward Cutrell, Christian Holz, Meredith Ringel Morris, and Mike Sinclair. 2018. Enabling People with Visual Impairments to Navigate Virtual Reality with a Haptic and Auditory Cane Simulation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173690
[54] Yuhang Zhao, Edward Cutrell, Christian Holz, Meredith Ringel Morris, Eyal Ofek, and Andrew D. Wilson. 2019. SeeingVR: A Set of Tools to Make Virtual Reality More Accessible to People with Low Vision. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300341
[55] Yu Zhong, Walter S. Lasecki, Erin Brady, and Jeffrey P. Bigham. 2015. RegionSpeak: Quick Comprehensive Spatial Descriptions of Complex Images for Blind Users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2353–2362. https://doi.org/10.1145/2702123.2702437