Toward Automatic Audio Description Generation for Accessible Videos

Yujia Wang
Beijing Institute of Technology
George Mason University

Wei Liang*
Beijing Institute of Technology

Haikun Huang
George Mason University

Yongqi Zhang
George Mason University

Dingzeyu Li
Adobe Research

Lap-Fai Yu
George Mason University

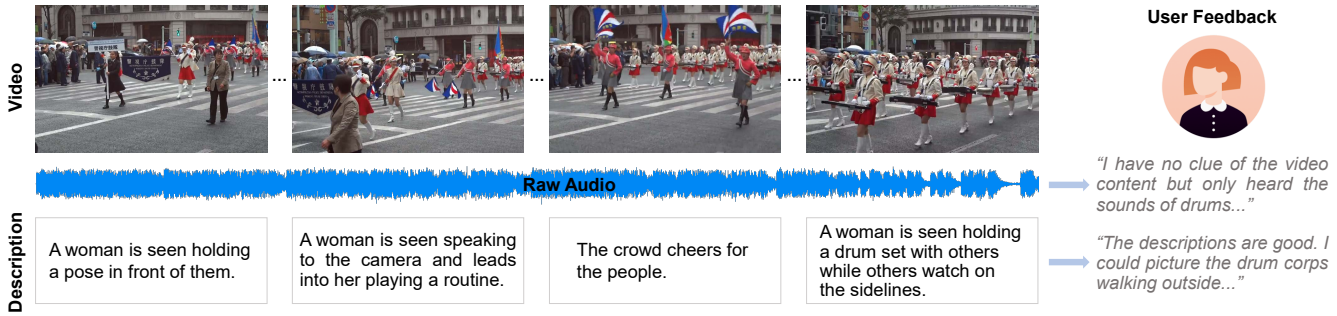


Figure 1: Our audio description system automatically generates audio descriptions for videos. Our user study shows that, with our audio descriptions, blind participants are more confident about what's happening in a video, whereas they expressed uncertainty listening to only the input raw audio.

ABSTRACT

Video accessibility is essential for people with visual impairments. Audio descriptions describe what is happening on-screen, *e.g.*, physical actions, facial expressions, and scene changes. Generating high-quality audio descriptions requires a lot of manual description generation [50]. To address this accessibility obstacle, we built a system that analyzes the audiovisual contents of a video and generates the audio descriptions. The system consisted of three modules: AD insertion time prediction, AD generation, and AD optimization. We evaluated the quality of our system on five types of videos by conducting qualitative studies with 20 sighted users and 12 users who were blind or visually impaired. Our findings revealed how audio description preferences varied with user types and video types. Based on our study's analysis, we provided recommendations for the development of future audio description generation technologies.

*Corresponding author.

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445347

CCS CONCEPTS

• **Human-centered computing** → **Accessibility systems and tools**; *Accessibility technologies.*

KEYWORDS

audio description, video description, audio-visual consistency, video captioning, sentence-level embedding, accessibility

ACM Reference Format:
Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3411764.3445347

1 INTRODUCTION

In 2019, videos are accounted for more than 60% of total internet downstream traffic [1]. Visual information is ubiquitously used for communication, yet people who are blind or who have low vision cannot directly consume video content. Audio descriptions (AD) explain the audiovisual events happening in a scene that are not audible to blind users. According to the American Council of the Blind, out of the 75 million video titles that Amazon Prime Video offers, only around 1,843 (0.0025%) are AD videos[2]. On the other hand, in the past few decades, we have seen steady progress in improving the coverage of closed captions, allowing people who are deaf and hard-of-hearing in enjoying videos. Thanks to deep

[1]https://www.sandvine.com/inthenews/netflix-falls-to-second-place-in-global-internet-traffic-share
[2]https://market.us/statistics/online-video-and-streaming-sites/amazon-prime-video/

learning and automatic speech recognition, platforms like YouTube automatically generate closed captions for all videos in spoken English [39]. In comparison, the percentage of videos with AD is significantly less than those with closed captions. Particularly, a vast majority of user-generated content shared online does not have AD, making them challenging for blind users to consume. Unlike automatic closed caption, audio descriptions are usually added manually and cost around $30 per hour of video, which is unscalable given the sheer amount of user-generated content [33]. Blind users find it valuable to watch videos, television, or movies independently, which, without an automatic AD generator, is hard to achieve [14]. Moreover, AD could also be used to enhance the visual experience for sighted users [40] (*e.g.*, enhanced memory for visual details) and be a powerful tool to be duly exploited in aiding literacy development, language acquisition, and language learning for vulnerable sighted viewers [43]. In this work, we built an automatic AD generation system and analyzed the challenges regarding current video understanding research.

To investigate the challenges of an automatic AD generation system, we compare it against a standard speech recognition system according to three aspects: when, what, and how to generate the output. In automatic speech recognition systems, all three aspects are fairly deterministic – a good caption system should output the spoken words (the what) faithfully and accurately (the how) at the audio onsets of each word (the when). For AD, none of the three aspects is as well-defined as in automatic speech recognition.

- *When* should an audio description be included? We believe that whenever there is an *inaudible* but *meaningful* event, an AD should be generated. However, it is unclear what is an inaudible meaningful event. For example, a camera zoom-in motion is inaudible but it may (or may not) tell a meaningful visual story; a scene transition is usually meaningful, yet considering the ambient audio alone it may (or may not) already be audible.
- *What* should the audio description say? In every video frame, many audiovisual events could happen, ranging from the main events taking place close to the camera, to the more subtle activities of people and objects in the background. The amount of content and level of detail users want from an audio description can vary considerably [51].
- *How* should we organize the synthesized descriptions? Raw video scene captions are usually repetitive and may be irrelevant to the main story of the video. Even human annotators may need to watch a video several times to create concise yet informative descriptions. Furthermore, for delivering pleasant user experiences, it is preferable to optimize the accuracy and fluency of audio descriptions.

To address the above challenges, we built *Tiresias*, a three-stage end-to-end automatic AD generation system. First, to predict *when* to insert an audio description, *i.e. Insertion Time Prediction* module, we developed an audiovisual inconsistency detection model which analyzes discrepancies between different modalities and different frames. Next, leveraging dense video captioning algorithms, we implemented an *AD generation* module to determine *what* actions/subjects/activities to describe. Third, we propose a *AD optimization* module to generate accurate and fluent output, which is fed to a text-to-speech engine for final rendering. § 3 explains how we extensively use state-of-the-art computer vision and natural language processing techniques in our implementation.

To evaluate the quality of our automatic audio descriptions, we conducted a user study with 20 sighted users and 12 blind or visually-impaired (BVI) users. We selected 30 YouTube videos covering different genres, including music, film, animals, sports, etc. Our predicted insertion positions match very well with where users felt uncertain. For example, we can detect the inconsistency between a drum-playing scene (visual) and sounds of people laughing (audio), where both sighted and BVI users expressed that audio description should have been added to enhance clarity. More broadly, the fully-automatic system *Tiresias* reduces confusion compared to the raw audio input. Specifically, 61.01% of the sighted users and 86.11% of the BVI users found that videos with our generated audio descriptions are less confusing than the raw audio.

