Diffscriber: Describing Visual Design Changes to Support Mixed-Ability Collaborative Presentation Authoring

Yi-Hao Peng
yihaop@cs.cmu.edu
Carnegie Mellon University

Jason Wu
jsonwu@cmu.edu
Carnegie Mellon University

Jeffrey P. Bigham
jbigham@cs.cmu.edu
Carnegie Mellon University

Amy Pavel
apavel@cs.utexas.edu
University of Texas, Austin

ABSTRACT

Visual slide-based presentations are ubiquitous, yet slide authoring tools are largely inaccessible to people who are blind or visually impaired (BVI). When authoring presentations, the 9 BVI presenters in our formative study usually work with sighted collaborators to produce visual slides based on the text content they produce. While BVI presenters valued collaborators' visual design skill, the collaborators often felt they could not fully review and provide feedback on the visual changes that were made. We present Diffscriber, a system that identifies and describes changes to a slide's content, layout, and style for presentation authoring. Using our system, BVI presentation authors can efficiently review changes to their presentation by navigating either a summary of high-level changes or individual slide elements. To learn more about changes of interest, presenters can use a generated change hierarchy to navigate to lower-level change details and element styles. BVI presenters using Diffscriber were able to identify slide design changes and provide feedback more easily as compared to using only the slides alone. More broadly, Diffscriber illustrates how advances in detecting and describing visual differences can improve mixed-ability collaboration.

CCS CONCEPTS

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

KEYWORDS

Accessibility; Presentation; Slides; Change descriptions and captioning; Visual design, Multimedia creation; Authoring tools

ACM Reference Format:
Yi-Hao Peng, Jason Wu, Jeffrey P. Bigham, and Amy Pavel. 2022. Diffscriber: Describing Visual Design Changes to Support Mixed-Ability Collaborative Presentation Authoring. In *ACM Symposium on User Interface Software and Technology (UIST '22), October 29–November 2, 2022, Bend, OR, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3526113.3545637

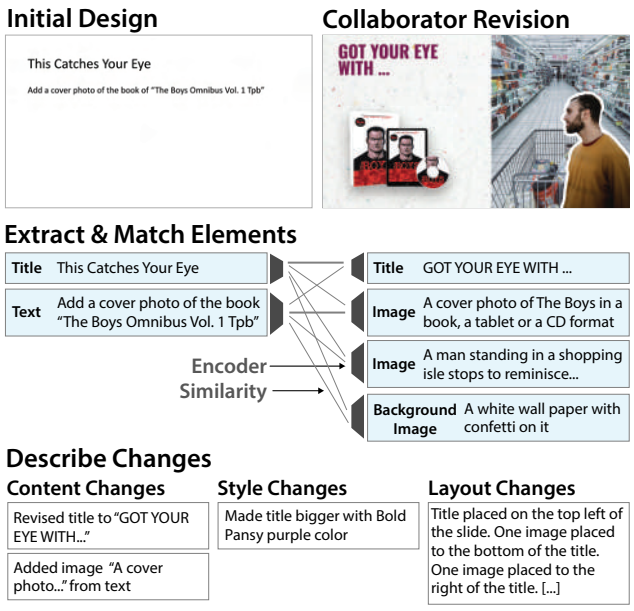

Figure 1: Diffscriber supports mixed-ability collaboration for BVI presenters and their collaborators. In particular, Diffscriber enables BVI presenters to better understand revisions to their presentations by describing changes.

1 INTRODUCTION

Slide-based presentations are ubiquitous and expected in professional and educational environments. As a result, blind and visually impaired presenters often need to author and deliver visual slide presentations. However, the industry-standard tools used to author slide presentations (*e.g.*, Microsoft PowerPoint, Google Slides, and Apple's Keynote) remain largely inaccessible to blind and visually impaired (BVI) presenters performing authoring tasks using screen reader software [35]. While current software can read out slide content if the slides are made in an accessible way, including text and image alt text, it is difficult for screen reader users to edit slides and evaluate the resulting visual results. At best, screen readers can relay unintuitive low-level descriptions of visual layout information (*e.g.*, the (x,y) locations of different visual elements).

Given these known challenges, we conducted a formative study with nine BVI presenters to understand how they currently author slide presentations. The BVI presenters in our study primarily author the content of their slides using text (either a document or a simple slide template), then hand off the text content to a sighted collaborator or hired assistant to author the slides. The collaborator then changes the: *content* of the slides (*e.g.*, by adding relevant images, rewording text to fit, or removing/adding content), and the *style* of the slides (*e.g.*, changing the layout of the information, adding text styles and decorative theme elements). While BVI presenters appreciated the expertise and efficiency of

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

ACM ISBN 978-1-4503-9320-1/22/10.
https://doi.org/10.1145/3526113.3545637

their collaborators, the limited communication after presentation changes made it difficult to understand the resulting design that they would present (*e.g.,* spatial layout of title and other elements on screen) and provided limited agency over the design process.

We present Diffscriber to support BVI authors' presentation authoring. Diffscriber provides BVI presenters access to the content and style changes made to their presentations, which directly supports collaboration with sighted collaborators. Using the underlying slide structure, Diffscriber first identifies correspondences between the content on the BVI presenters' original slides (either a text-based presentation or structured text document) and the content on the collaborator's revised slides. Then, Diffscriber generates descriptions of changes that occurred between the two documents including: revisions to the content (*e.g.*, addition, removal and replacement) and revisions to the style (*e.g.*, layout, and individual element properties). Diffscriber's interface lets users flexibly navigate between high-level change description summaries, individual change descriptions, individual slide element descriptions, and the corresponding editable slide elements. Using Diffscriber presenters can prioritize elements they may want to edit, and gain understanding of content and style revisions to provide feedback to their sighted collaborators.

We evaluated Diffscriber with BVI presentation authors reviewing changes to two professionally-redesigned slide presentations. We designed the evaluation scenario to mimic the primary form of collaboration identified in the formative study — asynchronous collaboration between a BVI presenter and sighted slide author. Using Diffscriber, presenters accurately identified more changes made to the presentation compared to using accessible slides alone (28.83 vs. 10.33 changes identified). Presenters also provided more feedback on the revised slides using than when using accessible slides alone (2.83 vs. 0.83 feedback provided). For slide authoring, participants unanimously preferred Diffscriber to accessible slides alone and to their prior experiences with PowerPoint. Participants expressed enthusiasm about using Diffscriber in the future to assess slide changes, communicate feedback, author slides, and learn from layout and style changes made during the revision process to inform future authoring tasks.

In summary, we contribute: a formative study of BVI presentation authors' collaborative design process, the Diffscriber system for identifying and describing changes to visual slide designs, and a user study with BVI presenters evaluating Diffscriber's change descriptions for visual slide designs.

2 RELATED WORK

As our work seeks to improve BVI presenter agency in the collaborative authoring process, our work relates to prior work that explores: the accessibility of presentations and slides, the accessibility of visual design authoring tools, and prior methods for describing images and their changes.

2.1 Presentation Accessibility

In slide presentations, authors use slides to reinforce concepts and provide visual aids. Presentations and their slides can be inaccessible to blind and visually impaired audience members when presenters do not describe their verbally slides [27], and if they do not make any shared slides such as PowerPoint slides annotated with alt text describing the visual content in an appropriate read order [19]. As accessible slides are important but time consuming to create, Ishihara et al. [13] and Sato et al. [34] created tools to make diagrams more accessible when accessing PowerPoint slides, and to turn PowerPoint slides into HTML. Peng et al. [26] created a method to detect and describe text and image elements in slide videos such that BVI audience members could flexibly access more information about presentation elements that were not described by the presenter. While such tools might be used to help blind presenters understand their slides, the tools have primarily been designed at making slides accessible for consumption rather than editing. That is, they make the content accessible, but they do not necessarily share anything about the slide style or layout if it is not clearly necessary (as recommended by accessibility guidelines on content accessibility [2]). In this project, we examine how to improve the accessibility of the slide presentation authoring process by surfacing both content and style changes.

