Slidecho: Flexible Non-Visual Exploration of Presentation Videos

Yi-Hao Peng
yihaop@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Jeffrey P. Bigham
jbigham@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Amy Pavel
apavel@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

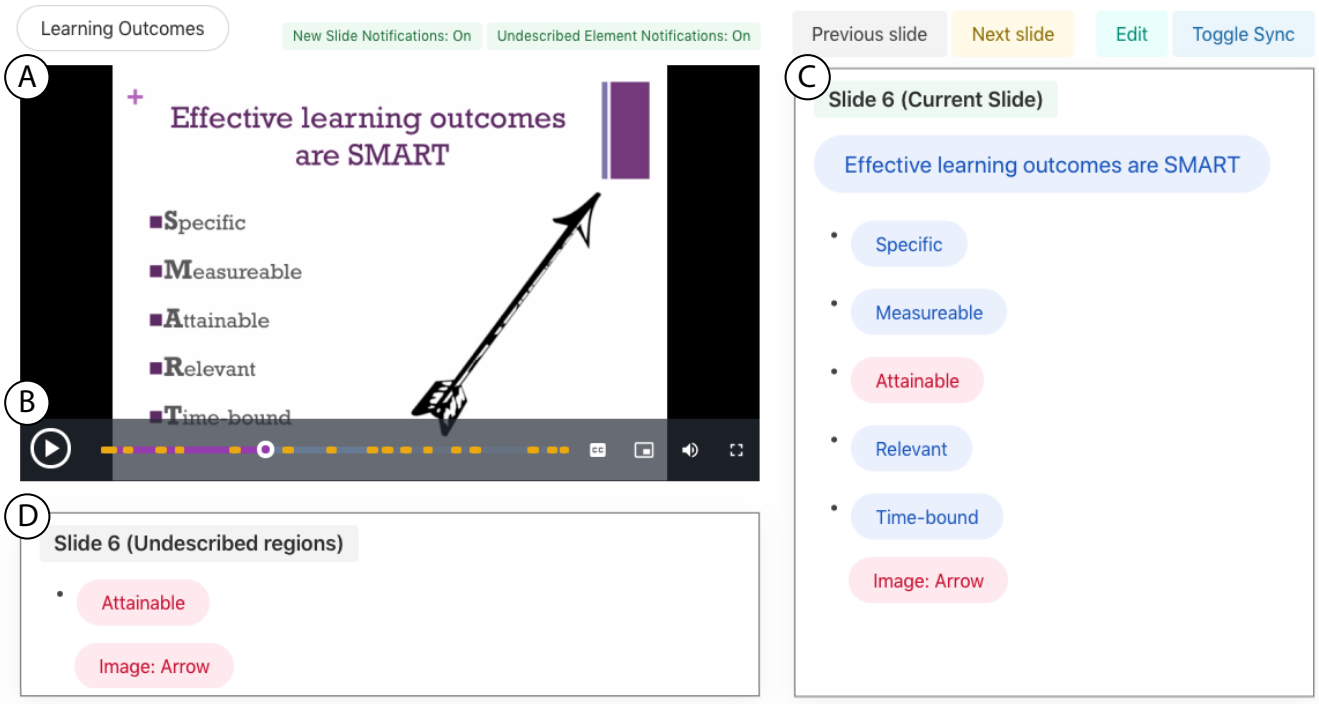

Figure 1: Slidecho features a (A) video player augmented with optional audio notifications for new slides, and undescribed elements, a (B) video timeline that lets viewers play/pause and navigate via slide boundaries, a (C) slides pane that enables viewers to navigate the slide structure (slides can optionally advance along with the video), and an (D) undescribed elements pane to quickly gain information about only the slide elements not present in the narration of a video. Edit mode ('Edit') lets people optionally correct Slidecho's automatically extracted slide boundaries, text elements, and image elements.

ABSTRACT

We present Slidecho, a system that enables non-visual access of the slide content in a presentation video on-demand. Slidecho automatically extracts slides and their text and image elements from the presentation video and aligns these elements to the presenter's speech. When listening to the video, Slidecho provides learners with audio notifications about slide changes and slide elements that are not described by the presenter. The learner can pause the video and browse the entire slide, or only the undescribed slide elements, to gain information. A technical evaluation with presentation videos in-the-wild shows that compared to the presenter's speech alone, Slidecho provides access to an additional 20% of total text elements and 30% of total image elements that were previously not described. Blind and visually impaired participants in our user study reported that it was easier to locate undescribed slide elements with Slidecho's synchronized interface than when browsing the video and extracted slides separately, and using Slidecho they read fewer slides that were fully redundant with the speech.

ASSETS '21, October 18–22, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-8306-6/21/10.
https://doi.org/10.1145/3441852.3471234

CCS CONCEPTS

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

KEYWORDS

Accessibility; Video; Audio description; Presentation; Slides; Multi-media consumption

ACM Reference Format:
Yi-Hao Peng, Jeffrey P. Bigham, and Amy Pavel. 2021. Slidecho: Flexible Non-Visual Exploration of Presentation Videos. In *The 23rd International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '21), October 18–22, 2021, Virtual Event, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3441852.3

1 INTRODUCTION

Videos are increasingly important for learning online across education, work, and hobbies. When recording educational videos such as course lectures, conference talks, and TED-style talks, many speakers use slides as visual aids. When speakers fail to describe key information in their slides, people who can not see the slides miss important information and visual jokes [52, 63]. This unequal access to information in recorded presentation videos adds to ex-isting inequalities in access to educational content for blind and visually impaired students [10, 11, 35].

To make presentation videos accessible to blind and visually impaired students, accessibility guidelines suggest that speakers [15, 40, 46]: describe the visual content during the presentation [40], add audio descriptions to the final video [15], and/or distribute accessible slides that include image descriptions and indicate the read-order for text [15, 40]. When speakers do not describe their slides, the speaker or a third party can add extended audio descriptions [15] that pause the video to describe important visual information in context of the narration (*e.g.,* pause the video after a speaker says "this insect" to describe that the slide depicts "a fruit fly"), but such recorded descriptions are time consuming to create and do not allow learners to flexibly access additional information. On the other hand, accessible slides let learners fully navigate the slide content and use the slide's titles, headers, and lists to skim for relevant information, but manually navigating the slides alongside the video to gain information is difficult (*e.g.,* a learner may hear "this insect" and read through all of the slides until they find "Image: A fruit fly"). Currently, most presentation videos online do not fully describe the visual content [52], contain audio descriptions [8], or provide corresponding slides (*e.g.*, neither TED videos [5] nor CHI conference videos [4] provide slides).

To make the visual content in presentation videos accessible, we present Slidecho[1], a tool that automatically extracts structured slides from the video and enables people to flexibly gain more information about slide content as they watch the video (Figure 1). Given a presentation video, Slidecho first extracts the slides along with their structured text and images, then uses Optical Character Recognition (OCR [1]) and image labeling [27] to provide initial descriptions that can be manually corrected using Slidecho's edit mode. Slidecho aligns the narration in the video and the content of the slides to determine when and what elements are not described by the speaker – or *undescribed elements*. As a person watches the video using Slidecho's interface (Figure 1A), Slidecho updates the current slide alongside the video and provides *audio notifications* to let users know about slide changes (*e.g.*, "Slide 1") and undescribed slide elements. When prompted by a notification or the presenter speech (*e.g.*, "You will use 'these' two essay prompts."), viewers can flexibly navigate through just the undescribed elements (Figure 1C) or all slide elements (Figure 1D). Users can optionally customize Slidecho by turning all audio notifications or slide and video synchronization on or off.

We evaluated Slidecho via a technical evaluation of each component of Slidecho's pipeline, and the novel interactions enabled by Slidecho via a user study. Our technical evaluation with presentation videos in-the-wild shows that compared to the presenter's speech alone, Slidecho provides access to an additional 20% of total text elements and 30% of total image elements that were previously not described. To evaluate the interactions enabled by Slidecho, we conducted a user study with 10 blind and visually impaired participants with prior experience viewing and/or creating slide-based content. Participants reported that Slidecho's audio notifications and synchronized slide and undescribed elements panes made it easier to locate visual content that was not described by the speaker than the traditional approach (our extracted slides and video consumed separately). Participants also read fewer slides that were fully redundant with the speech (compared to when using the existing approach) and expressed excitement about using the interactions enabled by Slidecho in the future for lecture videos, conference talks, and even live presentation.