Based on user interviews, we evaluated the performance of *Tiresias* with several quantitative metrics, summarized the subjective feedback from the interviewees, and shared our recommendation for future automatic AD generation systems and potential research directions within the space. Overall, 70% of the blind users reported that our automatic AD results are "somewhat helpful". Although we observed that the final AD output usually contains some missing events, blind users still prefer some descriptions to no information at all. Moreover, we observed considerable variation in the amount of content and level of detail each user wants; most existing audio descriptions, even those manually generated and curated, only offer a single version of the description. Hence, we recommend providing an interaction that allows users to select the amount of audio description they want. Moreover, we hope that our work will not only contribute to the design of audio description generation for social and video streaming platforms but also improve navigation skill training for blind children via multi-modal video games [47].

2 RELATED WORK

This research is related to prior works on (1) online accessibility for visual media, (2) audio-visual consistency, and (3) video description generation, especially for people with vision impairments. We briefly review the literature in these domains.

2.1 Online Accessibility for Visual Media and Audio Description

For static media webpages and documents, alternative text or alt text is commonly used to add captions to images. Traditionally, alt texts have to be manually curated by the authors with no automation assistance. Recently, Gleason et al. developed various tools to help generate alt texts of images [22], memes [21], and GIFs [20]. Their efforts in making GIFs accessible is among the most relevant projects and highlight the importance of accurate audio description for moving pictures. While they focused on GIFs, we investigated longer videos with *audio* content and built an automatic workflow to generate and evaluate audio description.

Audio description, is "the art of turning what is seen into what is heard; the visual is communicated through the human voice and descriptive language" (2001)[3]. Access is increased through such

[3]https://acb.org/adp/

audio description, which adds precise, concise verbal descriptions of visual media-about people, objects, scenes, body language, facial expressions, and colors [31]. Many researches have been conducted to evaluate the benefits of audio description on different video types (*e.g.*, science programs [48], film [45], and live theater [31]), different audiences (*i.e.* sighted users [40] and visually impaired users [35]), and different description styles (*e.g.*, cinematic language style [17], standard and creative style [55]). However, as an emerging form of inter-modal translation, the generation of audio description raises many new questions, *i.e.* audio describing a video is not simply a matter of substituting visual images with verbal descriptions, but should establish the links within and across different modes of expression (*e.g.*, links between visual images and image-sound links) [9]. Apart from the sheer amount and types of videos, another difficulty is that the online videos often adopt a variety of editing techniques. The effects carried by the edits need to be illustrated by the audio descriptions to support the BVI users in the creation of a coherent understanding of the video content [9]. Therefore, descriptions for online videos are typically authored by professionals manually rather than generated with automatic techniques.

Instead of generating audio descriptions from scratch, there have been promising research projects exploring the alternatives. Salisbury et al. proposed a real-time crowd-sourcing experience which brought human in the loop to compensate for the shortcomings of existing AI algorithms [46]. Guinness et al. leveraged reverse image search to collect similar images that have been annotated [23]. For videos uploaded by users, except memes and viral videos, we argue that the sheer amount and diversity of personal videos makes it impractical to leverage crowd-sourcing, reusing existing videos, or any human-in-the-loop process. For example, every minute 500 hours of videos are uploaded to YouTube[4], with varying lengths, topics, subjects, stories, etc. We believe a fully automatic AD generation tool can help blind and visually-impaired users to access online videos broadly. Therefore, we aim to build an early prototype of such a tool and evaluate its pros and cons so as to share recommendations for future development.

Videos that are accessible for people who are blind or visually-impaired are scarce. For example, Netflix was criticized for not making all the accessibility features available on its online streaming platform [27]. Since then, Netflix has been progressively rolling out captions and audio descriptions on streaming video [13]. Rohrbach et al. introduced a dataset for movies with AD and benchmarked video description methods in an evaluation [44]. Their work only looked at movies with scripts and their dataset includes short snippets of videos. In comparison, we focus on user-uploaded videos that have no scripts and could be much shorter or longer in length with varying recording quality.

2.2 Audio-Visual Consistency

Observing a tremendous amount of visual-audio combination examples, we learn the correlations between visual and audio unconsciously [18]. To know a sound, or to understand a sonic event, we have an urge to locate a sound source. That is why we may turn around upon hearing a sound behind us to find out what we have heard, or to confirm what we imagine has taken place [10, 49]. Therefore, we experience listening to events rather than sounds [18]. For visually-impaired people, it is important to train their listening skills to enhance vision-to-sound mapping early in their life, since sound is a major source of information [8, 19]. Such training helps them with general learning, understanding their surroundings, as well as obtaining critical safety information needed for travel [15].

In computer vision, audio-visual related research explores algorithms for achieving audiovisual understanding [4, 6, 24] akin to training visually-impaired people to understand their environments from audio. Researchers investigate different techniques along this direction, such as detecting audio-visual events [4], improving human speech consistency [34, 57], and generating scene-aware sound [25, 58]. Drawing inspiration from them, we developed an audiovisual inconsistency detection algorithm based on the audio and visual embeddings. The detected inconsistencies indicate the timestamps for inserting AD results.

2.3 Video Description Generation

In computer vision, video description generation refers to the automatic generation of natural language sentences that describe the visual content of a video [3]. Similar to image captioning, it requires not only recognizing salient objects, but also understanding actions and interactions [26]. Research on dense video captioning, such as C3D [53] and I3D [11], showed that multiple features can improve video captioning models [37, 38, 54]. Generating a comprehensive description for videos that contain multi-events happening simultaneously is an open problem.

The state-of-the-art methods tackle the problem of video description generation by *dense video captioning*, aiming to locate all events in time and adding captions for each event [28, 56]. A standard dense video captioning algorithm divides the problem into two sub-tasks: event detection and caption generation. An event proposal network finds a set of candidate proposals and a captioning network generates a caption for each proposal independently [29, 60]. While video description research is promising, its focus is not on filling the audiovisual gap to create accessible videos. By evaluating the existing algorithms, we want to identify the shortcoming and hence future avenues for AD research.

We apply similarity-based modeling to optimize the output from dense video captioning engines. Sentence similarity and perplexity modeling lie at the core of many natural language processing applications [41]. Starting from word embeddings [7, 32, 59], recent popular neural network methods evolved to use sentence-level embedding for distance metrics [5, 30, 36]. Complementary aspects of a sentence, *e.g.*, syntax, length, and word order, proved to improve the ability to measure semantic textual similarity [52]. Based on the existing findings, we optimize the generated description in the sentence level. In particular, to enhance user experiences of video with description, several factors need to be considered in the optimization process. For instance, the description should be relevant to the video topic; it should not be redundant and confusing.

3 AUTOMATIC AD SYSTEM

Our automatic audio description (AD) system consists of three modules, *i.e.* insertion time prediction, audio description generation, and

[4]https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/

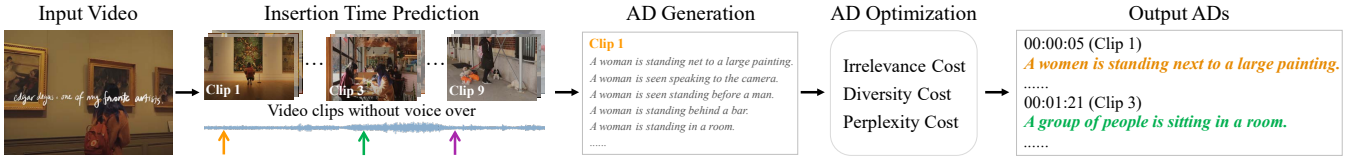


Figure 2: Overview of our proposed audio description system, *Tiresias*. It consists of three major modules: *Insertion Time Prediction*, *Audio Description Generation*, and *Audio Description Optimization*.

sentence-level audio description optimization, for making videos accessible. We describe our system in the following.

