HHS Public Access

Author manuscript

ASSETS. Author manuscript; available in PMC 2022 February 18.

Published in final edited form as:

ASSETS. 2021 ; 17: . doi:10.1145/3441852.3471201.

The Efficacy of Collaborative Authoring of Video Scene Descriptions

Rosiana Natalie,
Singapore Management University, Singapore

Joshua Tseng,
Singapore Management University, Singapore

Jolene Loh,
Singapore Management University, Singapore

Ian Luke Yi-Ren Chan,
Singapore Management University, Singapore

Huei Suen Tan,
Singapore Management University, Singapore

Ebrima H Jarjue,
University of Maryland, College Park, USA

Hernisa Kacorri,
University of Maryland, College Park, USA

Kotaro Hara
Singapore Management University, Singapore

Abstract

The majority of online video contents remain inaccessible to people with visual impairments due to the lack of audio descriptions to depict the video scenes. Content creators have traditionally relied on professionals to author audio descriptions, but their service is costly and not readily-available. We investigate the feasibility of creating more cost-effective audio descriptions that are also of high quality by involving novices. Specifically, we designed, developed, and evaluated ViScene, a web-based collaborative audio description authoring tool that enables a sighted novice author and a reviewer either sighted or blind to interact and contribute to scene descriptions (SDs)–text that can be transformed into audio through text-to-speech. Through a mixed-design study with $N = 60$ participants, we assessed the quality of SDs created by sighted novices with feedback from both sighted and blind reviewers. Our results showed that with ViScene novices could produce content that is Descriptive, Objective, Referable, and Clear at a cost of *i.e.,* US$2.81pvm to US$5.48pvm, which is 54% to 96% lower than the professional service. However, the descriptions lacked in other quality dimensions (*e.g.,* learning, a measure of how well an SD conveys the video's intended message). While professional audio describers remain the

rnatalie.2019@phdcs.smu.edu.sg .

gold standard, for content creators who cannot afford it, ViScene offers a cost-effective alternative, ultimately leading to a more accessible medium.

Keywords

Scene description; visual impairment; video accessibility

1 INTRODUCTION

Videos rely heavily on visual cues to convey information and thus are often not accessible to people with visual impairments. Audio descriptions, verbal commentaries of visual information in videos are critically used to increase access [17, 22, 39, 48] and improve the accessibility of instructional, educational, and entertainment videos [12, 15, 35]. Providing audio descriptions is increasingly essential as online video consumption rises by 20 – 40% during the COVID-19 pandemic in 2020 [1, 13]. However, providing high-quality audio descriptions remains challenging as hiring professional audio describers is both time-consuming and costly [52].

Technical solutions minimizing the time and cost have mainly focused on user interfaces that streamline the authoring process or automate part of it [10, 25, 26, 42, 52]. Notably, Kobayashi *et al.* suggested that involving novices in the process could be a reasonable, cost-effective alternative [26] though no cost estimates or in-depth quality characteristics were reported. We build on this prior work to explore more fine-grained questions like *"is it really cost-effective to involve novices in audio description authoring?"* and *"what are the dimensions of audio description qualities and how "good" are novice-created audio descriptions along those dimensions?"*

Motivated by the benefits of online collaborative authoring and peer assessment in improving work quality [27, 34], we investigate how sighted novices can author high-quality audio descriptions through collaboration with a sighted or a blind reviewer. To this end, we designed and developed ViScene, an interactive web-based application, illustrated in Fig. 1. Using ViScene, an author writes the scene descriptions (SDs)–textual descriptions of video scenes–and ViScene converts them into audio descriptions through text-to-speech (TTS) [2]. Descriptions can be written in scene segments where dialogues are absent (Fig. 1b). A reviewer can access the authored SD segments and provide feedback on quality in the form of open-ended comments at a scene-level (Fig. 1c), which the author can address subsequently. The interface also visualizes SD succinctness (Fig. 1d) based on the length of the generated audio descriptions.