2.2 Accessibility of Authoring Tools

Prior work has explored how blind and visually impaired content creators author visual designs including drawings [15], documents [45], maps [39], artboards [35], websites [17] and videos [36]. Such work has identified that while BVI content creators would like to author visual content (*e.g.*, for work, education and hobbies), it remains difficult to assess the current state of their work due to iterative changes to the visual content as well as lack of interpretable feedback on visual attributes [35]. Thus, prior work has explored creating physical 2.5D printed prototypes of digitally created visual designs like website layouts [17] and maps [39]. but each design revision requires reprinting the design to gain information about its state making the approach likely best suited for prototypes that require few iterations. Other work has explored shape displays as a method to communicate visual state, for instance when observing and designing 3D objects [40]. However, shape changing displays are currently low resolution such that presenting text or braille content on shape changing displays (as is required for text-dense designs) would be challenging.

Rather than requiring blind and visually impaired authors to create both the content and style of their visual designs alone, prior work has explored mixed-ability collaboration in the context of document editing [3]. Mixed ability collaboration supports an "interdependence" rather than "independence" approach to making content authoring accessible [1]. Specifically, it emphasizes social support as playing an important role to collectively create access with people. For collaborative document editing, mixed-ability collaborators value working together, but experience many challenges with the accessibility of the collaborative tool itself such as keeping track of the document state [3]. Compared to documents that feature primarily text, slides are visually rich in terms of their image content, layouts and styles such slides will likely be more challenging to edit synchronously. We explore change descriptions as an approach to support mixed-ability collaboration on visually rich designs.

To help presenters author well-designed presentations more intuitively given their content alone, prior work has explored directly

creating slides from text-based specifications (*e.g.*, Markdown or LaTeX-based slide generator), automatically suggesting alternative layouts for a given design (*e.g.*, DesignScape [23, 24], PowerPoint Designer [37]), and plain text descriptions (*e.g.*, Text2Slide [47], Doc2Slide [9, 42]). However, these tools, designed primarily to make authoring more efficient for sighted authors, rely on the presentation author to decide between multiple options [23, 24, 37] or verify the results of the automated system [9, 42]. In addition, automated tools may lack the expertise of sighted collaborators who in many cases have knowledge about the work itself and the visual design patterns in the domain. While our presentation change descriptions may benefit presenters using automated tools (*e.g.*, by describing the before and after changes), we aim to support blind and visually impaired presenters in their current presentation process.

2.3 Describing Visuals and their Changes

Image descriptions are important for digital accessibility. Prior work introduced approaches to generate image descriptions including crowdsourcing [32], reverse image search [11], and using a combination of methods [10]. With pre-trained image-text models [28], describing visuals automatically with AI/ML has become more feasible and has been deployed for Facebook's automatic alt-text [46] and Microsoft's Seeing AI [20]. However, prior studies have found the auto-generated descriptions are error-prone and impact the user's understanding of images [33]. Even when generated descriptions are corrected, details like color are left out, making it difficult for BVI people to fully understand the image [41, 49]. Beyond generating image captions, prior research in computer vision explores ways to generate descriptions of visual differences between two images including: detecting the appearance of people, cars, or birds in subsequent frames [8, 14, 25], or the image pairs from before and after PhotoShop editing [44]. While prior work compares photographs or renderings, we identify and describe visual changes in structured designs that feature text. Prior work also detected changes in slides to visualize slide version changes statically [7] or dynamically [6]. As such work was not designed for accessible authoring, the feedback visualizes the changes instead of describing them, and detects high-level changes rather than fine-grained changes to slide elements.

2.4 Mixed-Ability Collaboration

Supporting mixed-ability collaboration has been a critical part of inclusive workspace and education for people with diverse abilities. Prior work proposed the frameworks to articulate important awareness-related information in both synchronous [12] and asynchronous [43] settings of collaborative authoring (e.g, who, what, where, how, and why the changes were made). Based on these frameworks, recent works investigated mixed-ability collaborative document editing between blind and sighted writers [3, 5], and proposed tools [4, 16] to improve the collaborative awareness (e.g., by better indicating who has changed what). Our work extends the support of mixed-ability collaboration to the domain of visual design, and focuses on the scenario where the collaboration (between blind presentation authors and their assistants) is taken place to make inaccessible authoring process (e.g., visual style and layout changes) accessible.

ID	Gender	Age	Level of Vision	# Years	Present Frequency
P1	F	58	Totally blind	Since birth	Twice a week
P2	M	21	Light perception	Since birth	Once a month
P3	M	48	Totally blind	Since birth	Four times a week
P4	F	27	Light perception	Since birth	Once a month
P5	M	56	Totally blind	Since birth	Once every three months
P6	M	30	Low vision	Since birth	Twice a month
P7	F	31	Low vision	Since age 21	Once a week
P8	F	58	Light perception	Since birth	Once a month
P9	F	31	Totally blind	Since age 16	Once a month

Table 1: Participant's demographic information in our formative study, including gender, age, level of vision and years at the designated level of vision, and their frequency to present.

3 BACKGROUND AND FORMATIVE STUDY

While prior work has explored the current limitations for blind users attempting to lay out elements on an art board (e.g., a slide) [35], no prior work has investigated how blind people author slide presentations despite these limitations. To uncover current strategies for presentation authoring, we conducted remote semi-structured interviews with presenters who are blind or have a vision impairment. We also analyzed existing pre-revision and post-revision slides provided by two presenters to identify common changes.

3.1 Procedure

We recruited 9 people (5 female, 4 male, age=21-58) who are blind or have a visual impairment who had experience authoring and giving slide-based presentations (Table 1). We specified that the presentation authoring experience may include independent presentation authoring or collaborative presentation authoring. Each interview was 1-hour long, and we asked participants to share: their process for creating slide presentations (a recent example, and the general process), the challenges that they encountered during the authoring process, and their solutions for overcoming the challenges they encountered. We additionally asked about if they had any prior experience collaborating on presentations and their process for collaboration, and the difficulties as well as the corresponding solutions throughout the experiences. Finally, we asked whether they have used slide template to create slides by themselves and if they have used any of other alternative slide authoring tools such as structure-text (e.g., Markdown, HTML) or slide generation framework (e.g., Reveal.js, Marp). Participants were compensated $30. We recorded each Zoom session including the audio for the interview questions, analyzed the interview by grouping the interview notes into themes, and returning to the interviews to extract specific quotes and synthesize feedback by themes.

3.2 Findings

All participants reported that they collaborated on slides rather than authoring slides alone, even if they had prior experience independently authoring slides.

3.2.1 How did presenters author slides? The presenter composed the content of the presentation including text that should go on each slide and instructions for visuals to add (e.g., a screenshot of a specific website, a chart given some data). 7 presenters only shared the content using a Word document with similar structure to the

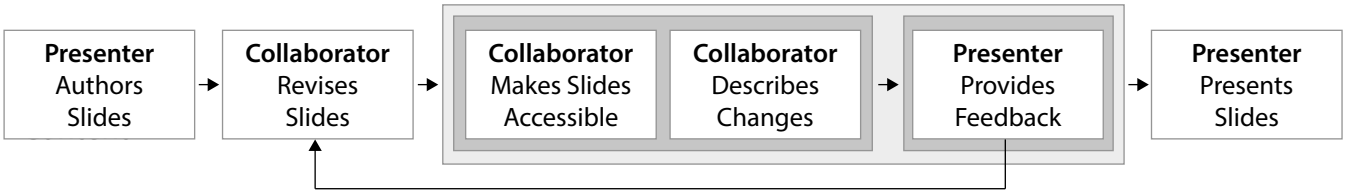

Figure 2: When collaboratively authoring slides all participants authored slides (*e.g.*, via a slide template or structured text) and handed them to the collaborator for revision. Some participants directly presented the slides, but others received additional information from the collaborator about the slides either by making slides accessible, describing changes, or both. After receiving this information, some presenters went through a round of iteration, while others presented the slides directly.

final slides (*e.g.*, included titles and bullet list texts as slide contents). The other 2 presenters authored their content in PowerPoint with a default slide template, which followed a similar content structure including the titles followed with bullet list text. The presenter shared their base presentation content with a sighted collaborator such as a work colleague (5 participants), a professional assistant (3 participants), or a friend (4 participants). The sighted collaborator then authored the visual design of the presentation slides.