In summary, we contribute:

- Slidecho, a system for extracting and flexibly exploring the slides from a presentation video.
- New interactions for flexibly gaining more information about video including: undescribed visual element notifications, and panes to explore the undescribed visual content.
- A user study comparing Slidecho's new interactions with the existing approach.

2 RELATED WORK

As Slidecho makes presentation videos non-visually accessible, our work relates to prior work on improving the non-visual accessibility of videos, visual content in education and presentations, and images (including those that commonly appear in presentations like photographs, charts, and diagrams). Slidecho's design also relates to prior work that aligns slides and videos.

2.1 Accessibility of Videos

To make videos more non-visually accessible, professionals traditionally create audio descriptions, or narrations of "important visual details that can not be understood from the main soundtrack alone" [2]. Guidance such as the Web Content Accessibility Guidelines (WCAG [15]), the Section 508 Rehabilitation Act, and the 21st Century Communications and Video Accessibility Act require at least *inline audio descriptions*, or audio descriptions that play alongside the video, placed within gaps in the speech [49]. Because inline audio descriptions are challenging to create, prior work proposed ways to make it easier to author descriptions including task-specific authoring tools [3, 14], tools for getting feedback on audio descriptions [47, 59], and a site to host descriptions [33]. Recent work also explores approaches that leverage Computer Vision and Natural Language Processing to quickly create audio descriptions by: detecting key visual content [21, 22], generating descriptions of the visual content [22, 56, 70, 73], and editing descriptions to fit within

[1]Slidecho is a portmanteau of "slide" and "echo" referring to our audio notifications for slide numbers and slide content that echo the transitions of the presentation slides as the presenter speaks.

the time between the speech [51, 70]. Inline audio descriptions that opportunistically place descriptions within speech gaps will not always make a video accessible because: some videos such as presentations rarely have pauses in the speech, and videos can be inaccessible even when speech is continuous [42]. To make videos accessible when gaps in speech are not available, authors can add *extended audio descriptions* that pause the video in order to play back description of the visual content (WCAG 1.2.7 [15]). Such descriptions are less common as they are time consuming to create and add video length, but some task-specific tools like YouDescribe [33] let people create them.

A core limitation of all audio descriptions is that they are a static summary of the visual content on screen, filtered through the constraints of the description type, and what the describer considers as more or less important. For educational videos, the needs of description may be particularly varied based on the task (*e.g.*, personal interest vs. finding an answer to a worksheet). Our work aims to help people to gain more information about visual content on-demand while preserving benefits of audio descriptions including: adding description at the time of the narration, and letting users know about missing visual content that is not obviously missing from the speech (*i.e.* providing access to unknown-unknowns).

2.2 Accessibility of Visual Content in Education and Presentations

The inaccessibility of tools for producing and consuming the visual content that is used in education for sighted learners is a long-standing problem for blind and visually impaired students and teachers alike (*e.g.*, graphical content in STEM classes [11, 26]). For instance, it can be challenging for blind presenters to create slides [26], and for blind and visually impaired students to understand a sighted presenter's presentation when it leaves out important visual content. To make live presentations more accessible to blind and visually impaired people, Hayden et al. proposed a note taking system that helps students magnify the lecture content while taking notes [30], and Peng et al. created a tool to help people remember to fully describe the content on their slides as people do not always remember to describe the slides even when instructed to do so [52]. After giving presentations, teachers can distribute accessible slides such as PowerPoint slides annotated with descriptions of visual content and in the appropriate read order [46]. Accessible slides are valuable but non-trivial to create. Ishihara et al. [34] and Sato et al. [60] created tools to make diagrams more accessible when accessing PowerPoint slides, and to turn PowerPoint slides into HTML respectively. Still, the presentation slides must be manually navigated alongside the video – a task that can be tedious (*e.g.*, when the slide content is mostly repetitive with the speech), and challenging (*e.g.*, if the slide boundaries are not obvious from the audio alone). Outside of formal education, raw PowerPoint slides (with or without an acceptable read-order and level of description) rarely accompany the presentation video – even in popular venues including TED-talks [5], online lectures, and ACM SIGCHI presentation videos [4]. Our work instead aims to extract accessible presentation slides from the presentation video, then let users flexibly navigate between the slides and presentation video.

2.3 Accessibility of Images

Slides in presentation videos contain a variety of text, image, and video elements. Prior work explored hvideos and accessible slides alone, the synchronizations and noti-fications provided by Slidecho improved the overall usability andaccessibility alike, as it enabled instant access to the existing and ad-ditional slide information just as users consumed the presentationspeech.ow to make a wide variety of images accessible on the web [12] or twitter [25] using a single approach per entire image (e.g., crowdsourcing [57], OCR [12, 25], or reverse image search [28]), or in some cases multiple approaches [24]. However, prior work for describing images collapses the structure of the image into a single text description that can be difficult to skim and browse for complex images that may have long descriptions (*e.g.*, an infographic, a flyer, or slide). Other work preserved the structure of the image. For instance, Pareddy et al. proposed an approach for exposing the structure of an underlying screenshot by capturing page structure at the time of the screenshot [50]. Our work aims to extract then let people navigate structured image content in the context of the narration about the image. Slidecho additionally classifies extracted slide elements as described or undescribed to assist users with information-seeking tasks.

2.4 Synchronizing Slides and Videos

Prior work aligned segments of a presentation video to parallel presentation slides, for purposes such as note-taking and supporting slide-based navigation of a video recording [36–38, 62, 65, 66, 72]. This primarily aligned entire slides rather than individual slide elements to the presenter's speech. For instance, Tsujimura et al. and Jung et al. consider real-time alignment of speech to slides for the purpose of sighted students glancing at the slides during the presentation to refer to the current speaker's location [37, 65]. However, the existing work is focused on cases where slides are already available, which is rarely the case for presentation videos online. Further, the prior work builds visual interfaces intended for sighted users glancing at the slide content during the lecture. We instead design a non-visually accessible interface aimed at preserving benefits of existing methods for making presentations accessible, leading to different design considerations (*e.g.*, image descriptions, audio notifications).

3 CURRENT PRACTICE AND GUIDELINES

A few key strategies exist for making presentations and recorded videos non-visually accessible: (1) describing the visual content during the presentation (also known as *embedded descriptions*), (2) creating *audio descriptions* after the presenter records their video, and (3) distributing *non-visually accessible slides*. To preserve the benefits of the existing approaches in designing new tools, we leverage established guidelines for describing/distributing slides [7, 19, 20, 40, 46, 48, 54, 67], and audio descriptions [2, 6, 18, 32, 33], developed by and in collaboration with blind and visually impaired people. We summarize guidelines related to non-visual access to presentation videos:

Embedding descriptions of presentation content: Describe relevant visual information on the slides including text, images,

graphics, and other visual aids (Embedding Guideline 1, or EG1). When describing visual content, use nouns instead of pronouns (EG2), and include context and regions of interest (EG3). Before playing a video, summarize the content, then narrate the action in short phrases as it plays, or use an audio-described video (EG4).

Audio descriptions: Describe important visual content (Audio Description Guideline 1, or AG1). Avoid describing content that can be inferred from the audio (AG2). Start general then progress to details (AG3). Avoid overlapping audio description speech with speech in the video (AG4). Do not "spoil" the video by describing surprising content before it appears (e.g., the speaker provides the answer to a riddle) (AG5). Use extended descriptions, which pause the underlying video, only when necessary (AG6).

Non-visually accessible slides: Set an appropriate reading order for elements (Slides Guideline 1, or SG1), and provide alternative text for non-text elements (*e.g.*, photographs, charts), unless the element is decorative (SG2). Group related graphic elements to make tab navigation easier (SG3). Add audio descriptions for videos in the slides (SG4).