We used ViScene and three types of videos to explore whether novice authors and reviewers can collaboratively create high-quality SDs cost-effectively. We conducted a mixed-design study with 60 sighted participants who used ViScene to write SDs. The participants acted as authors and split into three groups: *without-feedback (baseline)*, *sighted-feedback*, and *blind-feedback*. All participants authored SDs on the same videos over two sessions. A sighted or blind reviewer provided feedback between the sessions to those in the *sighted-feedback* and *blind-feedback* group. The reviewers provided multifaceted feedback on the

quality of SDs using a SD quality codebook that we developed through an extensive review of existing guidelines. Sighted and blind evaluators assessed the quality of the final SDs. Our analysis suggests that novice authors Descriptive, Objective, Referable, and Clear SDs with the feedback. Although our collaborative authoring approach took about an hour to author SDs for a one-minute video, we show that the authoring cost is still cheaper than hiring a professional (*i.e.,* US$0.85 - US$9.65 per video minute). In summary, this work contributes:

- The design and development of ViScene, a web-based collaborative SD authoring system.
- A concise codebook grounded in experts' guidelines that characterizes the quality of SDs. The codebook enables reviewing and evaluation of SDs and can be used in settings beyond collaborative authoring.
- Empirical results with 60 participants demonstrating the feasibility, value and limitations of mixed-ability (*i.e.,* sighted and blind people) collaborative SD authoring.
- Design recommendations for future SD co-authoring interfaces elicited through the evaluation of ViScene.

2 RELATED WORK

We introduce prior efforts related to video accessibility with a focus on audio descriptions and technologies that support their authoring.

2.1 Video Accessibility and Audio Description

Internationally, anti-discrimination related regulations mandate the provision of videos with audio descriptions [30]. In the U.S., for example, CVAA Title 2 [14] requires major broadcast and cable networks to make online videos accessible. Similarly, Section 504 and Section 508 of the Rehabilitation Act (which refers to WCAG [9]) mandate public entities as well as federally funded organizations to make videos that they publish to be accessible [9, 14]. See [30] for a comprehensive survey of recent audio description-related legislations around the globe. Although the current regulations do not widely enforce non-public entities to make video contents accessible, video streaming platformers like Netflix make a significant effort in making their contents accessible [35, 40, 41].

Decades of work have gone to making guidelines for authoring high-quality audio descriptions [39]. Perhaps one of the most comprehensive sets of guidelines had been offered by the Described and Captioned Media Program [15]. While its focus is on educational video content, the guideline discusses a broad range of topics on how to make videos accessible with audio descriptions. Its content often overlaps with other guidelines like the Audio Description Coalition's guidelines, which focus on cultural video content [12]. We have reviewed these guidelines and other existing ones [17, 22, 35, 48] to create a concise codebook that has captured the key qualities of audio descriptions, as described in Section 4.2. In our study, the codebook is used both by the minimally trained reviewers who provide feedback to novice authors as well as the evaluators who assess the quality of the generated SDs.

Despite these legislative, corporate, and grassroots efforts to make comprehensive guidelines, the number of inaccessible videos seem to be increasing as more content is being uploaded to online video platforms. Guidelines like WCAG 2.1 encourage making videos accessible [9], but there are hundreds of hours' worth of video content uploaded every minute to YouTube [11, 20] with the majority of them lacking audio descriptions. Unfortunately, we are not aware of any prior work that quantitatively analyzed the prevalence of videos without audio descriptions. But, it is clear that many videos remain inaccessible for people with visual impairments. This could be partially attributed to the cost and availability of professional audio description service [38, 52]. One estimates the cost to be "$12 per video minute to $75 per video minute" [45] and the turnaround time takes days to weeks [45]. This makes it hard for casual video creators to make videos accessible via paid services. Thus, much work is needed to make the audio description authoring process more cost-effective and labor efficient, while maintaining quality.

2.2 Technologies to Support Audio Description Authoring

Prior work has proposed technologies to support authoring of audio descriptions [7, 19, 25, 31, 50]. Kobayashi *et al.* created a script editor that allows novice authors to edit the audio description and modify synthetic speech parameters easily [25]. Branje and Fels also reported that it is feasible for minimally trained amateur SD authors to generate overall medium to good quality audio descriptions [7]. The 3PlayMedia's post-production tool allows its users to author textual descriptions of scenes and convert them into audio description through TTS [31]. Tools such as Able Player [43] can deliver the audio descriptions created using the HTML5's <track> tag; Web content publishers can add audio descriptions in textual VTT format [46]. The textual information can then be read out by a screen reader or TTS [18, 26]. Although prior work has offered easy access to tools that allow users to provide audio descriptions, the rationale behind what is considered as a "good" quality SD was not elaborated. Kobayashi *et al.* examined the quality of novice-created audio descriptions [26], but their main focus was examining overall quality, not looking deeper despite how the quality of audio descriptions is multifaceted. Solely looking into the overall quality prevents us from understanding fine-grained audio description qualities. We believe these qualities are useful for SD authoring training, feedback, and automating SD generation.