3.1 Insertion Time Prediction

As demonstrated in [18], humans are able to perceive events and environments based on sounds. For example, people can easily identify a noisy street or a pleasant seaside from sounds. Inspired by this, our system is designed to first segment the input video as two kinds of clips, *i.e.* audio-visual consistent clips and inconsistent clips. The inconsistent video clips will be fed into the description generation module.

Keyframe Extraction. We preprocess the input video into clips that do not have voice overs, since the final generated descriptions cannot overlap with these clips in audio track, otherwise it would be difficult to distinguish speech content and understanding the video content [16]. For the rest video clips, we further segment the clips using a robust keyframe extraction tool called Katna [1]. The keyframes are defined as the representative frames of a video stream, which provide an accurate and compact summary of the video content considering the changes in the scene setting, lighting, and human actions. All frames between two consequent keyframes constitute a video clip.

Audio-Visual Consistency. To determine whether a pair of a video frame and the audio clip correspond to each other or not, we use the proposed deep network [4] to train the two visual and audio networks from scratch. The learning for both visual and audio semantic information is performed in a completely unsupervised manner by simply watching and listening to a large number of unlabelled videos. The network structure consists of three distinct parts: the visual and the audio sub-networks which respectively extract visual and audio features, and the fusion network which takes these features into account to produce the final decision of consistency. This method achieves a consistency classification accuracy of 78% on Flickr-SoundNet Dataset and of 74% on Kinetics-Sounds Dataset. Please refer to [4] for more details. The inconsistent video clips will be fed into our description generation module.

3.2 Audio Description Generation

To generate high-quality descriptions for all events of interest of a video, the fine-grained video clips were fed into an attention-based model [12], which consists of a sentence localizer and a caption generator.

The sentence localizer provides a cross-attention multi-model feature fusion framework, which models the correspondence between the video clip and the caption. The involved attention mechanism contains two sub-attention computations, *i.e.* the attention between the final hidden state of the clip and the caption feature at each time step, and the attention between the final hidden state of the caption and the video features. For training, the video and event descriptions are available. The video and one of its event descriptions are fed into the localizer. A temporal segment prediction is obtained, which is fed into the generator to perform captioning on the frames soft clipped from each of the extracted video clips.

The dual network was trained on the ActivityNet Captions dataset [28] that has been applied as the benchmark for dense video captioning. This dataset contains 20,000 videos in total, covering a wide range of complex human activities. For each video, the temporal segment and caption sentence of each human event is annotated. On average, there are 3.65 events annotated for each video, resulting in a total of 100,000 events. We evaluated the description generation model with 5,000 video clips randomly selected from the ActivityNet Caption dataset and the YouCookII dataset. With the evaluation metrics of BLEU and METEOR, *i.e.* evaluating the quality of the generated text which has been machine-translated (the higher the score, the closer to human judgment), we achieve higher scores (9.56 for METEOR and 10.12 for BLEU) against other methods [28]. Note that the scores range from $[0, 1]$ but are scaled by 100 for visualization, following standard practices.

3.3 Audio Description Optimization

Descriptions that are redundant, confusing, grammatically wrong, or incompatible with the video topic have been shown to interrupt one's immersive video experience. The corresponding quality standards were drafted by the Advanced Diagnostic Imaging Audio Description Guidelines Committee in 2003 and serve as a useful set of guidelines for video describers [2]. Since several descriptions are generated for each video clip by the previous module, next we optimize the results based on the audio description guidelines to yield accurate and fluent descriptions as the final output.

To enable accurate description selection from multiple generated results, we formulate the process as optimization with several costs, considering relevancy to the video topic, diversity, and perplexity.

Formulation. We formulate our optimization problem on the sentence level. Let $S_i(n) = (s_1, s_2, ..., s_n)$ denote the generated descriptions for the i-th video clip. The final output descriptions S for the whole video is achieved by minimizing the cost function:

$C_{total}(S) = \sum_{i=1}^{N} (C_i(S_i) + \lambda_d C_d(S_i) + \lambda_p C_p(S_i))$, where N is the number of segmented video clips. $C_i(\cdot)$ is the irrelevance cost term that penalizes the descriptions irrelevant to the video topic. $C_d(\cdot)$ is the diversity cost term for evaluating the description content. $C_p(\cdot)$ is the perplexity cost term that penalizes the confusing descriptions. λ_d and λ_p are the weights, which are empirically set as 0.4 and 0.2 by default, to balance the cost terms. We leverage pre-trained sentence-level embeddings in our optimization task, which has demonstrated good accuracy in diverse transfer tasks.

Irrelevance Cost and Diversity Cost. When considering the generated candidate descriptions S_i for video clip i, the irrelevance cost term is considered for each description s_j independently from others. We used Google's TensorFlow Hub model, *i.e.* the Universal Sentence Encoder, to measure the relevance of descriptions to the video topic (determined by the video title). The higher the relevance is, the lower the irrelevance cost $C_i(\cdot)$. The encoder produces sentence-level embeddings for the texts, *i.e.* encodes the text into high-dimensional vectors, which are approximately normalized. Therefore, the semantic correlation of two texts can be trivially computed as the inner product of the encoded vectors. We use the same strategy to measure the diversity between two consecutive descriptions. The higher the diversity, the lower the diversity cost $C_d(\cdot)$ is. We aim to maximize the description diversity.

Perplexity Cost. To filter out uncommon expressions, linguistic errors, and not fluent sentences in the generated results, we use GPT [42] as a language model to assign a language modeling score by measuring how well a probability model predicts a sample. The score can also be regarded as the perplexity score of the sentence. The higher the score is, the lower the perplexity cost $C_p(\cdot)$.

Optimization. We apply a dynamic programming method to efficiently optimize the total cost function. Let $\langle i, S_i \rangle$ denote the current video clip i and the corresponding generated descriptions. Let $C_{total}(S_i)$ demote the optimal cost for this state (for video clip i). When choosing the next description, the current state can transition to a state in the next video clip $i+1$. The next state would be $\langle i+1, S_{i+1} \rangle$ where S_{i+1} is the new generated descriptions for the next video clip $i+1$. The optimal solution can be obtained by solving $S^* = \arg\min_{\langle S_i \rangle} C_{total}(S)$ using dynamic programming and back-tracking the descriptions generated for each video clip.

3.4 Implementation Details

We implemented our approach on an Intel Core i7-9700 machine equipped with an NVIDIA GeForce RTX GPU with 24GB graphics card memory. Our system, including *Insertion Time Prediction*, *Audio Description Generation*, and *Audio Description Optimization*, took about 45 seconds to process a two-minute video. After generating the optimized descriptions for the whole video, we applied Google Cloud Text-to-Speech to convert them into speeches and used FFmpeg to insert the speeches into the original video, which took about 200 milliseconds for each description.

To validate the robustness, since *Tiresias* works in a fully automatic manner, we ran it on 500 videos and shared it as an *AutoAudioDescription* dataset. The videos were sampled from the ActivityNet Captions Dataset [28]. Overall, it took about 420 hours for our system to generate audio descriptions to all the 500 YouTube videos totaling around 750 hours, which would cost about $150 if we ran our system through a typical cloud service like AWS.