We build on well-established guidelines to design a system that enables learners to flexibly gain additional information about visual content as they watch a recorded presentation video. While we cannot alter how the presenter in a video described their slides (as EG1-EG4 assumes), we can use guidelines about what (AG1, AG2, EG1, EG3, EG4 and SG2, SG4) and how (EG2, AG3-AG6) content should be described to inform (1) when additional description might be needed (*e.g.*, when the presenter has used a pronoun rather than a noun to describe content - EG2), and (2) how our system should add the additional description (*e.g.*, do not overlap description speech with video speech - AG2, and only pause the video for description when necessary - AG6).

4 INTERFACE

Slidecho makes presentation videos accessible by letting users quickly navigate the video alongside the extracted slides. Slidecho's interface consists of: (1) the *video pane* that lets users play back the presentation video (Figure 1A), (2) the *slides pane* that lets users navigate the slides alongside or separately from the video (Figure 1D), and the (3) *undescribed elements pane* that lets users navigate through only the elements that the speaker did not cover on the slide (Figure 1C). Slidecho's *edit mode* lets the presentation author or a third party manually adjust the text recognition, image descriptions, and slide boundaries.

4.1 Video Pane

Users can listen to and navigate the video using the video pane (Figure 1A). The user presses alt+shift+slash keys to play and pause the video. Users can play and pause the video when their screen reader is focused anywhere on the screen. Users may jump forward or backward in the video by one slide by pressing alt+shift+right bracket or alt+shift+left bracket keys, respectively. As the user listens to the presentation video, they may optionally select to also hear audio notifications for slide boundaries (plays back "Slide 3" when the transition to Slide 3 appears) so that users can keep track of when new content may be available to navigate. In addition, users can turn on two types of undescribed element alerts 2: (1) a short "extra slide elements" alert that plays then pauses the video when Slidecho detects the speaker has not described one-third of the slide content and is about to move on to the next slide, or (2) a read slide elements alert that automatically reads the slide elements that were not mentioned by the speaker at the end of each slide. For example, if the presenter gets to the end of a slide without describing "Attainable" when defining the acronym SMART (Figure 1), the system will pause the video and use speech to text to say "undescribed elements: 'Attainable', 'Image: Dart"'. After each undescribed element alert the video will pause so users can explore or press a key to continue playing.

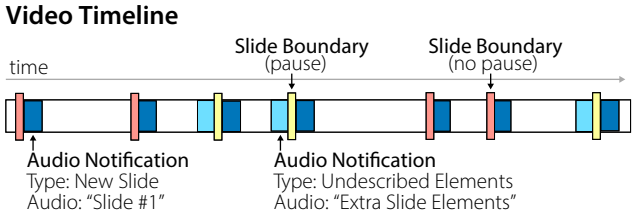

Figure 2: As the video plays back, Slidecho can provide two types of audio notifications (that can be toggled on/off by the user at anytime): (1) new slide notifications (dark blue), and (2) undescribed elements notifications (light blue). A pause (yellow) occurs after each slide boundary with undescribed elements to make it easy for users to explore the slide.

4.2 Slides Pane

The slides pane enables users to navigate by the extracted slides (Figure 1D). Each slide contains text and image slide elements, each with a description generated either automatically or edited after the fact by the presentation author or a third-party. As the user reads through the text and image elements, they can click on an element to jump to the point in the slide. For instance, if a user reads a slide with only the text "Outcomes or Objectives?", they may click on the element in the slide pane to play the video at the start of the sentence that mentions the slide element, in order to hear the (presentation) context of the element in the narration: "there is no consensus in higher education about the use of the words 'learning outcome' and 'learning objective' ". The user may use the slide pane to read all of the slides all the way through, or they may click "auto update" to have the slides synchronize with the video to allow easy access to the relevant slide as the user watches the video. For example, if a user listened to the speaker describe an acronym "SMART", they can jump to the corresponding slide pane (Figure 1D) and read the acronym text elements in order as a list: "Specific, Measurable, Attainable, Relevant, Time-bound."

4.3 Undescribed Elements Pane

The undescribed elements pane lets the user quickly navigate only the elements that were not mentioned by the speaker (Figure 1C). The undescribed elements pane appears only when the slides are

synchronized the video, and it will include all undescribed elements for the current slide displayed in the presentation video. For instance, if a user heard the speaker describe the acronym "SMART" while watching the video, they could jump to the undescribed elements pane instead of the slide pane to navigate to: "Attainable." (Figure 1C) If the undescribed element needs more (slide) context, the user can subsequently visit the full slide on the slide pane to read that element as part of the entire slide.

4.4 Edit Mode

Edit mode allows the end user, presentation authors, or other third-parties to improve upon Slidecho's automatic analysis of slide text and images. Users can click on any slide element to edit its description (e.g., to change the description of an image element from "graphical user interface or application" to "A poster of the movie Supersize Me with Morgan Spurlok"). In edit mode, users may edit the slide boundaries on the video timeline or toggle whether a slide element is undescribed or not.

5 ALGORITHMS

To power Slidecho's interface, we build a computational pipeline that given an input presentation video: (1) identifies the presentation video type and extracts the slide frames, (2) describes the text and image elements on the slide, (3) converts each slide into an accessible HTML format, (4) aligns the tokenized sentences of the transcribed speech with slide element descriptions, and (5) adjusts the slide time boundaries based on the alignment results to insert the audio notifications.

5.1 Extracting Slide Frames

To extract slides from the video, we first determine what frames in the video contain slides, and the position of the slides within the frames (*e.g.*, full screen slide, or slide with presenter talking-head view). We use Google Video's Intelligence API [27] to identify frames that contain the label "public speaking". If the label occurs for more than 90% of the video time, we assume the slides are on screen for the entire video – either full-screen or with an additional view of the presenter's head (*e.g.*, Zoom screen-share recordings with a front-facing camera). To determine the spatial boundary of the slide, we identify a bounding box that contains the maximum amount of text recognized by Optical Character Recognition (OCR [1]) across the video. We then detect shot boundaries by using an existing API [27] that considers edge-based differences [9] and OCR text distances [41] for the consecutive video frames similar to prior work [9, 13, 37, 74]. To assure the slide we extracted would not be mid-build (*e.g.*, the text was not done being animated in), we assigned the final frame of each slide interval (the time between adjacent shot boundaries in the video) as the representative slide frame. If the speakers did not appear over 90% of the time, we identify that the video likely includes the a mixture of slide-focused views (e.g., full-screen slide, or slide-focused camera) and other types of shots (e.g., presenters movements or audience reactions). In that case, we first apply the same shot boundary detection, and then remove the non-slide segments (labeled "public speaking" or "audience" [27]). For non-slide segments, we assign the corresponding slide by extending the time boundary of the closest previous slide segment and assigning the representative slide frame from original slide segments to the newly constructed ones.

5.2 Describing Slide Elements

Slidecho then creates descriptions for slide text and image elements. For text elements, we use OCR [1] to recognize text on slides. For image elements, we first segment out images from the rest of the slide following prior work [53]. Specifically, we first remove the recognized text from the slide frame using the recognized text bounding boxes produced with OCR (We remove text only to pick the bounding boxes, but we add the text back afterwards for the description and the interface). Then, we scan through the slide frame for pixel changes using a sliding window and identify the images as the regions with enclosed boxes bordered by regions with no visual content. If we do not find any image segments, the slide typically contains only text or a full screen image. We avoid generic descriptions (e.g., label "text") by only obtaining a description for the full slide if the number of text elements fell within empirically determined thresholds. Specifically, we extract the whole slide view as the single image element if the slide contains: few text elements (<5), objects or scenes (as recognized by a detection API [27], or many text elements (>100, e.g., a screen shot). Future work may get descriptions for all full slides and remove generic descriptions (e.g., "text"). After segmenting each image element, the system then recognizes the descriptions of image elements using Microsoft Cognitive Services Vision API to get the resulting captions with highest confidence score. The generated descriptions or text can be edited using Slidecho's edit mode.