Recent work has also investigated ways to combine manual audio description authoring with automation [10, 28, 42, 49, 52]. For example, Yuksel *et al.* created a workflow where the computer vision (CV) and natural language processing (NLP) model generates the initial description and volunteers post-edit [51, 52]. Their analysis focused on the time efficiency of generating the audio description and usability of the technology. Their quality assessment was centered around overall and topic understanding qualities. Campos *et al.* designed a way to automatically scripting audio descriptions using the pre-existing video scripts (*i.e.,* the blueprint of chronological rundown of the scenes) and subtitles [10]. Wang *et al.* [2021] attempted to fully automate the generation of audio descriptions, but the result still lacked some useful information for blind audiences (*e.g.,* the character actions, gender, places)[49]. Pavel *et al.*'s Rescribe automated the editing and revision processes to fit the content into limited space for audio descriptions using dynamic programming [42]. However, the

focus was on providing succinct descriptions and audio description's other quality facets remained under-explored (*e.g.,* descriptiveness, objectiveness). Overall, the effects of online human-human and mixed-ability collaborations in SD authoring have not been previously explored.

In this paper, we design and develop ViScene that supports the sighted novice to author SDs with the commentary feedback from either a sighted or a blind reviewer. The work is motivated by prior research demonstrating the benefit of collaborative writing and peer evaluation in improving one's writing [4, 21, 24, 27]. The mixed-ability collaboration in our work is a novel framing in the context of audio descriptions and is inspired from the prior studies of how blind and sighted people, when working in pairs, can co-create an accessible household [5] and workspace [6].

3 VISCENE

ViScene's user interface builds on the design of the existing audio description authoring tools (*e.g.*, [26]); it shows a video pane on the left and table that consists of closed caption (CC), scene descriptions (SDs), and comment feedback on the right (Fig. 1). The video pane (Fig. 1a) has a few subcomponents; a video player, CC and SD bars (Fig. 1b), and a video selector (Fig. 1f). In addition to typical video interactions (*e.g.*, play, pause), the user interface visualizes the segments of the presented video where CC and SD are offered, allowing a sighted user to browse the presence of speech in the video. SD bar also indicates the corresponding SD's succinctness; the bar's fill color turns red if the authored SD overlaps with a speech in the video (*i.e.,* CC), nudging the user to shorten the SD (succinctness visualization).

ViScene accommodates two types of users: authors and reviewers. Authors can watch and listen to a video and write SD segments in corresponding table rows for each scene (Fig. 1c). They can write and edit SDs under the Scene Description column only at time segments where there is no speech (*i.e.*, no closed captions). They have read-only access to Closed Caption and Feedback columns. Authored SDs are rendered to audio with the Amazon Polly TTS service [2]. ViScene dynamically computes the SD's audio length. The estimated length proportionally defines the width of the SD bar. To indicate the link between an SD and the corresponding SD bar, both a table cell and SD bar turns yellow when a user hovers over either of them with a mouse cursor.

Reviewers' interface is similar to authors' with minor differences. Just like authors, reviewers can watch and listen to the videos. They can also read SD written by the authors and write under Feedback column where they can provide comments to authors on how they can improve SD. The interface has an additional drop-down menu to select the author's work (Fig. 1g). They can select an author whose work they wish to assess. The interface is compatible with screen readers to ensure access among blind reviewers.

Using ViScene, an author and a reviewer take turns generating audio descriptions collaboratively. First, in the *authoring* step, an author writes SDs (referred to as *session 1* in our study). Then in the *review* step, a reviewer assesses the author's SDs and provide

pointers for improvements. Last, in the *revision* step, the author address the reviewer's comments (referred to as *session 2* in our study). We will define the terms *session 1* and *session 2* in more detail in Section 4.3.

4 STUDY METHOD

We conducted a remote user study to assess the quality and cost of creating SDs with ViScene.

4.1 Videos and Ground Truth Scene Descriptions

Videos' visual contents vary depending on the videos' intended audience, goal, and tone. To explore the effects of videos on the SD quality, we used three different types of videos– Explainer, Instructional, and Advertisement. We had two additional selection criteria in addition to video types: video length and availability of ground truth SD. We aimed to find videos that are about one to two-minute-long; we identified that such videos are suitable for a study that takes about one to two hours per session based on our pilot study. Furthermore, we wanted to find videos that come with high-quality SDs that could be used as a reference to evaluate the quality of participants-authored SDs.