After the first draft, most presenters (6 of 9) took the slides and directly presented them without further input. In particular, presenters mentioned that they would not communicate and collaborate further on presentations that were team presentations, or when there was little time for iteration. Other presenters communicated further about the slides with the collaborator. In these cases the presenter either: made the slides accessible, talked through the changes that they made to the slides, or talked through the accessible slides (both). The last case was rare as it required most effort and occurred for people working in accessibility-related organizations with the assistant hired by through an agency. In most cases, participants reported that collaborators chose to talk through the slides as it didn't require prior preparation. In other cases, the collaborator made the slides screen-reader accessible so the blind presenter could review the content and the presenter would look through the slides asynchronously and give feedback via email on what content they would like to change.

The last stage of the process was to revise the presentation based on the edited information. Participants again expressed that given current authoring tools they would ask their collaborators to do the edits as they were not confident they could edit the slides without affecting their visual design. For participants who knew how to change the text content in slide software, they would make the changes then ask people to check the results to confirm the layout and design were still coherent.

3.2.2 Current Challenges and Solutions. Participants expressed that some collaborators who initially agreed to help in the past might fail to provide consistently and thoroughly once starting the authoring process (e.g., in school). In addition, when blind presenters received a summary of changes through a verbal description, they were only able to participate in the collaboration based on the information provided. However, the assistant would need to keep track of and remember to describe each edit that they made to the content, layout, and style (such that many changes may not be communicated). In practice, participants mentioned that collaborators would focus on the content changes when prioritizing changes to describe in limited time. In addition, participants and their collaborators occasionally disagreed in terms of what changes would be important to describe or not, such that blind presenters would later find out about some undesired changes that were not mentioned by the collaborator. For instance, P8 described an experience where an audience member told them that the font is too small on the slides after the talk, but the font size was not described during the meeting with collaborator.

In other cases such that slide was the only medium for communication, participants did not always know if their collaborators were experts in: the presentation tool, visual design, or accessibility. The expertise in accessibility at times impacted participants' ability to understand the content on their slides. Accessibility professionals would also sometimes forget to add alt-text or correct read order after editing, or they would accidentally apply a slide template that introduced inaccessible elements (P9). Even when the slide was fully screen-reader accessible, the slide information focused on the content (*e.g.*, text and image alt text only), as other visual characteristics like position or color were important but difficult to access, interpret or compare. Presenters reported resolving these concerns by asking another person for a second opinion.

3.2.3 Types of slide changes. Participants shared examples of content, slide and layout changes during the interview that took place during collaboration. Two participants additionally shared with us their original slides or content specification as well as the revised slides for two presentations they had recently authored (total = 37 slides). We analyzed the before and after slide examples and identified the types of design changes that occurred between the two revisions (Table 2):

Content changes consisted of adding, replacing, or removing new slide elements. Content additions included adding text (*e.g.*, a title for a slide that did not have one), adding images either relevant to the content (*e.g.*, a presenter portrait) or irrelevant to the content (*e.g.*, an abstract squiggle graphic on the edge of the slide), and adding shapes. Content replacements included revising text (*e.g.* replacing "correct or mitigate" with "take steps to correct or mitigate accessibility issues encountered"), replacing a text request with an image (*e.g.*, "Screenshot of website accessibility checker report" replaced with an image of the accessibility checker report), and

Figure 3: Interface Layout Overview. The change descriptions interface augments the Google Slides' *editor view* with changes. As the author compares the changes between the previous and current slides, the customized view shows the elements on slides, the changes that have been made to the slides including layout, content and the style changes.

replacing text without a request with an image. Replacement relationships between elements were often one-to-many – for example, the collaborator split one bullet point into multiple horizontal list items — or even many to many. Finally, text was removed (*e.g.*, removing 1 of 5 bullet points). Observed content changes did not include image or shape removal or replacement as the presenter's initial slides were text-only, as also reported by all participants in the formative study.

Style changes consisted of changing styles for existing elements on the slide or defining style for a new element. The text style including typeface, font weight, font color, font style, etc. changed for all text elements between the original and revised slide. We also observed new shape elements (e.g., circles, squares to be positioned under text) with properties defined for fill and stroke.

Layout changes including changing the relative positioning of the title with respect to the rest of the body (*e.g.*, moving the title from the top to the left side), and changing the body layout (*e.g.*, changing a bullet list to a left to right list or row/column layout). We used title and list as basis to detect changes as they are the common structures existed in the slides,

We use the observed changes in BVI presenters' presentations to guide the changes we support in our system.

4 SYSTEM

Diffscriber enables BVI presentation authors to efficiently understand visual edits. Our system provides two key ways for authors to access slide information: (1) the *slide content list* (Figure 5) that allows authors to read the content of each slide element (*e.g.*, the text content or alt text for an image), the metadata for the element (*e.g.*, position and style), and any changes made to that element between the prior and current slide, and (2) the *change description list* that allows authors to read through a summary of changes, access lower-level change descriptions, and navigate to any changed element to obtain more information (Figure 4). We implemented our system as an extension for Google Slides and directly augmented the information on the existing authoring interface (Figure 3). To produce descriptions of slide changes, Diffscriber takes as input two Google Slides URLs, one before and one after the slide revisions. Alternatively, when editing a single Google Slides URL, users can

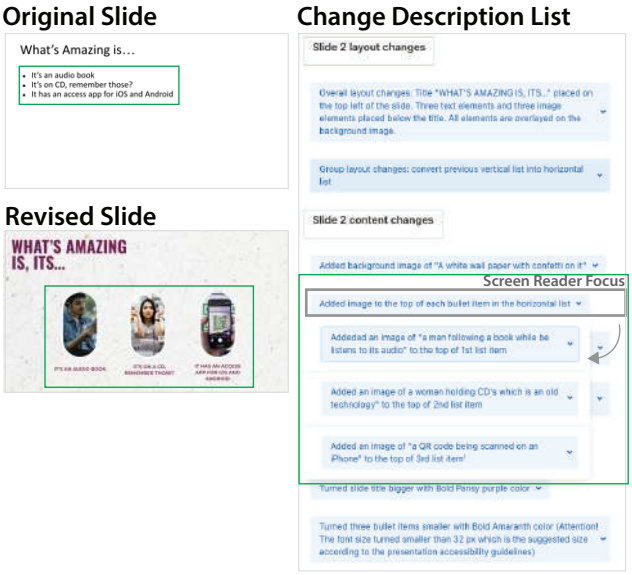

Figure 4: Layout, content and style changes (bottom two changes are style changes) for comparing the original slide and the revised slide. The body change (green box on the original and revised slide) is described in the layout changes as well as the content changes. In the content changes for this body element (green box), the addition of images is summarized in a "Change Summary". Readers can expand the summary to read individual changes.

Type	Operation	Changes
Content	Add	Add text
		Add content-relevant image
		Add content-irrelevant image
		Add shape
	Replace	Revise text
		Replace text request with image
		Replace text with image
	Remove	Remove text
Style	Change/New	Text: font, weight, size, color, background color
		Shape: type, fill, stroke, stroke weight
Layout	Change	Title to body layout
		Body layout

Table 2: Observed changes in two presenters' before and after slide. Observed changes did not include replacement or removal of images and shapes as the presenter's first draft's included only text.

select "save" to save a snapshot, and later select "compare" to compare the current version to the previously saved snapshot. While our design is applicable to many presentation tools, we used Google Slides due to its broad availability across operation systems and extensibility for development.