5.3 Converting Grouped Slide Elements into Web Elements

To format extracted slide elements as an accessible HTML format that users can skim and browse, we first group described elements and determine element read order. Our system supports three types of grouping structures: (1) text group, or a segment of plain text, (2) text list group, or segment of text with bullet points or other list icons where each point of item is the independent text group, and (3) image.

To form the initial structure of the text groups and the text list groups, we first use the block entities recognition provided by Google Vision API's dense text detection. In pilot tests the API's blocks alone were often inaccurate text groups that commonly appear on slides (*e.g.*, bulleted lists), so Slidecho further processes each text block to determine its final structure in the HTML slide. For each block, we first detect whether the block contains the title element or not by considering text size and position heuristics (as in [29, 37]). For each non-title block, the system first detects if there are any bullet-point style of symbols [71] at the beginning of any line. If so, we transform the block into an HTML unordered list using the spatial relationship between the first non-symbol characters of each line to determine list elements.

For the remaining blocks (no bullet symbol or title), we determine if the block is either a single text group, multiple text groups, or a non-bulleted text list. For each line of text in the block, if its initial character is lowercase, we add the line to the last group. If the initial character is uppercase, we create a new text group of type text list

or new text group depending on the condition met: (1) text list, if the first initial character is indented to the right of the above line, or if the first line of block is styled in bold or ended with colon, or (2) new text groups, if punctuation is found at the end of the above line. We represent the text groups in their corresponding HTML structures, and append "Image:" to image descriptions for images on the slide. After all slide elements are grouped and structured as web elements, we assign the read order for each of them based on the their distance from upper left-hand corner following prior work [52].

5.4 Aligning Slides to Speech

To help users flexibly navigate between the slides and video, and to easily identify the *undescribed slide elements*, we align the structured slide elements with the speaker's narration. To achieve the goal, we transcribe the speech using Google Cloud Speech-to-Text API, which also provides timing for each word and punctuation. Then, we use the sentence tokenizer from NLTK [44] to parse the text into sentences. To avoid starting the video mid-sentence when a user clicks on a slide element, we use the sentence as the minimum transcript unit for aligning slide elements. For each slide, we first the extract the slide's transcript sentences as the set of sentences that overlap in time with the slide boundaries (as initially detected by shot detection), and add to that the sentence immediately preceding the extracted set (to catch the speaker's transition to the next slide). Then, we encode each of the slide's text and image element descriptions as well as the slide's transcript sentences to get their embeddings using the sentence transformer [55] with the pre-trained RoBERTa model [43]. We compute the cosine similarity between all possible sentence embedding, slide element embedding pairs. Slidecho then assigns each slide element to the most similar sentence to determine its narration time. If similarity is equal for two sentences, we chose the sentence spoken first. Meanwhile, we determine an element is "undescribed" if no match exceeds an empirically-defined similarity threshold of 0.3. In the future, we could improve the classification of "undescribed" using features of the text, audio, and visuals (e.g., decorative elements may be less likely to be described).

Because Slidecho uses slide boundaries to insert audio notifications and update the slide shown in the slide pane alongside the video, we adjust the visually-detected slide boundary times to better fit the narration and avoid interrupting the speaker mid-sentence. Specifically, if we find an optimal element match in a sentence that starts before the visual transition time, or if the detected slide transition time is in the middle of a speaker's sentence, Slidecho adjusts the boundaries to the start time of the nearest prior sentence.

6 TECHNICAL EVALUATION

To assess the performance of Slidecho's algorithmic components on presentation videos in-the-wild, we conducted a technical evaluation of Slidecho with 20 online videos containing a total of 88.8 sampling minutes, 158 unique slides and 574 slide elements. The videos were randomly selected from the previous established dataset [52]. Our sampling achieves a set of videos with a range of presentation styles (e.g., TED talks, seminars, and course lectures) and domains (e.g., applied science, nature science, social science, humanities). To demonstrate the system performance under the scope of this work, we excluded videos where the slides were not fully visible by a human (e.g., due to occlusion or resolution), or the presentation itself contained videos. For each stage of the pipeline, we report quantitative metrics of accuracy by comparing Slidecho's labels to manually labelled ground truth annotations (e.g., correct slide boundary shots and elements; the best-matching narration sentence for each slide element) for our video set following prior works [37, 74], and a qualitative analysis of errors obtained by manually examining the algorithm's failures, and identifying common themes.

6.1 Slide Time-Boundary Detection

Our slide boundary detection (visual detection plus post-alignment adjustment) achieved an F1-score of 97.2% (97.1% precision, 97.6% recall). Boundary detection failures occurred only in live-presentation style videos that contain a mix of slide, presenter, and audience shots edited into a video. The detection missed slide boundaries when the presenter switched the slide, but the video did not provide a close up of the slide (Figure 3A). The detection added extra slide boundaries when the video contained multiple close up shots of a single slide separated by non-slide shots.

6.2 Recognizing Slide Elements

We evaluated the performance of the text OCR detection and image segmentation and recognition results respectively. For text elements, OCR [1] incorrectly recognized only 1.6% of characters (mostly special symbols) on the representative slide frames. For image elements, we reported our image segmentation accuracy and the description quality quality [25, 58] of the auto-generated image descriptions. Our image segmentation achieved an F1-score of 91% (precision: 87.6%; recall: 95.4%) for segmenting out images from representative slide frames. Our high recall (we extracted 95.4% of images on the slide) assures that users do not miss out on many images that were present on the slides – the lower precision implies that users may navigate over some images that were not segmented correctly. Such segmentation errors most often occurred when the prediction segmented many small image elements (*e.g.*, many human faces shown in Figure 3B), rather than grouping them into a single image element (*e.g.*, one speaker intended to represent the concept "the public"). Ideas from prior work on grouping together slide shape elements in slides could be applied in the future to improve the image segmentation step [34]. Microsoft's auto-generated scene descriptions for slide images that were correctly segmented, coded following prior work [25], showed that 18% of generated descriptions were irrelevant, 45% of them were relevant, 31% of them were good and 6% of them were great. For example, an irrelevant description recognizes the Wikipedia logo as "A closeup of a coin"(Figure 3C).

6.3 Grouping Slide Elements

We compared our predicted slide element groups (used to inform the structure of the slide HTML) to our ground truth slide element groups. Our method achieves a F1-score of 87.8% (91.5% precision, 89.0% recall). The errors were most common when there were few text elements (Figure 3D) on the slide (as above, images and shapes

Figure 3: Errors in Slidecho's pipeline for extracting slides from a video: (A) Slidecho did not detect a slide boundary because the video did not include a close-up of the slide [64], (B) Slidecho returned 13 image segments but the presenter intended one image to represent "the public" [16], (C) Slidecho described this Wikipedia icon as "a close-up of a coin" [69], (D) Slidecho did not segment individual elements [64], and (E) Slidecho introduced extra segmentation into a single sentence segment [61].

were occasionally grouped incorrectly), and when the text segments were particularly spread out (the initial block estimation would not include the separated content as shown in Figure 3E).

6.4 Element-Speech Alignment

We compared our slide element to speech sentence alignment results with ground truth alignment. Our slide element to sentence alignment achieves an 84.3% F1-score (87.4% precision, 82.1% recall, 20% of talks have an F1 score of >= 90%). The performance of the alignment was primarily presentation-dependent, such that Slidecho achieved high accuracy for some talks (maximum F1-score of 94%) and much lower accuracy for others (minimum F1-score of 75%). Well-structured talks in which the speaker mentioned each slide element in turn, and there was not much overlap in the narration used to describe slide elements, achieved high scores. More casual presentations where presenter's narration was more loosely related to the slide content (*e.g.*, a casual discussion might include redundant mentions of a slide element, or vague references to elements that do not use the same terminology).

In our ground truth labeling, only 14.5% of images had a description of the image present in the narration, and only a subset of those had a good or great auto-generated description. So, image element to sentence alignments were nearly non-existent (2% correct predicted alignments). But, images, unlike text, sometimes occupied the entire slide so users can rely on our high-quality slide boundaries to find coarse regions of narration that described the image.