- **Video 1 - Explainer Video** [1]: Web Accessibility Guide, Duration: 1 min 38 sec. (Fig 2a). A video about "Colour with Good Contrast," taken from the W3C website. The video explains the situations in which good color contrast of digital contents (*e.g.*, mobile application interface) is necessary.
- **Video 2 - Instructional Video** [2]: Origami Tutorial, Duration: 1 min 42 sec. (Fig 2b). An instructional YouTube video on how to make an origami bookmark corner. A person presents the folding movement step-by-step. Each instructional step is visually depicted without any verbal narrative, making this video inaccessible for people with visual impairments.
- **Video 3 - Advertisement Video** [3]: Subaru Commercial, Duration: 1 min (Fig. 2c). A Subaru car advertisement which shows a group of people exploring the scenic area with a Subaru Outback car. The video was introduced by the Audio Description Project (ADP) by the American Council of Blind (ACB) as an example advertisement that provides good audio descriptions.

The three videos that we selected represent the types of videos that are commonly available online. For example, videos like Web Accessibility Guide and Origami Tutorial belong to the educational video category that are among the most commonly available and most played videos on YouTube [3, 32]. These two videos had objective goals of conveying a concept or helping the viewers to follow instructions. To include a video that has a less defined goal with a less educational tone, we included Subaru Commercial, which was more geared towards entertaining the audience.

[1] https://www.w3.org/WAI/perspective-videos/contrast/
[2] https://www.youtube.com/watch?v=hO4J1GjPQFw
[3] https://www.youtube.com/watch?v=flu6u988kh0

Two of the three videos we chose (Explainer and Advertisement) came with the audio description. We treated them as a reasonable ground truth as they were provided or endorsed by the trustable sources. Web Accessibility Guide and its audio description was created by W3C-WAI [22] and Subaru Commercial's audio description was acknowledged by ACB [37].

Because Origami Tutorial did not have an audio description, we invited an origami expert to help us create it. He wrote an origami book "It's Just a Bit of Paper" [16] and published blog posts to teach origami for people with visual impairments [4]. We held a remote meeting in which he watched the video and verbally explained each instructional step. We later transcribed the instruction. During the meeting session, he invited his legally blind partner to follow his verbal instruction to create an origami bookmark corner. His partner had the prior general knowledge about origami, but it was her first time making the origami from the Instructional video. His partner was able to make the expected outcome of the Instructional video, which verified that the verbal instruction was useful for people with visual impairments. Thus we used the transcribed instruction as the ground truth for the Instructional video.

Because ViScene presents a textual SD of each scene as a row on the table pane Fig. 1, we had to segment the video into scenes manually. For Explainer and Advertisement videos, each video segment represented either the dialogue or the natural pause where the narration or dialogue was absent. For the Instructional video, we split the video into ten instructionally meaningful segments and two additional segments (intro and outro; Fig.2).

4.2 Scene Description Quality

We needed a way to instruct novice authors and reviewers on what constitutes the qualities of SDs and guide the evaluation of participant-created SDs' qualities. We create a concise codebook based on the existing guidelines for making audio descriptions via a literature review of the existing guidelines and way of assessments [8, 12, 15, 33, 35, 38, 44, 48, 50]. We did not directly use one of the existing guidelines for our study because of their length and foci. For example, DCMP's Description Key provides a very detailed, comprehensive set of audio description authoring guidelines [15]. However, it consists of 63 indicators; this is too long to instruct novice authors about the audio description quality quickly, and some items are too detailed for our purpose. Also, its focus is more on education, so we wanted to incorporate views of other guidelines (*e.g.*, Netflix's audio description style guide, which is applicable for entertainment videos [35]).

The resulting codebook had nine codes: *Descriptive, Objective, Succinct, Learning, Sufficient, Accurate, Referable, Interest* and *Clarity*, which we extended from Natalie *et. al*'s codebook [34]. The simplified codebook is shown in Table 1; see Appendix A for the full codebook with detailed descriptions and examples. We instructed authors and reviewers to keep these codes in mind in authoring and reviewing SDs. The evaluators used the codebook to assess whether SDs satisfied the quality criteria described in each code.

[4] http://www.itsjustabitofpaper.co.uk/

4.3 Experimental Design

We evaluated the efficacy of supporting novice audio describers to author good scene descriptions using ViScene through a remote user study. We turned to the remote user study due to COVID-19 restrictions, and we used Microsoft Teams for remote communication with our participants. We invited each participant to two sessions; they were asked to use ViScene to independently write SDs in the first session, and revise their SDs in the second session based on review comments. Two members of the research team (one sighted and one legally blind) reviewed the authors' work. Three other members of the research team (two sighted and one fully blind authors) evaluated the quality of SDs created through author-reviewer collaboration.

The study was a 3×2 mixed-design study, where the feedback (its presence and a type of reviewer) was the between-subjects factor (*i.e., without-feedback*, *sighted-feedback*, and *blind-feedback)*, and session was the within-subjects factor (*i.e., session 1* and *session 2*). For the scene descriptions generated by the participants in *sighted-feedback* condition, a sighted reviewer provided comments on the SD quality using the codebook in Table 1. Likewise, for the SDs generated by the participants in *blind-feedback* condition, the blind reviewer provided comments on the SD quality using the codebook.