Slide content list: The slide content list displays all the elements in each slide (Figure 5). The elements are sorted based on their

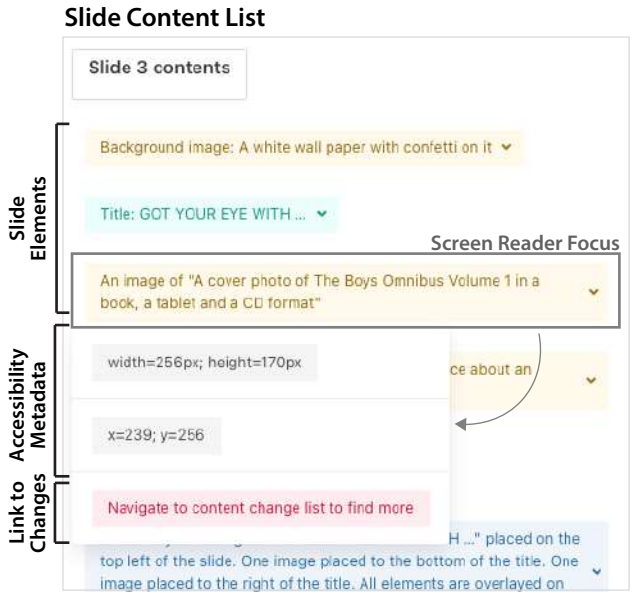

Figure 5: The Slide Content List enables BVI authors to navigate the content of each slide element and access position and style information about each element on demand. In addition, they can jump to where the element appears in the changes to find out how the element has changed since the last revision. This slide content list displays content for the slide depicted in Figure 3

Figure 6: The Change Description List enables BVI authors to navigate each change. For each change, they can access additional information about the relative position of the element with respect to the nearest neighbor element, and get more information about that element. All changes are also linked to the elements in the Slide Content List so that authors can get more information about the element metadata.

distance from upper left-hand corner (smallest distance first). To obtain the elements for the slide content list, we retrieved the underlying slide element information using the public Google Slides API. However, access to slide element metadata is limited with the API, so we also used a selenium-based web crawler to extract additional information from the slide Document Object Model (DOM). We combined the element information from both sources based on the unique element ID.e To allow users flexibly review the visual properties of the elements on-demand, each element is an interactive component. Specifically, we displayed the slide content (text content, or image descriptions extracted from from assistant-provided alt-text or auto-generated image captions [18]) as the first level of the hierarchy and other attributes (e.g., position, size) as the second layer of the hierarchy (Figure 5). All changes are linked to the Slide Content List, such that if there has been a change to an element, the user can additionally navigate to the associated changes using the "Navigate to content change list to find more" button.

Change description list: The change description list displays detected changes in three separate sections: layout, content and style changes (Figure 4). Similar to the slide content list's slide elements, each change description element is also an interactive hierarchical component. Using the change description list, authors can first browse summarized descriptions that all received a similar change (*e.g.*, "Added image to the top of each bullet item in the horizontal list"). Then, authors can browse the lower level changes (*e.g.*, Added image of "a QR code being scanned on an iPhone" to the top of the 3rd list item") (Figure 4). For any individual change, authors can click on the element to learn about relative positioning of the changed element (*e.g.*, "Added image to the bottom of the slide title") then flexibly navigate to more information about the related element (*e.g.*, "Slide title: 'GOT YOUR EYE WITH ..."') (Figure 6). Similar to the slide content list, each slide change is linked back to its corresponding element. For example, an author could flexibly learn more about the image by clicking on the linked slide element and listening to the accessibility metadata (Figure 5). For each change section, we ordered the change types based on estimated importance. For example, for text changes, as BVI presenters authored the slide content, we considered the text changes of "removal" and "addition" to be relatively more important than "replacement" changes. For replacement changes, we ranked replacing text with an image higher than replacing text with text as most text-text replacements represented small rewording changes. Within each change type, we then sorted the elements with distance-based read order similarly as slide content list.

Edit mode: With identified changes, BVI presentation authors can turn on the edit mode to modify the slides. Considering the primary edits that blind presenters usually perform, our system currently supports the editing text content. By activating the edit mode, users can directly edit the content in our interface and updates the slides by clicking the update button. Users can optionally use the navigate button in the element's second level to find the corresponding text box in the slide editor and revise the content directly.

5 ALGORITHMIC METHODS

Diffscriber detects and describes changes between slides for BVI presenters by *(i)* establishing correspondence between slide content, *(ii)* detecting relevant changes to content and appearance, then *(iii)* producing a hierarchically-summarized description.

5.1 Correspondence

The first stage of our system establishes a correspondence between the elements on two slides. We segment each slide's content into individual elements, which we classify either as textual elements or visual elements (*e.g.,* figures, diagrams, images) based on the properties extracted from the public Google Slides API and the slide DOM.

Our approach to computing correspondence is informed by similar methods for predicting word-alignments in NLP [31]. Specifically, elements are featurized using a neural embedding model then compared with those from another source (*i.e.,* slide). For each original slide (with M elements) and edited slide (with N elements), we create a cost matrix $C \in \mathbb{R}^{MxN}$ where $C_{i,j}$ indicates the alignment score between element i on the original slide and element j on the edited slide. Similar to previous work [31], we use the cosine similarity score, which provides a normalized metric for alignment strength.

The cost matrix C is used to extract likely correspondences by identifying element-element pairs that have high similarity relative to other candidates. Arg-max matching is a simple approach that produces a correspondence between two elements i and j if $C_{i,j} = \mathbf{max}(C_{i,:}) \wedge C_{i,j} = \mathbf{max}(C_{:,j})$. This produces a one-to-one mapping between elements on the source and target slide, since any correspondence must be the most likely for both elements. However, since it is possible for one element on the source slide to map to multiple on the edited slide (*e.g.,* revising a block of text into several bullet points), we use a variation of this approach called itermax [31], which repeatedly applies arg-max matching to generate multiple possible correspondences.

5.2 Edit Classification

We construct a classifier to detect the edits that we observed in our formative study (Table 2). We opt to use rule-based methods to detect content changes from the correspondence matrix and associated element metadata.

Content changes Content changes involve adding, replacing, or removing images or text. Adding new content that isn't grounded in the original content (*e.g.,* a decoration or content-irrelevant image) leads to "unmatched" items on the final slide. On the other hand, adding elements relevant to source material can be detected by searching for correspondences between content and textual content. Text revisions are detected using a similar strategy by locating correspondence between source and edited text elements. BVI presenters may also leave instructions in their slides for additional content *e.g.,* "place a photo of me here." Though our current system does not differentiate between instructions and textual content, we put the mapping between correspondent text and image element as unique class to generate text-image specific descriptions (e.g. add [image description] from [text]). Finally, we detect text removal or instructions that were not followed by identifying elements in the source slide without a correspondence match.

Style edits Style edits can indicate changed emphasis and importance. Since the original slides authored by BVI presenters include little style information, we focus primarily on detecting and extracting attributes on the edited slide (See Table 2 for a list of attributes). Elements on the edited slide with detected style attributes were automatically associated with the original element using their predicted correspondence match.

Layout changes Diffscriber supports two types of layout changes on the edited slide: *(i)* title to body layout and *(ii)* body layout. Title to body layout changes may occur when the slide assistant uses a non-vertical template (*e.g.* title text is positioned on the left and content is positioned on the right) to create a slide. To detect this change, we approximate the reading order of slide elements by ordering them top-down, left-to-right and check if the title element is first.