6.5 Reflection

Slidecho's near-perfect performance for text recognition and slide boundary detection (powered by state-of-the-art services) enables users to view the slide text in its exact form at any time alongside the speaker's narration. The ability to view the exact slide text is important as the slide text significantly summarizes the content in the narration and makes it easy to digest the key points (the narration is typically at least 2x longer in word-length than the slide text). In our sample, presenters neglected to mention 1 in 5 of their text elements, such that Slidecho would make 20.4% of all text on slides newly accessible.

On the other hand, only 37% of all image descriptions had 'good' or 'great' quality. While image description accuracy could be improved in the future (via future description algorithms, manual expert editing, or crowdsourcing), 82% of errors consisted of descriptions that were relevant but incomplete such that users could recognize the need for additional visual information. On the other hand, some of the remaining 18% of 'irrelevant' descriptions could be misleading (*e.g.* in Macleod et al. participants imagined reasons for irrelevant images [45]). Still, speakers often did not describe their images (85% of all images were not described), and Slidecho was able to make 30.1% of all slide images newly accessible (i.e. the speaker did not describe the image and it had a 'good' or 'great' description). In the interface, each of these undescribed images would trigger an audio notification and the user could read the image description to identify more information.

7 USER STUDY

Slidecho's design is informed by existing methods for making presentations more accessible to blind and visually impaired learners including distributing accessible slides along with the video (to enable simultaneous browsing) and adding extended audio descriptions (to add information just-in-time). We conducted a user study with blind and visually impaired people to learn:

- How do blind and visually impaired learners use the slides alongside the video to gain more visual content?
- How will Slidecho's speech/video synchronization and audio notifications impact blind and visually impaired learner's ability to locate relevant visual content?

7.1 Methods

Materials: We selected two popular videos of a similar style and intended for a general audience (a 3.5-min TED-talk about a 30-day challenge [17], and a 5-min TED-talk about online collaboration [69]). We selected short, broadly understandable, and slide-based TED talks such that participants had adequate time to fully explore the lecture content using our interface. We chose clip boundaries to assure a similar amount of described and undescribed elements in each clip. We selected one additional short, general-audience lecture video titled "Learning Outcomes" [68] for a tutorial phase. None of these lectures were paired with accessible slides, so we ran all three videos through Slidecho to obtain the extracted slides and slide-to-video alignment. We used Slidecho's edit mode to manually correct image descriptions for the clips to keep accuracy consistent. In total, we edited image descriptions on 4 of 20 slides (7 of 13 total images) to bring image descriptions to a level of

"good" or above (e.g., changed "a close up of a coin" to "the logo of Wikipedia"). Meanwhile, we did not edit other system results. To evaluate how Slidecho's novel features (synchronized slides and video along with the corresponding audio notifications) would help viewers navigate relevant visual content, we created two versions of Slidecho: (1) the Slidecho interface with the extracted slides and video side-by-side without synchronization (*i.e.* "**no-sync**") to mimic current practice, and (2) the full Slidecho interface with the extracted slides synchronized to the video with audio notifications (*i.e.* "**sync**").

Procedure: We conducted remote user studies via Zoom with 10 blind and visually impaired participants to learn how people would use our interfaces to navigate videos and slides together to learn more about the visual content. During each 1 hour long study, we first asked participants a few demographic questions (*e.g.*, age, gender, how they would describe their visual impairment) and asked participants about their prior experience with both recorded and live slide presentations. Then, we provided a 10-minute tutorial of the two interface versions using the lecture-style "Learning Outcomes" video. After the tutorial, participants spent time watching the content of each of two TED videos, each with a different interface (sync or no-sync). We randomized the order of the videos and interfaces that participants saw together to make sure half of the participants watched each video with an interface different from the other half of participants did. We told them to stop when they felt they had fully consumed the presented content. After participants viewed each video, we asked two questions: one that required only the visual content and one that required only the audio content. Specifically, we asked "when is the national novel writing month?" (speech only, answer: November) and "what does the speaker day look like after he passed the thirty-day plan of sugar control?" (slide only, answer: a stack of chocolate) for the 30-day challenge talk. For the online collaboration talk, we asked "What are the two barriers for doing large-scale language translations for free?" (speech only, answer: lack of bilingual people and motivation) and "What was the speaker's favorite translation sentence?" (slide only, answer: "Please apologize for your stupidity. There are a many thank you"). Then, we conducted an interview about their experiences with our interfaces both quantitatively and qualitatively. We compensated participants $30 for the 1 hour session via a choice of PayPal or Amazon Gift Card.

Participants: We recruited the participants using an email list that the authors had access to (the participants were originally recruited from social media). Participants ranged from 23-55 years old (3 female, and 7 male) and described their visual impairments as blind (8 participants) or low-vision (2 participants). All participants had experience watching slide-based presentations in person and online. Participants who consumed presentation videos online mentioned that they had watched: TED talks (6 participants), course lectures (3 participants), tech talks (3 participants), workshop recordings (2 participants), Khan Academy videos (1 participant), and other explainer or instructional videos (2 participants). Half of the participants mentioned that they rarely found the corresponding accessible slides for the talk. All of the participants used our interfaces with a screen reader (5 JAWS, 3 NVDA and 2 VoiceOver) and 1 of them used a braille display as an additional modality to access information.

Measure and analysis: We recorded each Zoom session including the audio for the interview questions, and screen-sharing when participants used the interfaces. We analyzed the interview by grouping the interview notes into themes, and returning to the interviews to extract specific quotes. To assess how each interface was used, we tallied whether participants used each interaction type while watching the recorded session videos. We also collected 7-point Likert scale ratings (the higher, the stronger the metric score) with respect to: "helped me locate the undescribed slide elements", "helped me keep updated information across videos and slides", "the mental effort required", "the level of distraction interface caused", and "the level of accessibility improvement the interface provided for the slide presentation videos".

Study Limitations We manually corrected the auto-generated image descriptions to assess the interactions provided by Slidecho. Despite this limitation, Slidecho is currently accurate for 80% of the slides in the study (*e.g.*, the slides contain only text or already have accurate image captions). To accommodate necessary corrections we designed our interface to allow manual expert editing. In the future, we expect manual correction to be less necessary as image descriptions improve over time and by reusing portions of speaker's narrations as descriptions, our system could make it easier to add slide alt text. In addition, we selected short TED-style lectures for the study. These slide-based lectures are accessible to a general audience, and allowed exploration of narration, text, and images — that together account for many slide-based talks. Exploring the impact of talk content (*e.g.*, a technical talk with complex images and diagrams), length (*e.g.*, a 5 minute talk vs. 90 minute lecture), and context (*e.g.*, watching an entire lecture vs. watching a short segment) could be important future research.

7.2 Results

We report on the benefits and trade-offs of having the synchronized slides with audio notifications, and the side-by-side slides and video in terms of interface use, preference, and scenarios for future use.

Interface Use: Participants spent 6.38 minutes (σ=1.98) consuming each video. All participants played the entire video with both interfaces. Participants read significantly more redundant slide elements with the no-sync interface ($\mu = 8.50$, $\sigma = 0.50$) than they did with the sync interface ($\mu = 3.90$, $\sigma = 2.17$) ($F = 38.55$, $p < 0.01$ via One-way ANOVA). Participants also spent significantly more time ($\mu = 7.30$ minutes, $\sigma = 2.33$) using the no-sync interface than they did using sync interface ($\mu = 5.46$ minutes, $\sigma = 0.68$) ($F = 5.20$, $p < 0.05$ via One-way ANOVA). Despite spending less time viewing the video content with the sync interface, all participants answered the factual questions about the audio-only content (1 question per video) and visual-only content (1 question per video) correctly in both conditions.