4 EVALUATION

With an IRB approval, we recruited 20 sighted participants and 12 blind or visually-impaired participants. The 20 sighted participants (p01-p20) reported normal or corrected-to-normal vision, no color-blindness, and normal hearing. These sighted participants were equally split between men and women, and they averaged 32.10 years old (min = 19, max = 57). The 12 blind or visually-impaired participants (P21-P32), as shown in Table 1, are diverse in terms of gender, age, and occupation. These participants have a range of visual impairments as well as varied video streaming platforms with online videos. Four BVI participants (P21, P26, P29, and P31) did not complete the study due to technical difficulties. Their responses are included in the interviews, but not in the quantitative evaluation. Please refer to supplementary materials for sighted participant demographics.

4.1 Study Procedures

Due to COVID-19, we conducted all the interviews and studies virtually via Zoom. We chose 30 videos from the *ActivityNet Captions* dataset covering 5 different video types, *i.e.* music, film, pets and animals, vlog (further divided into people and activities, comedy, and DIY), and physical (further divided into sports, dancing, and fitness). The length of each video ranges from 30 seconds to 3 minutes. Each video has two sets of corresponding data: (i) *Video Set*, *i.e.* original video and the same video with automatic audio descriptions, for sighted users to evaluate; and (ii) *Audio Set*, *i.e.* the original audio track and the same audio with automatic audio descriptions, for both sighted and blind users. The only difference between the *Video Set* and the *Audio Set* is whether visual frames are available.

We designed and implemented a program (illustrated in supplementary materials). We ran our study in two passes. In particular, for BVI participants in the first pass, we asked the participants to watch/listen to the original input video/audio. We asked them to raise their hand when they needed *additional information*; in other words, when they felt that the video did not have sufficient information to help them understand the scene. In the second pass, we asked the participants to watch/listen to the same video/audio plus our automatically-generated audio descriptions. Just like in the first pass, the participants raised their hands for *additional information* that our AD output does not capture. In addition, we also ask them to raise their hands and let us know if the generated AD is *confusing*, *redundant*, or *grammatically problematic*. For the sighted participants, we sent them the executable program and instructed them to perform the same operations as BVI participants by clicking different buttons, *e.g.*, clicking "A" for *Additional information*, etc.

We randomly selected 3 samples from *Video Set* and *Audio Set* and categorized the results into 3 categories.

- SV for the sighted participants who watched the *Video Set* samples.
- SA for the sighted participants who listened to the *Audio Set* samples.

Table 1: Demographics of user study participants. "Vis. Exp." refers to the visual experience, including "CTB" (Congenital Total Blindness), "CB-LC" (Congenital Blindness with some light/color perception), "ATB" (Acquired Total Blindness), "AB-LC" (Acquired Blindness with some light/color perception), and "CLV" (Congenital Low-Vision).

ID	Gender	Age	Occupation	Vis. Exp.	Video Platforms	Video Types
P21	F	34	Unemployed	ATB	YouTube, Netflix	Music, fitness, series
P22	M	75	Retired	CB-LC	YouTube, Netflix, TV	Movie
P23	F	69	Retired	CTB	Netflix, Amazon Prime, Hulu	Documentary, movie, sports
P24	M	29	Teacher	CLV	TV, Netflix	Documentary, sports
P25	F	65	Manager	CB-LC	TV	Movie
P26	M	51	Student	CTB	YouTube	News, sports
P27	F	37	Office Assistant	ATB	Serotek	Movie, TV shows
P28	F	16	Student	CLV	YouTube	Education videos
P29	F	17	Student	CLV	YouTube, Netflix	Cartoon
P30	F	60	Retired	CLV	YouTube,Amazon Prime, TV	Movie, news, documentary
P31	M	59	Retired	CLV	YouTube, TV	Movie, sports
P32	F	49	Retired	AB-LC	JW.org, YouTube, TV	Drama

- BVI for the blind or visually-impaired participants who listened to the *Audio Set* samples. Based on the following pre-study interviews, we choose the lowest common denominator to only show the samples from *Audio Set* to all visually impaired participants:
  - Similar to blind people, low-vision participants (P24, P28-P31) have difficulties manipulating computers without others' accompany. Therefore, they usually used phones (with a small screen) to listen to videos, as P31 stated, *"The screen reader on phone is easier to use than on the computer."*
  - Our recruited adult low-vision participants (P24, P30, P31) stated that they only listen to news, sports, and documentaries, which are fully described by the announcers. For example, P30 stated, *"If the videos come up like on a news web page, they're kind of embedded into the story and the presenter would describe the content in detail."*
  - Some low-vision participants (P24 and P28) pointed out that they would watch a video on TV or a bigger computer monitor if they want to learn something from the video. However, they indeed need others' help to get the details, especially for students.

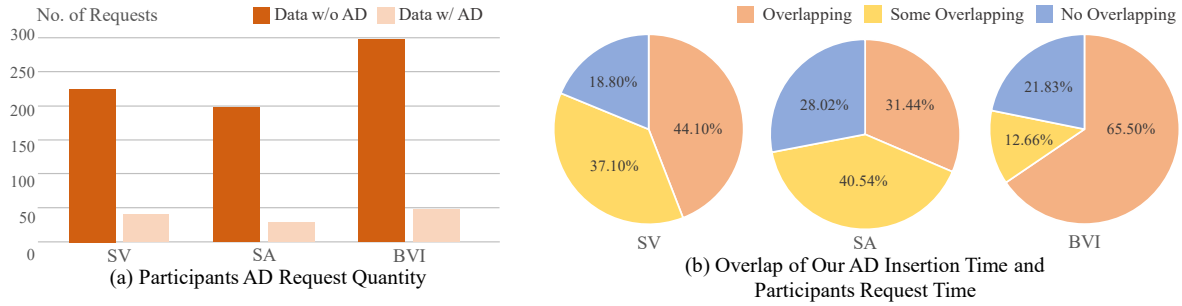
We included the *Video Set* for the sighted participants to investigate if there is a difference introduced by the videos. We asked the BVI participants to communicate by raising their hands so that they could bypass the difficulty of operating mobile phones or computers.

4.2 Quantitative Results and Analysis

We analyzed the results from our user study guided by the following research questions: Did the participants find the automatic AD results helpful? If so, in which specific ways our automatic AD is helpful for people who are blind or visually-impaired? On the other hand, what are the shortcomings of our implementation? What are the similarities and differences between the sighted and blind or visually-impaired participants? Figure 3 and Figure 4 visualize the aggregated statistics in SV, SA, and BVI categories. Next we discuss 4 observations from the charts.