To detect layout changes in the body, we infer the grouping structure of elements such as bulleted lists. Groupings from the original slide are extracted from the HTML scraped from the presentation interface. From our observations, groupings in source content were always shown as bulleted lists, so we looked for items in the HTML that belonged to the same list node. Since groupings on edited slides can appear in different HTML nodes, we use the original as a starting point, then add any elements that had high similarity (based on similar style and positional alignment) to ones in the group and remove elements in the group that do not contain a matched correspondence. Once groupings from the original and edited slide are detected, we fit element locations to a grid (approximate number of rows and columns) by calculating average element size with respect to the bounding box of the group. As original presenter slides have minimal single-column layouts (as observed and reported in the formative study), we automatically signal a layout change if a group on the edited slide contains more than one column.

5.3 Description & Summarization

Each of our detectors produces a natural language description and additional metadata. Natural language descriptions are generated using pre-defined templates, which provides greater control and predictability compared to other text generation approaches [21].

We aim to provide summaries of the visual design information such that each component of the design description: (1) maximizes the amount of new important information, (2) minimizes redundant information. Presenters may optionally gain more information about each individual element (e.g., the visual edits) given our high level summaries.

(1) *High-level overview* - A summary of edits that are consistent between elements of the same group (e.g. bullet list), or a simplified description of an edit that replaces long descriptions of content with their structure
(2) *Fine-grained* - Comprehensive description of all detected edits where each edit is described using a sentence.

Table 3 shows the templates we used to generate natural language descriptions from detected edits. Finally, we use several strategies to

Type	Templates
Add	"added text ..."
	"added figure ..."
	"added figure ... from text ..."
Replace	"revised ... to ..."
	"title changed to ..."
Remove	"removed text ..."
	"removed image"
Layout	"layout of group ... changed ...'
	"layout of slide changed ..."
Style	"turn text to style of ...'

Table 3: Templates used to generate natural language descriptions from different types of edits.

further improve the quality of generated descriptions under certain conditions.

Location-based References: For changes that involved adding a new element to the slide, we also included a location indicator to help BVI presenters spatially reason about its placement. We identified the *anchor* element (*i.e.,* an element that is referenced by the original slide) that is closest to target and describe the target's position relative to the anchor.

Structural References: By default, textual elements are described using their content and visual elements are described with using alt-text. However, this can lead to long descriptions that are difficult to digest. If the user modified content that was originally in a detected group (*e.g.,* bulleted list), we generate alternative descriptions that reference their location in the group

Instructional References: Often, BVI presenters may include instructions in their materials that can reference visual content. If a correspondence is detected between originally-provided text and a visual, we incorporate it into our description.

6 TECHNICAL METHOD SELECTION

Since our system's performance depends heavily on the correspondences generated between the source and modified slide, we evaluated several approaches for representing and matching slide elements. The main purpose of our evaluation was to measure the performance of our system under realistic constraints that could be introduced in the slide-authoring process, such as missing alt-text.

6.1 Dataset

To our knowledge, there is no dataset of before-after presentations created by BVI presenters publicly available (due to the inaccessibility of current authoring software), and many of the people we asked discarded the original version of slides as it would not be useful for presenting. We contacted participants from our formative study to collect a dataset of 61 examples (pairs of slides), consisting of text-based slides (initially authored by them) and the modified versions (created by assistants). We received four pairs of presentations (original and edited) that had varying topics, content, and authoring. S3 is the longest presentation and includes several instances where BVI presenters left instructions in the source materials. S4 was mostly automatically generated using PowerPoint's design suggestion feature and did not contain many visual figures. Refer to our supplementary material for a more detailed overview of each presentation. For each example, we created a mapping between the original slide elements and edited slide elements.

Method	S1	S2	S3	S4	# Overall
ST + alt	0.68/0.88/0.77	0.81/0.89/0.85	0.69/0.83/0.76	0.85/0.96/0.9	0.75/0.88/0.81
ST + auto	0.67/0.88/0.76	0.8/0.89/0.84	0.64/0.81/0.71	0.87/0.96/0.91	0.73/0.87/0.79
CLIP + alt	0.68/0.84/0.75	0.78/0.84/0.81	0.75/0.83/0.79	0.9/0.96/0.93	0.78/0.86/0.81
CLIP + auto	0.65/0.81/0.72	0.78/0.850.81	0.72/0.8/0.76	0.88/0.96/0.91	0.76/0.84/0.8
CLIP	0.61/0.84/0.71	0.78/0.85/0.81	0.63/0.81/0.71	0.83/0.96/0.89	0.70/0.85/0.77

Table 4: Results of our technical evaluation on correspondence accuracy (Precision/Recall/F1 value). Overall, all element featurization methods performed at a similar level, however human-provided alt-text led to improvements on slides with more visual content (S3).

6.2 Method

We investigated different methods of featurizing slide elements that would allow us to compute semantic similarity between them and generate correspondence. Sentence transformers are machine learning models that were trained by associating semantically similar text (*e.g.,* paraphrase detection) and are useful for generating embeddings for variable-length text [30]. Text elements were directly fed into the model, and image elements were either featurized using provided pre-authored alt-text or the output of an off-the-shelf image captioning model [18]. We hypothesized two potential drawbacks to this approach: *(i)* the added complexity (*i.e.,* requires running two models when no alt-text is available), and *(ii)* some information might be "lost" when transferred from one modality to another. Thus, we also evaluated CLIP, multi-modal model that was trained to associate both image and text [29], *e.g.,* an image of a cat would have a similar representation to the text "cat." Using this model, we explored the previous conditions (alt-text and auto-generated captions) and also used the model's built-in image encoder.

We used each approach to featurize elements on the original and edited slides then generated a set of correspondences using our matching algorithm. To measure correspondence quality for each approach, we used definitions of precision, recall, and F-1 score from word-alignment evaluation [22], since they consider one-to-many mappings. Table 4 shows the results of our technical evaluation. Generally, recall was higher than precision, meaning that BVI presenters were likely to learn about the majority of edits that were performed. While lower precision may potentially lead to false-positive descriptions, in many cases, there is little difference in how the change is described. Given that all methods performed at a similar level, we chose to use the CLIP model since it performed well overall (and especially with alt-text) and introduced the last complexity into our pipeline due to its single-model architecture.

The vast majority of content on the original slides were textual and only a small number were instructions for adding images. S3 contained the most image references in our dataset, which may reveal greater differences between image featurization methods (automatic captioning vs image encoder). Thus, the presence of human-authored alt-text had the greatest benefit in this case, as automated methods are often unable to capture the semantic meaning or purpose behind figures. In general, we found that automated methods

could still produce accurate correspondences. Nevertheless, alt-text is still essential for generating informative and meaningful change descriptions, especially those that involve or otherwise reference image content.

7 USER STUDY

To assess the the efficacy of Diffscriber, we conducted a user study with 6 BVI presentation authors reviewing two sets of revised slides using two different interfaces: the full Diffscriber interface (Diffscriber), and the Diffscriber interface without change descriptions (Accessible Slides).

7.1 Method

Materials: We used two presentations, authored as first drafts by a BVI presenter, as our example presentations. One presentation was titled "Graphic Audio" and the presentation contained a short pitch for a website that features audiobooks with rich movie-like audio (the original presentation featured 10 slides with around 1-3 lines of text each). The second presentation was titled "Cloud Storage" and contained a workshop presentation about storage options for businesses (the original presentation featured 14 slides with 2-6 lines of text each). Both presentations were intended for a general audience. We anonymized the original slides and hired a professional slide designer on UpWork to create a slide revision. From these revised slides, we selected 5 slides per presentation to demonstrate a range of change types (*e.g.,* additions, revisions, removal, layout changes, style changes). We make the size of the sources of the cropped images as same as the cropped results to make the data consistent between visuals and actual structures. We also remove invisible elements such as the blank text box or transparent shape to cleanup the redundancy in the slide structure. We sent the original version of slides 3 days before the study (both Word and PowerPoint format) to let participants get familiar with the topics and the slides as if they are the prsentation authors.