Using the no-sync interface, all participants played the entire video and read all of the slide elements, either by: (a) reading all of the slide elements at once after watching the video (6 participants), (b) watching a portion of the video, reading a portion of

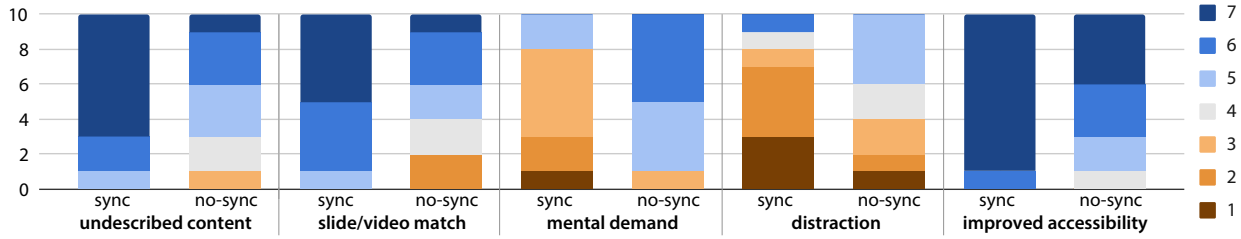

Figure 4: Likert Scale ratings from the user study for 5 questions (helped me identify undescribed elements, helped me find matching slides and videos, mental demand, distraction, and improved accessibility) from 1-low to 7-high and 2 interfaces (sync and no sync). Full table of ratings included in the Supplementary Materials.

the slide elements, then watching the end of the video, and reading the rest of the slide elements (2 participants), or (c) watching the video and reading the slides at the same time (1 participant with a braille display, and 1 participant with a screen reader). P6's process exemplified the most common approach: P6 played the video until the speaker said "So I'll let you read, this person starts by apologizing", P6 paused the video and read through all slide elements on prior slides until they reached the corresponding slide (Slide 3), read Slide 3, then played the video to the end before reading the rest of the slides. All participants viewed slide content that was redundant with the presenters speech, and most (8 participants) viewed the slide content before or after the time it was mentioned by the speaker (e.g., reading and laughing at a verbal joke minutes after the audience laughed in the video). P1 uniquely played the screen reader and the video at the same time such that the speech track of the screen reader and video overlapped entirely.

With the full Slidecho interface (sync), 8 participants viewed the slide content when and only when they heard the "extra slide content" audio alert indicated they were missing information about the visual content. As one participant described:

> "if it says there's extra content, I will look at the text of the slide. If there's not, I'll just listen along. Cause there would sometimes be a stretch of like two or three slides, whether it wasn't and where, what he was saying, encompassed everything." – P8

When participants paused the video to explore the slides, 8 participants visited the undescribed elements pane first, then subsequently viewed the slides pane to read the undescribed elements in some cases (*e.g.* to understand the context of a standalone video). Compared to no-sync, all participants read fewer slide elements with our interface, as the audio notifications made it so participants did not have to manually search for unknown-unknown visual content. In addition, when participants did read the slide elements, they read the slide elements closer to the time of the slide's narration. For instance, P2 beat a sighted audience to the visual joke – P2 heard the joke set-up, paused the video, read the image, and laughed before unpausing the video to hear the audience laugh.

Interface Preferences: Participants provided both qualitative and quantitative feedback about their interface preferences. Overall, 8 participants preferred the sync version of Slidecho and the remaining two participants (P4 and P9) preferred the side-by-side no-sync interface. Among two participants who preferred the side-by-side no-sync interface, P4 (who uniquely consumed the video and slide content simultaneously) reported that the no-sync allowed more flexibility to navigate backwards and forwards through the slides while the video was playing. P9 (who read the slides first then played the video) told us that she preferred the no-sync interface, because she would prefer anything she was more familiar with. 6 participants said they would most prefer to have both interface options available.

Participants rated their ability to identify undescribed slide elements as significantly better with the sync Slidecho ($\mu = 6.60$, $\sigma = 0.66$) than with the no-sync interface ($\mu = 5.10$, $\sigma = 1.14$) ($p < 0.01$ via dependent t-test) (Figure 4). P5 specifically liked that the interface told them where the undescribed content would be to navigate, reporting "if it wouldn't tell me that it was there, I wouldn't know." P1 shared that they would view undescribed content more often with the sync Slidecho interface because: "I wouldn't have to make a decision as to whether or not I cared about the undescribed content when it's easier to find" (P1). Participants also rated the ease of navigating to the most relevant slide for a given video time as higher for Slidecho ($\mu = 6.40$, $\sigma = 0.66$) than with the no-sync interface ($\mu = 4.70$, $\sigma = 1.62$) ($p < 0.05$ via dependent t-test). As P5 explained, "I didn't have to navigate the keyboard near as much as I did with [no-sync]". Participants also reported that they gained agency when being able to get the complete and updated information: "The updated slides let me feel that I was really participating in the talk and I didn't have to be worried about if I was left behind.' (P6)

Surprisingly, participants also rated the more complex Slidecho(sync) interface as significantly less mentally demanding ($\mu = 3.00$, $\sigma = 1.18$) than the no-sync interface ($\mu = 5.30$, $\sigma = 0.90$) ($p < 0.01$ via dependent t-test). P2 mentioned that when reflecting on the mental demand of Slidecho "I have to pause [the video] still at the right time, but it's a lot easier to know when to do that." However, they pointed out that they would like to add a system mode for Slidecho that would automatically pause at the end of all slides (Slidecho only automatically pauses for slides where there are more than 33% unmentioned elements, which occurred for 60% of the slides in our two video samples). Similarly, participants found Slidecho (sync) interface to be less distracting ($\mu = 2.40$, $\sigma = 1.50$) than the no-sync interface ($\mu = 3.70$, $\sigma = 1.35$) , but the difference was not significant ($p = 0.064$ via dependent t-test). P8 explained that the audio notifications and pauses "I don't think it's really

distracting to me at least. I think that what it's doing is, is worth the stop start." On the other hand, P5 reported that the unsynchronized interface was more distracting because needing to search for a relevant slide to match speaker's narration would take you out of what the speaker was saying.

Overall, participants rated the synchronized interface as improving the accessibility of the presentation video significantly better ($\mu = 6.90$, $\sigma = 0.30$) than the non-synchronized interface ($\mu = 6.00$, $\sigma = 1.00$) ($p < 0.05$ via dependent t-test). As explained by P7, "I think the interface with synchronized information and corresponding alerts provided me more granular access to some of the mysterious points at the most closest moment as it can possibly be" (P7). P3 expressed how Slidecho's improved accessibility enhanced their engagement with the content: "Knowing that there will be audio notification to indicate what I have missed help me keep engaging in the content itself. Usually when I found there were things being inaccessible in the video and continued for a certain amount of time, I might just skip the content. The interactions make me stay focused as the audience since I knew I can review what I need to additionally know without putting too much effort to seek by myself" (P3).

Interface Improvements: Participants also provided some suggestions on how we can improve the current interactions. A few participants commented on Slidecho's text-to-speech synthesis: "I would prefer only hearing the audio feedback from the JAWS instead of some voice generated from the web since I already felt there were too much information scattered around the website nowadays" (P4). Slidecho only uses text-to-speech for audio notifications (*e.g.*, "Slide 1", "Extra slide content"), and in the future we could use participant suggestions to include a status pane with header that alerts audiences about the undescribed content through text. For the audio notification, we could also explore non-speech sounds. Though our participants customized Slidecho (*e.g.*, one participant turned off new slide notifications), participants suggested additional customization options including: changing the hot keys for changing interface panes, and changing the notification content and speed. Participants also wanted to toggle between synchronized slides (that support video-first exploration) and unsynchronized slides (that support slide-first exploration). We provide this toggle in the updated version of the interface.

Past Experiences and Future Use: In the pre-task interview, 7 participants reported they had received slides before or after a presentation in the past, but that in most cases these slides were still not completely accessible (e.g., lack of image descriptions, wrong read order). For online presentations, participants reported that it was even difficult to find any available slides. Reflecting on their prior experience, 4 participants mentioned during the formative interview that they would always prefer to play the video first then read through the slides afterwards (*e.g.*, as compared to audio descriptions, or interleaving navigation with watching). After using Slidecho, all 4 of these participants preferred and wanted to use the synchronized interactions that interleave the video and slides in the future.