Automatic audio descriptions may reduce the requests for additional information. In the first pass, we asked the participants to indicate requests for additional information while watching/listening to the original input. Then in the second pass, we asked the participants to do the same but updated the input to include our automatic AD output. Figure 3(a) shows a significant drop among all 3 cohorts, regardless of whether the participant was sighted or blind or if the sighted participants saw the visual frame or not. In particular, with our generated audio descriptions, the demand for additional descriptions is less than 20% of that without our AD. As many participants reported, *"some insertion position marked before could be omitted because I could exactly know what happened in the video through the descriptions."* However, the familiarity of the video content after the first pass would implicitly affect the users' perception of the video content in the second pass, *i.e.* users may have a general understanding of the video content based on some sound details in the first pass, while they may pay more attention to other aspects of video content in the second pass that they did not think deeply in the first pass. For example, as P24 raised his hand twice and stated in the first pass that *"It does like that a shopping cart seems to be rolling on through the beginning of the video. I want to know whether the people near a grocery store."* After the second pass, P24 changed his initial questions and raised his hand once, *"I know people will take the shopping cart out of the parking lot or out of the store. So I want to know either there's a random shopping cart, or they're looking at a store."*

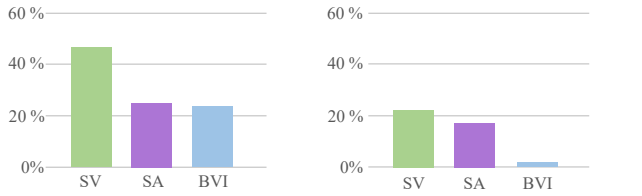
65.50% of the requests from the blind participants overlapped with our automatic AD.. When a participant indicated requests for additional information, we also recorded the timestamp. In our automatically generated AD, we also extracted the insertion times. We show further that the insertion time differences between the participants' indications and our results are significant, as shown in Fig. 3(b). We consider the results within a 2 seconds time difference as overlapping, as users stated that there might be a slight reaction delay. When the time difference is between 2-4 seconds, we consider it as a partial overlapping. If the time difference is more than 4 seconds, we regard it as no overlapping. Under this metric,



(a) Participants AD Request Quantity

(b) Overlap of Our AD Insertion Time and Participants Request Time

Figure 3: Quantitative results of AD insertion request. (a) shows the sharp drop in requests with our automatic audio descriptions in all three SV, SA, and BVI groups. (b) shows the percentage of AD insertion position overlap between our results and users' requests, indicating the effectiveness of our insertion time prediction module.



(a) AD Confusion

(b) AD Redundancy & Grammar Errors

Figure 4: (a) The percentage of reported AD confusion among SV, SA, and BVI groups. We see a relatively consistent result from SA and BVI, which both only have access to the audio stream. The SV group which also has access to the visual stream shows a significantly higher confusion rate. We attribute this to an inaccurate depiction of visual activities. (b) The percentage of reported AD redundancy and grammar errors. The BVI group reported significantly lower numbers than the sighted group. From interviews, we realize most blind viewers do not mind hearing repeated descriptions, whereas sighted viewers behave the opposite way.

the overlapping of the sighted participants' results achieved 44.10% (SV) and 31.33% (SA). The overlapping is even higher at 65.50% for the BVI category. In detail, for low-vision participants and other visually impaired participants, the results of overlapping achieved 53.48% and 68.27% respectively, which both higher than the SV and SA overlapping results.

In order to further explore the impact of different reaction delays (i.e. the interval setting) on the analysis results, we analyzed the results with different intervals. Compared with the result of 2s-interval, the result of 1s-interval or >=3s-interval has substantial fluctuations. For example, the results "no overlapping" of SA participants would greatly increase from 31.44% to 88.50% with 1s-interval, or reduce from 31.44% to 12.39% with 3s-interval. Therefore, we mainly show the analysis results with 2 second intervals. Overall, our AD insertions that have good or partial overlaps are above 70% for SA and SV; and 79% insertion times are close to the BVI participants' demands.

The generated AD might not convey an accurate depiction of the visual activities. Figure 4(a) shows that the SV group (sighted users who watched the *Video Set*) indicated most confusion about the descriptions. Almost half of the results (46.72%) were labeled as confusing. On SA and BVI, the confusion percentage is much lower at 25.56% and 24.62% (25.58% for low-vision participants and 24.19% for other visually impaired participants), respectively. We wondered what had caused the much higher confusion rate in the SV group. Some SV participants reported they had received both visual and auditory information, so they could easily see and hear the confusing audio descriptions. For example, a barrel was incorrectly described as a dog. On the other hand, when the SA or BVI participants encountered similar scenarios, they were usually more "curious" than confused about the description since they were much less confident of what was actually happening in the video.

Blind participants are generally okay with redundant and repeated AD, while sighted participants complained about excessive redundancy. There is a stark contrast between the reported redundancy of the sighted participants and that of the BVI participants. As shown in Figure 4(b), around 20% of the descriptions were labeled by the sighted participants as redundant results (24.00% for the *Video Set* and 18.46% for the *Audio Set*), while only 1.01% by the BVI participants. When we asked the BVI participants about the 1.01%, we found that those were related to grammar inconsistency from the AD generation NLP model rather than redundancy, further reducing the 1.01% to zero. This 20% versus 0% difference is intriguing. Take a dancing video as an example. A sighted participant noted that she would only like to know the special movements of the dancers rather than the repeated descriptions saying 'they are dancing in a circle'. She said that if the dancers were doing the same movement, there was no need to describe it. However, almost all BVI participants (both low-vision and other visually impaired participants) shared the opposite opinion by giving a comment like *"I didn't mind the redundant descriptions because I could picture what the descriptions were saying."*

4.3 Post-Study Interview Findings

We interviewed each participant with open-ended questions to further analyze the results. We list the questions in the Appendix. We group our findings into 3 topics.

What is missing in our audio descriptions? Overall, both sighted and BVI participants stated that whenever they could not get access

Table 2: Results of cross-type analysis, where "✓" specifies the type of description content wanted. (M=Music, AN=animals, F=Film, AC=Activity, C=Comedy, D=DIY, S=Sports, D=Dancing, F=Fitness)

Video Content	M	AN	F	AC	C	D	S	D	F
Event/Scene									
People Present	✓	✓	✓	✓	✓		✓	✓	✓
Text	✓		✓	✓		✓	✓		✓
Interaction		✓	✓	✓	✓		✓	✓	✓
Place	✓	✓	✓	✓	✓		✓	✓	✓
Weather		✓	✓	✓	✓				
People									
Quantity	✓		✓	✓	✓		✓	✓	
Position			✓	✓	✓		✓		
Props		✓	✓	✓	✓	✓			
Action/Reaction	✓	✓	✓	✓	✓	✓	✓	✓	✓
Relationship			✓	✓	✓		✓	✓	
Gender	✓	✓	✓	✓	✓	✓	✓	✓	✓
Race			✓	✓			✓	✓	
Name	✓		✓				✓	✓	✓
Identity/Role	✓		✓		✓		✓	✓	
Facial Expression	✓	✓	✓	✓	✓			✓	✓
Body Shape/Size			✓	✓	✓		✓	✓	✓
Hair Style/Color			✓		✓				
Eye Color			✓						
Unique Features			✓	✓	✓		✓	✓	
Other Factors									
Background Info	✓	✓					✓	✓	
Storytelling/Logical			✓	✓	✓	✓	✓	✓	
Camera Movement			✓		✓				

to visual content, they wanted to know everything that happened in the video and a simple yet comprehensive description of the video content. Many participants made a comment like *"I want to know more about the scene. Was it a room or was it outside? How many people were there? What were they doing? Were they teenagers or adults? What were they wearing? Were they next to each other? Whenever such a description would help paint the picture of what the person was doing, that would help."* For BVI participants, they wanted to get more detailed information about the scene or background, as P21 noted, *"Because the people were climbing the mountain, I would like to know more about the background, like the roads, trees, etc."* However, due to the lack of paired static audio-visual objects in the training dataset and the temporal features of image-audio data compared to video-audio data, in the audio-visual consistency model, our *Insertion Time Prediction* module may miss some necessary insertion positions.