Participants: We recruited 6 BVI presentation authors for around 1 hour long study on Zoom using mailing lists. Participants were ages 22-58 (2 female, 4 male) and all had experience consuming and authoring presentations (Table 5). We compensated participants $30 for completion of the study.

Procedure: We first asked a series of demographic and background questions about their experience consuming and authoring slide presentations. We then provided a short tutorial of our two interfaces (a slide from one of our previous collected slides with the topic about learning goals): the full Diffscriber interface (Diffscriber), and the Diffscriber interface without change descriptions (Accessible Slides). After the tutorial, we invited participants to continue exploring the tutorial slides or and ask questions. Next, participants reviewed the two different presentations, each with a different interface. We randomized the interface order and presentation order. For each presentation, we asked participants to identify and assess the quality of the revision and provide any feedback or questions for the professional slide designer for the next iteration. During the task we provided participants with both versions of the slides (the original BVI presentation author slides and the revised slides). We limited time on each slide pair to 5 minutes (25 minutes total per presentation). After each review task, we asked a series of

ID	Gender	Age	Level of Vision	# Years	Present Frequency
P1	F	28	Light perception	Since birth	Twice a month
P2	M	55	Light perception	Since 45	Once every two months
P3	M	22	Light perception	Since birth	Once a month
P4	F	58	Totally blind	Since age 20	Twice a month
P5	M	44	Light perception	Since age 30	Once a month
P6	M	56	Totally blind	Since birth	Once every three month

Table 5: Participant's demographic information in our user study, including gender, age, level of vision and years at the designated level of vision, and their frequency to present.

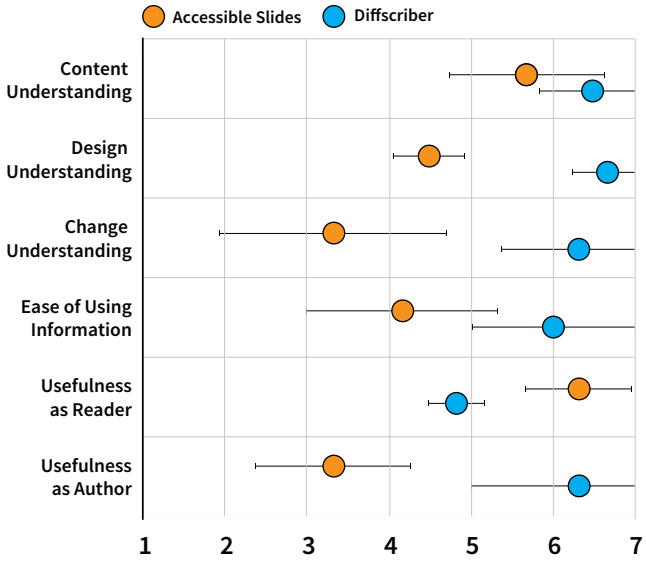

Figure 7: Participants rated Diffscriber (in blue on bottom) and Accessible Slides (in orange on top) from 1 (least) to 7 (most). Error bars display 95% confidence intervals.

Likert scale questions about how the interface supported them in understanding the slides (the content, the design, the changes), how useful the interface would be for slide tasks (consuming, authoring), and if interface helped them make use of the information provided. At the end of the study, we asked about the usefulness of different description types. Finally, we asked open-ended questions about how the interfaces compared to their current slide consumption and authoring process, and if and how participants might make use of either interface during their authoring process in the future. We audio and screen recorded each study.

7.2 Results

Identifying changes. Participants identified significantly more changes using Diffscriber ($\mu = 28.8$, $\sigma = 2.40$), than when using Accessible Slides ($\mu = 10.3$, $\sigma = 1.80$) ($t(10) = 12.11$; $p < 0.00001$). For assessing slide changes, all participants also preferred Diffscriber to Accessible Slides, and participants rated Diffscriber ($\mu = 6.33$, $\sigma = 1.21$) significantly higher than Accessible Slides ($\mu = 3.33$, $\sigma = 1.75$) for understanding changes ($t(10) = 3.45$; $p = 0.0062 < 0.01$). When using Accessible Slides to review slide revisions, participants read through the content of the slide elements and rarely checked the

	Participants	P1	P2	P3	P4	P5	P6	**Mean**	**SD**
Changes	Acc. Slides	13	8	12	12	8	9	10.33	2.25
	Diffscriber	31	29	23	30	29	31	28.83	2.99
Feedback	Acc. Slides	2	0	1	2	0	0	0.83	0.98
	Diffscriber	2	4	3	3	2	3	2.83	0.75

Table 6: The number of identified changes by each participant (Changes) and the number of feedback comments (Feedback) when viewing slides in the Accessible Slides condition (Acc. Slides) and the Diffscriber condition. The amount of changes identified and feedback provided were significantly higher for Diffscriber than for Accessible Slides.

element design information (*e.g.*, x,y position and text styles). While participants were allowed to compare the edited slides with original one, 3 participants switched back and forth between the original presentation and the revised presentation, while 3 participants only read the elements on the revised presentation. Participants mentioned that using Accessible Slides, they were able to notice added images (P4, P5) and to figure out other changes to the content by switching back and forth (P1), but it was difficult to remember the original slide presentation (P6). When using Diffscriber to review slide revisions, 5 of 6 participants assessed slide changes using the high level change summaries, only inspecting lower-level changes or individual elements 1-2 times throughout the viewing session. P5 preferred to instead to navigate using individual slide elements, reading changes embedded underneath each element (rather than navigating through the high level changes). P5 suggested providing the opportunity to toggle off the types of descriptions that they are not interested in (*e.g.*, the change summaries in their case). When asking about participants' impression on incorrect change descriptions, 2 participants expressed it might be challenging but still possible for them to locate some incorrect description given the amount of information we provided. For instance, P3 mentioned that "if I saw there is a yellow circle added to the bottom of one text while the other was added to the left, then I probably could guess that there might be something wrong there to describe the position". Overall. all of them think the tool could be a tool that encourage more discussions between them and their collaborators despite the errors.

Authoring and collaborating on presentations. Participants produced more suggestions and feedback after reviewing slides using Diffscriber ($\mu = 2.83$, $\sigma = 0.602$), compared to Accessible Slides ($\mu = 0.833$, $\sigma = 0.787$) ($t(10) = 3.96$; $p = 0.0027 < 0.01$). In addition, participants rated Diffscriber to be significantly more useful as a presentation authoring tool ($\mu = 6.33$, $\sigma = 1.21$) than Accessible Slides ($\mu = 3.33$, $\sigma = 1.63$) ($t(10) = 3.96$; $p = 0.0047 < 0.01$). All participants wanted to use the system in the authoring process, and found both tools to be preferable to their prior experience using PowerPoint (the current most accessible slide authoring tool). P4 expressed that both systems were preferable to PowerPoint for reading and authoring tools but: *"more dramatically for authoring to get more context"* as they were excited to be able to *"make sense of changes and more actively participate, and put more of a stamp or personal style on it like some specific background image or font type"*. P1, P2, P3, P4 each mentioned that they would use the tool to understand the status of design changes so they would be able to have more input into the design process. P6 wanted to use the interface to *"show the information I mostly rely on other people to tell me"*. P3 mentioned they would particularly like the tool for creating a high quality presentation (*e.g.*, for a job interview) when they would want to *"know the content delivery plus have a visual understanding of changes and their impact on the presentation when I go to present"*.