In the post-task interview, the ideal context for the use of each of interface varied between participants. When discussing about the context of future use of the sync interface, 2 participants mentioned they would use it for recreational videos (e.g., some of the TED talks), 4 of them will use it for class lectures and 1 for tech talk. As for the no-sync interface, 5 participants would like to use them (along with the note-taking), 1 for recreational video and 1 for more detailed instructional videos. 3 participants replied that they would like to use both interfaces at any applicable scenario.

8 DISCUSSION

8.1 Supporting Complex Elements and New Domains

Slidecho supports presentations with structured text and images. The performance of Slidecho provides especially accurate slide segmentation, text recognition, and grouping of common text structures (*e.g.*, lists, paragraphs) for presentation videos in our sample. In future work, we will support more complex text structures including tables and equations. To do this, we will need to better recognize the text characters and the relationships between them to provide high-quality structured HTML of the content. In addition, in the future we could also use existing non-visual approaches to navigate structured image and text content such as diagrams and graphs in the context of Slidecho, as well as consider how to describe such content given the specific context of the speaker's narration. Slidecho also does not yet support GIFs or videos that appear in slides – but prior work may be used to improve the accessibility of this media [23, 51, 70]. While we explored only presentation videos, Slidecho could be applied to many types of videos that feature static frames for showing text and images including: late night news programs that use slides as visual aids, documentaries, explainer videos and photo slideshows. Our interactions for supporting non-visual explorations could also be further applied to more dynamic videos as well – replacing slide segmentation with visual *event segmentation*. Future work might also examine the applications of these interactions to live lectures (*e.g.*, by exploring near-realtime algorithms for detection and segmentation) and eventually other types of video streams.

8.2 Creating Audio Descriptions for Presentations

Audio descriptions provide people information about the visual content in a video in context with the narration (*e.g.*, a character says "Look at this!" and the audio description immediately says "Cindy holds a tangerine."). In the user study, participants used Slidecho to gain visual information in context with speaker's narration in the current slide, before moving on to the next slide and video segment – thus, viewing visual information closer to the narration context than they did with the current approach (watching the entire video, and then watching the slides afterwards). However, there is room to improve the timing of the descriptions of visual content that the speaker is missing. For instance, speakers occasionally mention multiple elements on the screen before the slide (*e.g.*,"I'll let you read this..." for the first slide element, followed by "Here is the next question..." referring to the second element). Using Slidecho in these cases, viewers still need to remember what reference refers to what element (even if for a shorter amount of

time than the prior approach). In the future, we could detect visual references [42] (*e.g.*, "this" and "here") then place descriptions of the undescribed context within a pause and provide just-in-time descriptions as an additional option to users. We could further consider humor (*e.g.*,indicated by laughs) and suspense (*e.g.*, open questions in presentations) to make sure we time the descriptions in a way that preserves enjoyable aspects of the presentation. We may also consider exporting such descriptions to a standalone video, but additional work would be required to make sure the undescribed element would be understandable from the audio alone (*e.g.*, a list element might not make sense without its heading).

8.3 How Can We Provide More Accessible Presentation Media?

Presentations are a key medium for informative content. While it is urgent and important to make presentations accessible, they can provide a challenging starting point in terms of trying to use Computer Vision or even novice work (*e.g.*, via Amazon Mechanical Turk) to describe their underlying visual content. The videos we selected in our study were accessible to a general audience, but many presentations require deep domain expertise (*e.g.*, a microscope slide of a cell, or a factory process diagram) to understand the media and prioritize visual content to create an informative description. Thus, we can not rely on fully automated solutions becoming available soon for some types of content. How should we move forward? Prior work suggests tools for helping presenters describe their slides at the time of the presentation [52] and creating accessible slides [34]. To add additional descriptions to images after the fact, Communitysourcing [31] or Learnersourcing [39] may be good approaches to collaboratively create better descriptions for media in the context of the learners or area experts that are also consuming the presentation content. Within our own community of expert presenters (*e.g.*, CHI, ASSETS) we may consider encouraging presenters to release their accessible slides. In the future, we could consider tools to collaboratively edit descriptions of the slide content (to gain interpretations from different points of view) and interfaces like Slidecho to help people consume the high-quality accessible slides more quickly.

9 CONCLUSION

Many presentation videos remain inaccessible to people with visual impairments because the visual content is not described by the speaker, and the slides are unavailable (or inaccessible). We present Slidecho, a system that makes presentation videos accessible to people with visual impairments. Slidecho's algorithmic pipeline extracts structured and accessible slides from the presentation and aligns the slide elements to the speaker's narration. Slidecho's interface instantiates new interactions that augment the plain video interface with synchronized slide information and audio notifications to alert users to undescribed elements. By allowing participants instant access to additional slide information as they consumed the presentation speech, Slidecho improved the accessibility of presentation videos.

ACKNOWLEDGMENTS

This work was supported by the Adobe Research Fellowship and the National Science Foundation. We also thank our study participants and reviewers for their thoughtful feedback.

REFERENCES

[1] 2018. Tesseract OCR, Google Open Source. https://opensource.google.com/projects/tesseract. Accessed 2018-03-28.
[2] 2020. American Council of the Blind, Audio Description Project, Guidelines for Audio Describers. https://www.acb.org/adp/guidelines.html.
[3] 2021. 3PlayMedia. https://www.3playmedia.com/
[4] 2021. ACM SIGCHI Channel. https://www.youtube.com/channel/UCEfEi-IMiB87UsxY3765P6w.
[5] 2021. TED talks. https://www.ted.com/talks.
[6] N. Reviers A. Remael and G. Vercauteren. [n.d.]. Pictures painted in Words: ADLab Audio Description Guidelines. https://dcmp.org/learn/captioningkey/624.
[7] Shadi Abou-Zahra and EOWG Participants. 2020. How to Make Your Presentations Accessible to All. https://www.w3.org/WAI/teach-advocate/accessible-presentations/#preparing-slides-and-projected-material-speakers.
[8] Tania Acosta, Patricia Acosta-Vargas, Jose Zambrano-Miranda, and Sergio Lujan-Mora. 2020. Web Accessibility Evaluation of Videos Published on YouTube by Worldwide Top-Ranking Universities. *IEEE Access* 8 (2020), 110994–111011.
[9] Don Adjeroh, MC Lee, Nagamani Banda, and Uma Kandaswamy. 2009. Adaptive edge-oriented shot boundary detection. *EURASIP Journal on Image and Video Processing* 2009 (2009), 1–13.
[10] Roqayah Ajaj. 2020. Navigating the World of Higher Education as a Blind or Visually Impaired Student: Unequal Opportunities for Academic Success. (2020).
[11] Edward C Bell and Arielle M Silverman. 2019. Access to math and science content for youth who are blind or visually impaired. (2019).
[12] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. 2006. WebInSight: making web images accessible. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*. 181–188.
[13] Arijit Biswas, Ankit Gandhi, and Om Deshmukh. 2015. Mmtoc: A multimodal method for table of content creation in educational videos. In *Proceedings of the 23rd ACM international conference on Multimedia*. 621–630.
[14] 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.
[15] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* (2008).
[16] Paolo Cardini. 2012. Forget multitasking, try monotasking. https://www.ted.com/talks/paolo_cardini_forget_multitasking_try_monotasking.
[17] Matt Cutts. 2011. Try something new for 30 days. https://www.ted.com/talks/matt_cutts_try_something_new_for_30_days.
[18] DCMP. 2020. Description Key. https://dcmp.org/learn/captioningkey/624.
[19] Eleanor Dickson, Chelcie Juliet Rowell, and Yasmeen L. Shorish. 2016. Guide to Creating Accessible Presentations. https://www.diglib.org/dlf-events/2016forum/guide-to-creating-accessible-presentations/.
[20] Vocal Eyes. 2018. Making your conference presentation more accessible to blind and partially sighted people. https://vocaleyes.co.uk/services/resources/guidelines-for-making-your-conference-presentation-more-accessible-to-blind-and-partially-sighted-people/.
[21] L Gagnon, C Chapdelaine, D Byrns, S Foucher, M Héritier, and V Gupta. 2010. Computer-Assisted System for Videodescription Scripting. In *Proceedings of Computer Vision Application for Visually-Impaired (CVAVI), a satellite workshop of CVPR*.
[22] Langis Gagnon, Samuel Foucher, Maguelonne Heritier, Marc Lalonde, David Byrns, Claude Chapdelaine, James Turner, Suzanne Mathieu, Denis Laurendeau, Nath Tan Nguyen, et al. 2009. Towards computer-vision software tools to increase production and accessibility of video description for people with vision loss. *Universal Access in the Information Society* 8, 3 (2009), 199–218.
[23] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey P Bigham. 2020. Making GIFs Accessible. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 367–376.
[24] 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.
[25] 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*. ACM, New York, NY, USA, 1–12.
[26] A Godfrey, R Jonathan, and M Theodore Loots. 2015. Advice from blind teachers on how to teach statistics to blind students. *Journal of Statistics Education* 23, 3 (2015).
[27] Google. 2021. Google Video Intelligence API. https://cloud.google.com/video-intelligence/.