Both sighted and BVI participants stated that they wanted to get deeper interpretations and subtle relationships beyond the audiovisual content. For example, the reason for people's reaction, as P22 said, *"I know somebody was playing drums. I heard people laughing, but I don't know why the people were laughing. The audio descriptions should be able to make me laugh on it at the same time as sighted people."* This is a challenge in existing video dense captioning research. Since the model was trained on the dataset with activity annotations only, it failed to explain the relationships in a scene.

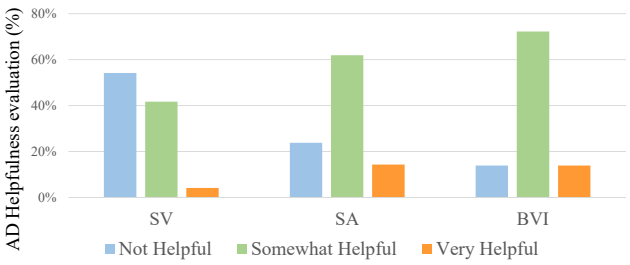


Figure 5: Audio description experience of participants.

Analyzing the feedback with respect to description preferences, we observed that BVI people generally have low engagement with videos uploaded to social media platforms. In some instances, the low level of engagement was related to their familiarity with different video types. For instance, for music videos, due to different levels of music theory understanding, some participants desired to know more about the background of a piece of music, as P11 and P24 expressed, *" who wrote the music or who composed it, why did the author or composer make the music?"*, while others were more concerned about the instruments. For sports video, low engagement stemmed not only from inadequate descriptions but also from different user preferences. P21 expressed that *"I just care who won the game"*, while P23 and P36 stated that *"I would like to know the details of the match. If someone did a move, like against another person, I would like to know his corresponding strategies."*

We also observed considerable diversity across sources in terms of participantsâĂŹ desires for the amount of content and the level of detail they want in audio descriptions. We offer a nuanced view of different content needs with different source video types in Table 2. For each type, we specify all kinds of content from a lengthy list of options that at least one of our participants thought was important to describe. We group these findings around three key themes that are commonly the central focuses of a video composition: event/scene, people, and other objects. Notably, for some video types, the amount of content desired in audio descriptions was greater than that of other video types. For instance, we noted that participants want to have the largest amount of content available to them for films, activity-related videos, whereas they requested a smaller amount of description content for animal and DIY videos.

We further observed that there are other factors that may impact the amount of content and/or the level of detail that is included in a description. For instance, we noted that the level of vision the participants have influences the amount of content they want. One sighted participant (P08) said *"I will ignore objects that are not related to the current event"*, while the BVI participants (P27 and P30) stated, *"I want to get more details, like how the ball was transferred from one player to another. "* P26 was more special, stating that he wanted to know more details of the players' clothes in sports videos, *i.e. "I want to know the color of the players' cloth, the brand names of their shoes, as well as the Polo."* Moreover, participants showed

different opinions towards the description density. P30 noted that *"The description should describe what happened when people were not talking, because maybe there was something flashing on the screen that somebody could just see their silence."* In contrast, we heard from P32 that the audio description should not take over the video content but *"should be simple and be said with a soft voice".*

Why is our audio description not very helpful? After the participants watched/listened to the video with our generated audio descriptions, we asked them about their overall experiences. Specifically, we asked them to rate how well our audio descriptions helped them understand the video content, by giving a rating of "Not Helpful", "Somewhat Helpful", or "Very Helpful". As shown in Fig. 5, most sighted participants who watched the *Video Set* claimed that *"It is easy to distinguish the wrong recognition result in the description"*, resulting in a high percentage of "Not Helpful" (54.17%) following with 41.67% of "Somewhat Helpful". For the 4.17% of responses with the rating "Very Helpful", the participants gave comments like *"The description describes the event precisely."*

23.81% of sighted participants and 13.33% of BVI participants who listened to the *Audio Set* rated the results as "Not Helpful". Based on their feedback, we believe there are three main reasons for their ratings, all of which result from the shortcomings of existing video understanding methods:

- *Audio descriptions do not match the video topic or title.* It is uncommon for online videos to show only human activities. Many videos are remixed with a landscape video montage, food sharing, or shopping. Such videos are sometimes titled with irrelevant and concise words, making it difficult to determine the relevance between the description and the video content at the sentence level. For instance, P29, a low-vision participant, noted, *"There were a lot of people sitting in a bar I guess because the description included the words piano and beer, but the video title was about washing the car. I don't really get that anybody was actually washing the car in the video."*
- *Audio descriptions do not match the sound.* For instance, the description does not match the character's attributes in the video, *e.g.*, gender, as both sighted and BVI participants noted, *"The description described a woman who was making a sandwich but I actually heard a man's voice."*
- *Audio descriptions are not logical.* For instance, P23 and P28 noted, *"The description said they are climbing up on the sidewalk and this does not make any sense to me. I think a sidewalk is something you just walk on, not climb."* Similarly, P32 stated, *"I'm not confused about climbing on the sidewalk. I would like to know why the sidewalk was difficult to climb? What were they using to climb the sidewalk?"*

A larger percentage of BVI participants, *i.e.* 76.67% (71.43% for low-vision participants and 77.42% for others, versus 61.90% for sighted participants) rated the descriptions as "Somewhat Helpful". This is probably because the BVI participants were used to changing their initial thoughts based on the descriptions. As a low-vision participant (P24) stated that *"I thought it was one badminton court. But now I think there were multiple courts, like multiple games going on at once. Because I heard the description saying the words of tennis and horse."* and a blind participant (P32) expressed *"I thought the man was climbing a mountain. I'm thinking the man is climbing up on a car because the description mentioned the car."* In comparison, the sighted participants usually doubted the content, thus more likely giving the "Not Helpful" rating.

Although most of the audio descriptions were rated as "Somewhat Helpful", we observed a variety of concerns on whether an audio description of an event lacked useful information. Overall, almost every participant stated that the provided audio descriptions did not fully meet their expectations, *i.e.* , they did not get all the information they wanted as shown in Table 2. In general, the description only described the activities that occurred in the video, but lacked more detailed information about the scene and people, for instance, scene attributes (weather, etc.) and personal attributes (name, gender, race, facial expression, etc.), which were stated as important by both the sighted and BVI participants. Therefore, the results rated as "Very Helpful" represent a small portion for both the sighted and BVI participants, 14.29% and 16, 57%, respectively.

How do our automatic AD results compare with a manual AD from Netflix? We selected 5 Netflix video clips[5] with audio description and conducted a smaller-scale study following the same procedure described in § 4.1. To the best of our knowledge, the audio descriptions from Netflix were manually created. The quantity of descriptions is 1.5 times of what the users requested. The overlaps between participants' requests and manual results in terms of insertion time are 66.67% for SV, 70.00% for SA, and 88.89% for the BVI participants, noticeably higher than those of our automatic algorithm at 44.10% for SV, 31.44% for SA, and 65.50% for the BVI participants. Regarding the description quality, very few (only two) descriptions are thought to be confusing. As P23 stated, *"This description is good but is ahead of the action."*. Such issues can be explained by the intention of preventing the description from colliding with the background sound in the video. Moreover, the manually generated descriptions rarely have redundant descriptions or grammar errors, thus no participants reported such problems.