When reflecting on the types of changes that would be most useful for authoring (1-7, 7 means very helpful), participants rated the **content changes** as most useful on average ($\mu = 6.50$, $\sigma = 1.22$). All participants noted that the content changes were important. For example, participants wanted to make sure the "content is consistent" (P2), "know what your collaborators changes are and make sure everything is good" (P3), and be able to suggest relevant edits (P4, P6). Participants rated the **style changes** as second most useful ($\mu = 6.00$, $\sigma = 1.22$), and the **layout changes** as third most useful ($\mu = 5.83$, $\sigma = 1.22$) for authoring. For style and layout changes, participants expressed enthusiasm around learning more about how presentations are authored such that they could author presentations and provide feedback in the future. For example, P3 and P4 expressed that they would be interested in using the layout change information to learn how people are changing layouts on the slide such that they can learn to design slides in the future. P4 was also interested in learning more about the styles on the slide from an aesthetic point of view, for example if the styles were similar or consistent across elements. P1, P2, and P6 reported that style and layout information were helpful in terms of learning more about the context of the content. Finally, participants rated the **accessibility metadata** (*e.g.*, raw text styles and x,y position) as fourth most useful ($\mu = 5.33$, $\sigma = 1.97$). Participants mentioned that the metadata was useful as it was necessary for authoring slides (P1,P2,P3,P4,P6), and P1 and P6 that such information was difficult or "painful" (P1) to get from PowerPoint currently. However, this information was still limited (P4) and the absolute rather than relative values (*e.g.*, for x, y coordinates) were not that useful (P5).

Observing description errors. The majority of description errors resulted from inaccurate detection of correspondences between prior and current slides. While correspondences were accurately detected for 77% of slide elements, errors occurred most often for slide elements with uncommon text (eg, a unique URL or proper noun), as the system failed to match sparse encodings for uncommon text to encodings for other slide elements. Other correspondence errors occurred when authors replaced slide text with an image depicting the text (eg, replaced text "35%" with an image of a hand-drawn "35%"), as the system failed to match the image encoding with the text encoding (Figure 8). In the future, we could improve the system by adding optical character recognition to extract text from images before computing element encodings. Only 3% of slides with accurately detected correspondences included a description error and these errors were due to ambiguous spatial relationships between slide elements.

While participants did not encounter errors during our study tasks, we shared examples of slides with common errors during the tutorial stage. When participants encountered example errors, they were able to identify the errors as the correspondence description did not make sense (e.g., the URL for an accessibility website, an

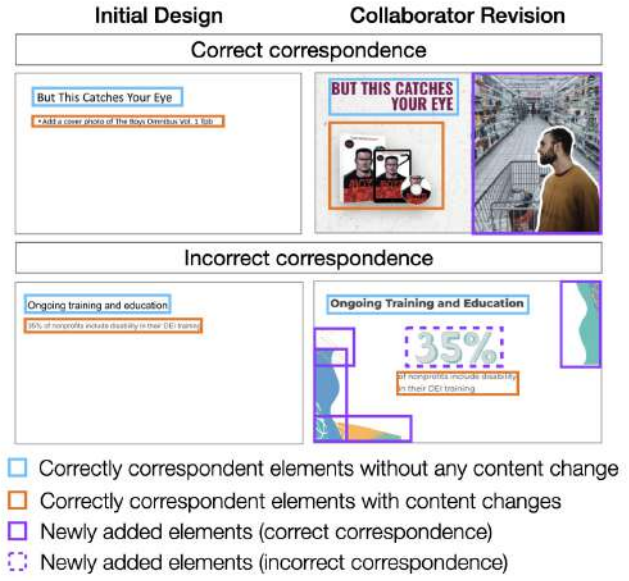

Figure 8: Two pairs of slides show the correct and incorrect correspondences between initial and revised slides (the assistant replaced the texts in the original slides with the image elements). The first slide titled "But this catches your eye" has all correct correspondences for non-revised, revised, and newly added elements. The second slide titled "Ongoing training and education" has an incorrect element correspondence, where the assistant made text "35%" from the sentence "35% of non-profit include disabilities in their DEI training" a separate word art image (caption: the white and the green numbers are in the shape of letters) and was recognized as a newly added element rather than a correspondent one.

example of uncommon text, was replaced with a star icon). Participants mentioned that even with errors, the descriptions contained enough information to improve discussion with their collaborators.

8 DISCUSSION AND FUTURE WORK

Diffscriber demonstrates that identifying and describing visual changes supports BVI presenters' presentation authoring. Supporting BVI content creators on mixed-ability teams is especially important given the ubiquity of visual information today. Diffscriber introduces a number of opportunities for future work:

Extending edits and summaries across multiple slides. Diffscriber currently assumes the number of slides is constant and describes changes between pairs of individual slides. As collaborators typically maintained the overall structure of BVI presenters' slides, this assumption covers a common use case. However, Diffscriber is unable to handle cases where a collaborator modifies slide ordering, merges slides, or separates one slide's content into multiple slides. In the future, we will investigate how to extend our matching approach to multiple slides rather than single slides. We will also explore the opportunity for change summaries across multiple slides rather than just within the same slide (*e.g.*, a style changed consistently across the entire presentation).

Cross-platform support. Diffscriber is implemented as a Google Slides plugin and relies on metadata specific to the platform's API and web-based scraping. While our target platform (Google Slides) is a popular tool for slide authoring, we aim to support additional slide-authoring tools in the future. We conducted some promising initial experiments using a visual extraction techniques following prior work [26, 38]. Operating on the visual content alone could make our approach independent of the particular tool used (and an increasingly popular approach in accessibility [48]).

High-quality alt text. The quality of descriptions generated by our system depends on the accessibility of the slide content itself (*e.g.*, the presence of alt text for slide figures). The materials provided to us by participants in our formative study were authored by assistants accustomed to working with BVI presenters and contained high-quality alt-text. However, assistants unfamiliar with accessibility may produce low quality alt text (*e.g.*, the filename, a few words). Diffscriber still operates without alt text using ML-based substitutes (*e.g.*, automated image captioning), but these are not as good as human-provided alt text.

Detecting additional changes. Diffscriber describes changes when moving from text-based presentations to full presentations (*e.g.*, with images and decorations). Our multimodal model could support the detection between visual edits beyond text to image translation by directly comparing the embedding in visual domain. We have conducted some preliminary explorations on detecting image resizing and displacement and the results showed that our system could identify such changes without introducing significant errors. In addition, prompted by comments in the user study expressing interest in the perceptual impact of changes (*e.g.*, legibility) we are exploring descriptions to make our system more useful for in-situ visual editing tasks such as an element-overlapping detector that informs users when there are occluded elements and an out-of-bounds detector that notifies users the when the elements move off of the page (*e.g.*, such as when adding text to a text block).

Supporting new collaboration types. Describing visuals and their changes supports BVI presenters in understanding the visual content on slides when collaborating with sighted co-authors. We designed the system to support existing collaboration that occurs asynchronously, or synchronously (e.g., verbally over a call), with only one collaborator directly editing the slides at a time. Future work can explore how to use our system for independent editing beyond text, collaboration between a BVI presenter and an automated design tool [23], and live editing (e.g., by streaming updates rather than requiring a user query). As slide authoring itself becomes more accessible, BVI presenters may want to edit slides at the same time as other collaborators (as is currently the case for document editing [4, 5]). Future work can explore how to extend our system to improve collaborative awareness to support synchronous and large group slide editing [12, 43], as prior work explored for mixed-ability collaboration in document editing [4, 5]. For example, our system could provide attribution details along with change descriptions. Given that many aspects of editing visual designs are less accessible than editing document text (e.g., editing spatial layouts, colors, styles), additional studies will be needed to understand

group dynamics in mixed-ability teams during synchronous visual design editing.

Evaluating visual change descriptions. We evaluated the usefulness of our descriptions for BVI presenters in a user study, and evaluated the performance of our system by assessing change detection (the source of most description errors). Future work could evaluate the quality of change descriptions quantitatively at scale by comparing automatically generated change descriptions to human-generated change descriptions (e.g., authored by crowd workers or experts in design or accessibility). We could recruit people (e.g., crowd workers, BVI presenters, sighted design experts) to compare human-generated descriptions to automated descriptions by rating metrics such as accuracy and coverage, and we could also use automated metrics [14] to evaluate the overlap between human generated and auto-generated descriptions.