[28] 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.

[29] Tessai Hayama, Hidetsugu Nanba, and Susumu Kunifuji. 2008. Structure extraction from presentation slide information. In *Pacific Rim International Conference on Artificial Intelligence*. Springer, 678–687.

[30] David Hayden, Dirk Colbry, John A Black Jr, and Sethuraman Panchanathan. 2008. Note-taker: enabling students who are legally blind to take notes in class. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. 81–88.

[31] Kurtis Heimerl, Brian Gawalt, Kuang Chen, Tapan Parikh, and Björn Hartmann. 2012. CommunitySourcing: engaging local crowds to perform expert work via physical kiosks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1539–1548.

[32] HHS. 2020. Provisional Guidance for Audio Description (AD). https://www.hhs.gov/web/section-508/making-files-accessible/accessible-audio-description/index.html.

[33] The Smith-Kettlewell Eye Research Institute. [n.d.]. YouDescribe. https://youdescribe.org/

[34] Tatsuya Ishihara, Hironobu Takagi, Takashi Itoh, and Chieko Asakawa. 2006. Analyzing visual layout for a non-visual presentation-document interface. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 165–172.

[35] Dhruv Jain, Venkatesh Potluri, and Ather Sharif. 2020. Navigating Graduate School with a Disability. In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–11.

[36] Gareth JF Jones and Richard J Edens. 2002. Automated alignment and annotation of audio-visual presentations. In *International Conference on Theory and Practice of Digital Libraries*. Springer, Springer, New York, NY, USA, 276–291.

[37] Hyeungshik Jung, Hijung Valentina Shin, and Juho Kim. 2018. Dynamicslide: Reference-based interaction techniques for slide-based lecture videos. In *The 31st Annual ACM Symposium on User Interface Software and Technology Adjunct Proceedings*. ACM, New York, NY, USA, 23–25.

[38] Min-Yen Kan. 2007. SlideSeer: A digital library of aligned document and presentation pairs. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*. ACM/IEEE, New York, NY, USA, 81–90.

[39] Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of CHI*. ACM, 4017–4026.

[40] Richard E Ladner and Kyle Rector. 2017. Making your presentation accessible. *interactions* 24, 4 (2017), 56–59.

[41] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, Vol. 10. Soviet Union.

[42] 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 (To Appear)*. ACM, New York, NY, USA, 1–4.

[43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692* (2019).

[44] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics.*

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

[46] Microsoft. 2021. Make your PowerPoint presentations accessible to people with disabilities. https://support.microsoft.com/en-us/topic/make-your-powerpoint-presentations-accessible-to-people-with-disabilities-6f7772b2-2f33-4bd2-8ca7-dae3b2b3ef25.

[47] 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*. 1–4.

[48] Bureau of Internet Accessibility. 2018. Tips for Making Your Presentations Accessible. https://www.boia.org/blog/tips-for-making-your-presentations-accessible.

[49] 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.

[50] Sujeath Pareddy, Anhong Guo, and Jeffrey P Bigham. 2019. X-Ray: Screenshot Accessibility via Embedded Metadata. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 389–395.

[51] 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* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864

[52] Yi-Hao Peng, JiWoong Jang, Jeffrey P Bigham, and Amy Pavel. 2021. Say It All: Feedback for Improving Non-Visual Presentation Accessibility. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[53] M. R. Rahman, S. Shah, and J. Subhlok. 2020. Visual Summarization of Lecture Video Segments for Enhanced Navigation. In *2020 IEEE International Symposium on Multimedia (ISM)*. 154–157. https://doi.org/10.1109/ISM.2020.00033

[54] Allison Ravenhall. 2018. Inclusive Design For Accessible Presentations. https://www.smashingmagazine.com/2018/11/inclusive-design-accessible-presentations/.

[55] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. https://arxiv.org/abs/1908.10084

[56] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3202–3212.

[57] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.

[58] Elliot Salisbury, Ece Kamar, and Meredith Morris. 2017. Toward scalable social alt text: Conversational crowdsourcing as a tool for refining vision-to-language technology for the blind. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.

[59] José Francisco Saray Villamizar, Benoît Encelle, Yannick Prié, and Pierre-Antoine Champin. 2011. An adaptive videos enrichment system based on decision trees for people with sensory disabilities. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. 1–4.

[60] Daisuke Sato, Hironobu Takagi, and Chieko Asakawa. 2006. Accessibility evaluation based on machine learning technique. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*. 253–254.

[61] Derek Sivers. 2010. Keep your goals to yourself. https://www.ted.com/talks/derek_sivers_keep_your_goals_to_yourself.

[62] Ranjini Swaminathan, Michael E Thompson, Sandiway Fong, Alon Efrat, Arnon Amir, and Kobus Barnard. 2010. Improving and aligning speech with presentation slides. In *2010 20th International Conference on Pattern Recognition*. IEEE, IEEE, New York, NY, USA, 3280–3283.

[63] Terrill Thompson. 2009. Audio Description and the JW FLV Player. https://terrillthompson.com/11.

[64] Julian Treasure. 2014. How to speak so that people want to listen. https://www.ted.com/talks/julian_treasure_how_to_speak_so_that_people_want_to_listen.

[65] Shoko Tsujimura, Kazumasa Yamamoto, and Seiichi Nakagawa. 2017. Automatic Explanation Spot Estimation Method Targeted at Text and Figures in Lecture Slides.. In *INTERSPEECH*. ISCA, René, CARRÉ, France, 2764–2768.

[66] Qiyam Tung, Ranjini Swaminathan, Alon Efrat, and Kobus Barnard. 2011. Expanding the point: automatic enlargement of presentation video elements. In *Proceedings of the 19th ACM international conference on Multimedia*. ACM, New York, NY, USA, 961–964.

[67] World Blind Union. 2012. Guidelines created by the World Blind Union (WBU) on how to make the use of PowerPoint and other visual presentations accessible. https://www.ifla.org/publications/guidelines-created-by-the-world-blind-union-wbu-on-how-to-make-the-use-of-powerpoint-an.

[68] GMCTL UofS. 2013. Learning Outcomes. https://www.youtube.com/watch?v=2q-wPQhpIkQ.

[69] Luis von Ahn. 2011. Massive-scale Online Collaboration. https://www.ted.com/talks/luis_von_ahn_massive_scale_online_collaboration.

[70] Yujia Wang, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, and Lap-Fai Yu. 2021. Toward Automatic Audio Description Generation for Accessible Videos. (2021).

[71] Wikipedia. 2021. Bullet. https://en.wikipedia.org/wiki/Bullet_(typography).

[72] Wang Xiangyu, Subramanian Ramanathan, and Mohan Kankanhalli. 2009. A robust framework for aligning lecture slides with video. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, IEEE, New York, NY, USA, 249–252.

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

[74] Baoquan Zhao, Shujin Lin, Xiaonan Luo, Songhua Xu, and Ruomei Wang. 2017. A novel system for visual navigation of educational videos using multimodal cues. In *Proceedings of the 25th ACM international conference on Multimedia*. 1680–1688.