Our findings underscore the types of description content that may be desired universally across different video types (Table 2). For instance, our participants consistently wanted to learn about actions cross all video types. This aligns with the content provided by the manual generated description. Almost every participant stated that they enjoyed the video with such descriptions and found that the video content was more accessible. Therefore, extending prior findings, our work also reveals the consistency between the BVI participants' expectations and the detailed information provided by manual generation.

5 RECOMMENDATION FOR NEXT-GENERATION AD SERVICES

Participants in our user study expressed that our system offered more video types compared to what they typically watch online, which was limited to movies and other curated content. Our findings offer convincing evidence that existing video understanding research and audio description implementations are inadequate to address the diverse demands from BVI viewers. Now we discuss

[5]Project Power, Extraction, Step Sisters, Rotten-The Avocado War, The Last Dance

recommendations and future directions for better audio description (AD) generation and interaction model.

We need a more context-aware prediction model to decide when to insert an audio description. Currently, we use the audiovisual inconsistency to predict if a visual event is inaudible and it works in some simple cases. However, we find that consistent audiovisual events may also require descriptions sometimes so that users can further understand the video content. For example, as P22 stated, *"I hear a dog is barking but I don't know why, is the dog afraid of the person or is the dog just playing?"* In this example, the auditory and visual data streams are consistent, yet the viewer still wanted additional information on the broader context. This is relevant to the training algorithms and the corresponding dataset, which were designed on the objective factors rather than a deep understanding of visual content. Alternatively, descriptions of reaction reasoning could be used to redesign the evaluation metrics, when authoring audio descriptions that are used for training AI models, and to support the inclusion of relevant details depending on video context.

We need smarter video understanding beyond just activity recognition. For the generation of audio descriptions, we used dense video captioning, which was trained on activity-focused datasets. A consequence was that our automatic descriptions only covered a small set of those listed in Table 2. Out of all the entries in the first column, our activity-focused model performed well on "Interaction" and "Action/Reaction" and ignored most other entries. For example, rarely did the descriptions include weather or place. The fine details of the people appearing in the scene were also missing in the video captions, such as gender, name, facial expression, body size, etc. A better AD system would benefit from a more robust video understanding module that can correctly predict all the aspects that viewers typically care about.

We need interactive levels of details for audio descriptions. Audio descriptions have been defined as a static track that one makes a binary choice, enable or disable. We envision a next-generation AD system with multiple levels of details. This has several implications.

(1) Viewers can fluently switch the level of audio description they desired. In most verbose level, probably it has all the fine details of everything on the screen, which might be overwhelming for some blind viewers. Therefore, they can fine-tune the level of details they want and find the perfect length for their consumption and engagement with the video.
(2) We will need an algorithm to prioritize all the descriptions and rank them accordingly. This is impossible without a good semantic understanding of the audiovisual events and the relationships among all the subjects.
(3) Audio descriptions, like all accessibility features, benefit many users. For example, a sighted user might be using a smartphone while listening to a video playing on TV. In this case, a minimal level of audio descriptions can notify the viewer about scene transitions or other significant updates.

We need more accurate sentence-level scene descriptions. We find that some output sentences appeared unreasonable and illogical, which was caused by inaccurate sentence-level description. For instance, P23 stated that *"If a person is playing the harmonica, how can he smile to the camera?"*. When we review this footage, we realize that the musician was indeed playing harmonica and the smiling came from a misclassification. In another instance, the audio descriptions outputted "He uses a lawn to cut the leaves", which is against common sense. We believe there is room for improvement in scene understanding, natural language processing, and in particular, generating accurate and contextual sentence-level descriptions.

6 CONCLUSION

The lack of accessible videos on social media platforms and video streaming platforms is a major barrier for participation by people with visual impairments. Our automatic audio description generator attempts to integrate promising methods for generating audio descriptions into one tool that users can use for different types of video. We conducted user studies with both sighted and blind participants and analyzed the feedback on the automatically generated audio descriptions. With our generated audio descriptions, the demand for additional descriptions reduced to less than 20% of the original input. More than 85% of BVI participants reported our system was at least "somewhat helpful" to their video experience. Despite the promising statistics, we also see that there is a big gap between the automatic results and the manually generated AD by Netflix. We further discussed several recommendations on how we should improve our automatic workflow. In particular, we believe the next-generation audio description would have multiple levels of details and would react to the viewer's interactive requests for more or fewer descriptions. In the interviews, the BVI participants repeatedly expressed that inaccessibility was their primary concern and they often had to find workarounds for videos without audio descriptions, *e.g.*, asking a friend or family, which affected their independence. Making all user-generated online videos accessible is a challenging and important long-term goal. Our work described the pros and cons of existing tools and sheds light on future endeavors.

Limitations. Due to COVID-19, we conducted our studies remotely via Zoom. In our pre-interview questionnaire, most of our 12 participants preferred Zoom's mobile app (with small screen size) or phone dial-in with no visuals at all, over Zoom's desktop app (with larger screen size). To accommodate their requests and offer the same calibrated experience, we chose the lowest common denominator to only show the audio sets to all participants with visual impairments, which is a limitation that the participants have low visions could be tested on the Video Set data by using their residual vision. It would be helpful to recruit more low-vision participants who usually watch videos and conduct experiments to explore the differences of accessing videos between them and blind people.

We conduct the study in a two-pass manner to explore whether the generated audio descriptions help participants to access the video content. However, the familiarity of the video content after the first pass would implicitly affect the participants' perception of the video content in the second pass, *e.g.*, they may pay more attention to other aspects of the video content. We are interested in designing a more comprehensive process, for example, recruiting more participants and separating the quantitative evaluations and subjective interviews.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable input, study participants for their involvement. Yongqi Zhang was supported by an NSF Graduate Research Fellowship. Lap-Fai Yu was supported by an NSF CAREER Award (award number: 1942531).