9 CONCLUSION

In this paper, we introduce Diffscriber, a system for identifying and describing visual presentation design changes to support collaboration between BVI presenters and slide-authoring assistants. The design of our system is grounded in a formative interview study that we conducted to uncover the current practices behind BVI slide authoring. Even with accessible slides (*i.e.,* those containing alt-text and correct reading order), we discovered that the current workflow presented difficulties for BVI presenters to contribute to the slide design and creation process. Using participant feedback and source materials provided to us, we categorized the types of necessary slide operations and built a system capable of recognizing and describing these changes. We conducted a technical evaluation and a usability study which show that Diffscriber *(i)* accurately associates corresponding elements between slides and *(ii)* provides an effective interface for BVI presenters to reason about assistant-authored slides through hierarchically-summarized descriptions generated by our system. Finally, we discuss and explore avenues for future work that could allow BVI people to author other types of visual content more effectively and collaboratively.

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] Cynthia L Bennett, Erin Brady, and Stacy M Branham. 2018. Interdependence as a frame for assistive technology research and design. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 161–173.

[2] 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).

[3] Maitraye Das, Darren Gergle, and Anne Marie Piper. 2019. " It doesn't win you friends" Understanding Accessibility in Collaborative Writing for People with Vision Impairments. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[4] Maitraye Das, Thomas Barlow McHugh, Anne Marie Piper, and Darren Gergle. 2022. Co11ab: Augmenting Accessibility in Synchronous Collaborative Writing for People with Vision Impairments. In *CHI Conference on Human Factors in Computing Systems*. 1–18.

[5] Maitraye Das, Anne Marie Piper, and Darren Gergle. 2022. Design and Evaluation of Accessible Collaborative Writing Techniques for People with Vision Impairments. *ACM Transactions on Computer-Human Interaction* 29, 2 (2022), 1–42.

[6] Laurent Denoue, Scott Carter, and Matthew Cooper. 2018. SlideDiff: Animating Textual and Media Changes in Slides. In *Proceedings of the ACM Symposium on Document Engineering 2018*. 1–4.

[7] Steven M Drucker, Georg Petschnigg, and Maneesh Agrawala. 2006. Comparing and managing multiple versions of slide presentations. In *Proceedings of the 19th Annual ACM symposium on User Interface Software and Technology*. 47–56.

[8] Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge Belongie. 2019. Neural naturalist: generating fine-grained image comparisons. *arXiv preprint arXiv:1909.04101* (2019).

[9] Tsu-Jui Fu, William Yang Wang, Daniel McDuff, and Yale Song. 2021. DOC2PPT: Automatic Presentation Slides Generation from Scientific Documents. *arXiv preprint arXiv:2101.11796* (2021).

[10] 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.

[11] 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.

[12] C Gutwin. [n.d.]. A descriptive framework of workspace awareness for real-time groupware. *Computer Supported Cooperative Work (CSCW)* 11 ([n. d.]).

[13] 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.

[14] Harsh Jhamtani and Taylor Berg-Kirkpatrick. 2018. Learning to describe differences between pairs of similar images. *arXiv preprint arXiv:1808.10584* (2018).

[15] Martin Kurze. 1996. TDraw: a computer-based tactile drawing tool for blind people. In *Proceedings of the second Annual ACM Conference on Assistive technologies*. 131–138.

[16] Cheuk Yin Phipson Lee, Zhuohao Zhang, Jaylin Herskovitz, JooYoung Seo, and Anhong Guo. CHI 2022. CollabAlly: Accessible Collaboration Awareness in Document Editing. (CHI 2022).

[17] Jingyi Li, Son Kim, Joshua A Miele, Maneesh Agrawala, and Sean Follmer. 2019. Editing spatial layouts through tactile templates for people with visual impairments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086* (2022).

[19] Microsoft. [n.d.]. Microsoft Acccessibility. https://support.microsoft.com/en-us/office/make-your-powerpoint-presentations-accessible-to-people-with-disabilities-6f7772b2-2f33-4bd2-8ca7-dae3b2b3ef25.

[20] Microsoft. [n.d.]. Microsoft Seeing AI. https://www.microsoft.com/en-us/ai/seeing-ai.

[21] Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alexander Berg, Tamara Berg, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. 747–756.

[22] Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29, 1 (2003), 19–51.

[23] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. Designscape: Design with interactive layout suggestions. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1221–1224.

[24] Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Learning layouts for single-pagegraphic designs. *IEEE transactions on visualization and computer graphics* 20, 8 (2014), 1200–1213.

[25] Dong Huk Park, Trevor Darrell, and Anna Rohrbach. 2019. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4624–4633.

[26] 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*. 1–12.

[27] 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.

[28] Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966* (2020).

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[30] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[31] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. *arXiv preprint arXiv:2004.08728* (2020).

[32] 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.

[33] 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.

[34] 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.

[35] Anastasia Schaadhardt, Alexis Hiniker, and Jacob O Wobbrock. 2021. Understanding Blind Screen-Reader Users' Experiences of Digital Artboards. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[36] Woosuk Seo and Hyunggu Jung. 2021. Understanding the community of blind or visually impaired vloggers on YouTube. *Universal Access in the Information Society* 20, 1 (2021), 31–44.

[37] Azure Cognitive Services. [n.d.]. Microsoft PowerPoint Designer. https://support.microsoft.com/en-us/office/create-professional-slide-layouts-with-powerpoint-designer-53c77d7b-dc40-45c2-b684-81415eac0617.

[38] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. 2021. LayoutParser: A unified toolkit for deep learning based document image analysis. In *International Conference on Document Analysis and Recognition*. Springer, 131–146.

[39] Lei Shi, Yuhang Zhao, Ricardo Gonzalez Penuela, Elizabeth Kupferstein, and Shiri Azenkot. 2020. Molder: an accessible design tool for tactile maps. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[40] Alexa F Siu, Son Kim, Joshua A Miele, and Sean Follmer. 2019. shapeCAD: An accessible 3D modelling workflow for the blind and visually-impaired via 2.5 D shape displays. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 342–354.

[41] 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.

[42] Edward Sun, Yufang Hou, Dakuo Wang, Yunfeng Zhang, and Nancy XR Wang. 2021. D2S: Document-to-slide generation via query-based text summarization. *arXiv preprint arXiv:2105.03664* (2021).

[43] James Tam and Saul Greenberg. 2006. A framework for asynchronous change awareness in collaborative documents and workspaces. *International Journal of Human-Computer Studies* 64, 7 (2006), 583–598.

[44] Hao Tan, Franck Dernoncourt, Zhe Lin, Trung Bui, and Mohit Bansal. 2019. Expressing visual relationships via language. *arXiv preprint arXiv:1906.07689* (2019).

[45] Lourdes M Morales Villaverde. 2014. Facilitating blind people to independently format their documents. *ACM SIGACCESS Accessibility and Computing* 108 (2014), 38–41.

[46] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. 2017. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1180–1192.

[47] Haijun Xia. 2020. Crosspower: Bridging Graphics and Linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 722–734.

[48] Xiaoyi Zhang, Lilian de Greef, Amanda Swearngin, Samuel White, Kyle Murray, Lisa Yu, Qi Shan, Jeffrey Nichols, Jason Wu, Chris Fleizach, Aaron Everitt, and Jeffrey P. Bigham. 2021. Screen Recognition: Creating Accessibility Metadata for Mobile Applications from Pixels. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Virtual, McVirtualand).

[49] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. 2017. The effect of computer-generated descriptions on photo-sharing experiences of people with visual impairments. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22.