REFERENCES

[1] 2019. Katna: Tool for automating common vide keyframe extraction and Image Autocrop tasks. https://katna.readthedocs.io/.
[2] 2020. Guidelines for Audio Describers. http://www.acb.org/adp/guidelines.html.
[3] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. 2019. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys (CSUR)* 52, 6 (2019), 1–37.
[4] Relja Arandjelovic and Andrew Zisserman. 2017. Look, listen and learn. In *ICCV*. 609–617.
[5] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. (2016).
[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. 2016. Soundnet: Learning sound representations from unlabeled video. In *Advances in neural information processing systems*. 892–900.
[7] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research* 3, Feb (2003), 1137–1155.
[8] Ann Bigelow. 1991. Spatial mapping of familiar locations in blind children. *Journal of Visual Impairment & Blindness* 85, 3 (1991), 113–117.
[9] Sabine Braun. 2011. Creating coherence in audio description. *Meta: Journal des traducteurs/Meta: TranslatorsâĂŹ Journal* 56, 3 (2011), 645–662.
[10] Richard Brown. 2013. *Consciousness inside and out: Phenomenology, neuroscience, and the nature of experience*. Springer Science & Business Media.
[11] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.
[12] Xuguang Duan, Wenbing Huang, Chuang Gan, Jingdong Wang, Wenwu Zhu, and Junzhou Huang. 2018. Weakly supervised dense event captioning in videos. In *Advances in Neural Information Processing Systems*. 3059–3069.
[13] Katie Ellis et al. 2015. Netflix closed captions offer an accessible model for the streaming video industry, but what about audio description? *Communication, Politics & Culture* 47, 3 (2015), 3.
[14] Deborah I Fels, John Patrick Udo, Jonas E Diamond, and Jeremy I Diamond. 2006. A comparison of alternative narrative approaches to video description for animated comedy. *Journal of Visual Impairment & Blindness* 100, 5 (2006), 295–305.
[15] Pat Fletcher. 2002. Seeing with sound: A journey into sight. *Retrieved September* 21 (2002), 2015.
[16] Louise Fryer. 2016. *An introduction to audio description: A practical guide*. Routledge.
[17] Louise Fryer and Jonathan Freeman. 2013. Cinematic language and the description of film: Keeping AD users in the frame. *Perspectives* 21, 3 (2013), 412–426.
[18] William W Gaver. 1993. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology* 5, 1 (1993), 1–29.
[19] Nicholas A Giudice and Gordon E Legge. 2008. Blind navigation and the role of technology. *The engineering handbook of smart technology for aging, disability, and independence* 8 (2008), 479–500.
[20] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris M Kitani, and Jefrey P Bigham. 2020. Making GIFs Accessible. (2020).
[21] Cole Gleason, Amy Pavel, Xingyu Liu, Patrick Carrington, Lydia B Chilton, and Jeffrey P Bigham. 2019. Making Memes Accessible. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 367–376.
[22] 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.
[23] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. 2018. Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–11.
[24] Di Hu, Feiping Nie, and Xuelong Li. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 9248–9257.
[25] Haikun Huang, Michael Solah, Dingzeyu Li, and Lap-Fai Yu. 2019. Audible panorama: Automatic spatial audio generation for panorama imagery. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.
[26] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*. 4634–4643.
[27] R Kingett. 2014. The Accessible Netflix Project Advocates Taking Steps to Ensure Netflix Accessibility for Everyone. *The Accessible Netflix Project* 26 (2014).
[28] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.
[29] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2018. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7492–7500.
[30] Lajanugen Logeswaran and Honglak Lee. 2018. An Efficient Framework for Learning Sentence Representations. In *International Conference on Learning Representations*. 1–16.
[31] John Miers. 1995. Audio description–seeing theater with your ears. *Information Technology and Disabilities* 2, 2 (1995).
[32] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
[33] Chris Mikul. 2010. Audio description background paper. *Ultimo NSW: Media Access Australia* (2010).
[34] Tae-Hyun Oh, Tali Dekel, Changil Kim, Inbar Mosseri, William T Freeman, Michael Rubinstein, and Wojciech Matusik. 2019. Speech2face: Learning the face behind a voice. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7539–7548.
[35] Jaclyn Packer, Katie Vizenor, and Joshua A Miele. 2015. An overview of video description: history, benefits, and guidelines. *Journal of Visual Impairment & Blindness* 109, 2 (2015), 83–93.
[36] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2018. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features. In *NAACL 2018 - Conference of the North American Chapter of the Association for Computational Linguistics*.
[37] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4594–4602.
[38] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. 2017. Video captioning with transferred semantic attributes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6504–6512.
[39] Becky Parton. 2016. Video captions for online courses: Do YouTubeâĂŹs auto-generated captions meet deaf studentsâĂŹ needs? *Journal of Open, Flexible, and Distance Learning* 20, 1 (2016), 8–18.
[40] Elisa Perego. 2016. Gains and losses of watching audio described films for sighted viewers. *Target. International Journal of Translation Studies* 28, 3 (2016), 424–444.
[41] Zhe Quan, Zhi-Jie Wang, Yuquan Le, Bin Yao, Kenli Li, and Jian Yin. 2019. An efficient framework for sentence similarity modeling. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 4 (2019), 853–865.
[42] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
[43] Aline Remael and Gert Vercauteren. 2011. Basisprincipes voor audiobeschrijving voor televisie en film [Basics of audio description for television and film]. *Antwerp: Departement Vertalers and Tolken, Artesis Hogeschool* (2011).
[44] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision* 123, 1 (2017), 94–120.
[45] Pablo Romero-Fresco and Louise Fryer. 2013. Could audio-described films benefit from audio introductions? An audience response study. *Journal of Visual Impairment & Blindness* 107, 4 (2013), 287–295.
[46] Elliot Salisbury, Ece Kamar, and Meredith Ringel Morris. 2017. Toward Scalable Social Alt Text: Conversational Crowdsourcing as a Tool for Refining Vision-to-Language Technology for the Blind.. In *HCOMP*. 147–156.
[47] Jaime Sánchez, Mauricio Saenz, and Jose Miguel Garrido. 2010. Usability of a multimodal video game to improve navigation skills for blind children. *ACM Transactions on Accessible Computing (TACCESS)* 3, 2 (2010), 1–29.
[48] Emilie Schmeidler and Corinne Kirchner. 2001. Adding audio description: Does it make a difference? *Journal of Visual Impairment & Blindness* 95, 4 (2001), 197–212.
[49] Merrie Snell. 2015. *Lipsynching: popular song recordings and the disembodied voice*. Ph.D. Dissertation. Newcastle University.
[50] Joel Snyder. 2014. *The visual made verbal: A comprehensive training manual and guide to the history and applications of audio description*. American Council of the Blind, Incorporated.
[51] Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. " Person, Shoes, Tree. Is the Person Naked?" What People with Vision Impairments Want in Image Descriptions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[52] Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079* (2018).
[53] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In

Proceedings of the IEEE international conference on computer vision. 4489–4497.
[54] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*. 4534–4542.
[55] Agnieszka Walczak and Louise Fryer. 2017. Creative description: The impact of audio description style on presence in visually impaired audiences. *British Journal of Visual Impairment* 35, 1 (2017), 6–17.
[56] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. 2018. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7190–7198.
[57] Yujia Wang, Wenguan Wang, Wei Liang, and Lap-Fai Yu. 2019. Comic-guided speech synthesis. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–14.
[58] Yujia Wang, Liang Wei, Li Wanwan, Li Dingzeyu, and Lap-Fai Yu. 2020. Scene-Aware Background Music Synthesis. In *ACM Multimedia*, Vol. 38.
[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
[60] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8739–8748.

A PRE-STUDY INTERVIEW QUESTION

1. Demographics information:
   - Age
   - Gender
   - Occupation
   - Level of vision
2. What platforms do you use to access videos?
3. What types of video do you usually watch?
4. What is your experience of video added with audio descriptions?
5. What are the major barriers for accessing the videos?

B DURING-STUDY INTERVIEW QUESTION

1. After watching/listening to the original video/audio,
   - based on what saw/heard, could you describe the video content?
   - for what reason did you think the video currently needs to be added with audio descriptions?
   - what is your expectation for the audio description? What information do you want to get from the descriptions?
2. After watching/listening to the video with audio descriptions,
   - for each description, did you think the descriptions are confusing?
   - for each description, did you think the descriptions are redundant or have grammar errors?
   - what was your experience watching/listening the video added with audio descriptions? Did you think the descriptions help you understand with the video content?
   - what else information should the description provide?