4.3.1 Participants and Reviewers.—We recruited sixty participants as the SD authors via university listserv and word-of-mouth. We randomly divided them into three groups (*i.e., without-feedback*, *sighted-feedback*, and *blind-feedback*), resulting in twenty participants in each condition. We recruited only the participants who had no experience in writing the SD to evaluate if novice authors could generate good audio description using ViScene. We also ensured that the authors did not have any disabilities (*e.g.*, visually-impaired, deaf and hard-of-hearing). We randomly assigned the participants to each condition.

Two members of the research team acted as the reviewers and gave comments to the SDs that the participants authored after *session 1*. One of them was a sighted reviewer and the other one was legally blind. We did not recruit reviewers from outside the research team, because we wanted a dedicated and motivated reviewer, controlling the quality and variability of the review comments. The blind reviewer used a screen reader to read SDs provided by the participants. Neither of the reviewers had prior experience in authoring or reviewing SDs. We trained the reviewers by asking them to understand the criteria of a good SD that we described in our codebook in Table 1. The research team leader trained the two reviewers to give good comments while they assessed the first two SDs from the participants. The leader then instructed the reviewers to give comments for all the remaining SDs independently, and instructed them to be objective, descriptive, and consistent in giving comments. We show excerpts of the comments that the reviewers gave to authors (Table 2).Neither the authors and the reviewers were exposed to the ground truth ADs that accompanied the three videos.

4.3.2 Procedure.—Each participant authored SD for three different videos using ViScene. We counterbalanced the sequence of the video to minimize the learning effect. In the first session, we introduced the basic concept of audio descriptions and SD (*e.g.*,

how they support people with visual impairments), the motivation of the study and the task. We briefed the participants on ViScene's interface, the topic of each video, and audio description qualities. Between *session 1* and *session 2*, a reviewer assessed the quality of the SDs using the codebook in Table 1 and provided comments to the participants using the ViScene's reviewer interface. After the review was done, we invited the participants for *session 2* to revise the SD. We did not disclose the source of comments to the participants (*i.e.,* sighted reviewer or blind reviewer) to minimize bias on the perceived importance of comments. ViScene logged key interactions (*e.g.*, task_start, task_end), which allowed us to calculate how much time authors and reviewers took to complete their tasks.

4.3.3 Sighted Researcher's Assessment based on the Codebook.—

Using the codebook in Table 1, two members of the research team evaluated the SDs. They performed manual binary evaluation. If an evaluator approved that the quality of SD had satisfied a quality facet, they gave '1' to the SD. If they felt that the SD did not satisfy the quality facet, they gave '0' to the SD, indicating a rejection. The sighted evaluator used the nine codes in Table 1—*Descriptive, Objective, Succinct, Learning, Sufficient, Accurate, Referable, Interest, and Clarity.* The unit of analysis was at a video-level. For each video, the evaluators read/listened to the SDs for the entire video and assessed whether each facet of SD qualities was satisfied. We assessed quality at a video-level as opposed to evaluating each scene because the *Learning, Sufficient, Interest, and Clarity* qualities could not be assessed at a scene-level. We randomized the sequence of the SD to minimize the evaluators' potential bias from review *conditions* and *sessions*. The sighted evaluators watched the videos with ground truth audio descriptions to make an informed evaluation.

4.3.4 Blind Researcher's Assessment based on the Codebook.—

A totally blind member of the research team acted as the blind evaluator. Similar to the sighted evaluator, the blind evaluator used the codebook to perform the binary evaluation. The unit of analysis was at a video-level. We asked the blind evaluator to assess the quality of the SD from the perspective of the end-user of the audio description. In the assessment by a blind evaluator, seven codes—*Descriptive, Objective, Succinct, Learning, Sufficient, Interest, and Clarity.* Because the blind evaluator could not justify the Accurate and Referable variables without seeing the videos, they did not assess them. Just like the sighted evaluators, the blind evaluator was familiarized with the ground truth audio description. The order of the evaluated SDs was randomized.

5 RESULT

We recruited sixty participants (15 male and 45 female) as SD authors. The participants' ages ranged from 19 to 37 years old (Mean = 22.8, SD = 3.16). None of them except for one participant had prior experience in authoring SDs. The participant who claimed to have prior experience previously participated in one user study in which she authored one description on one scene. Because this was a minor experience, we did not exclude her from the study.

The nine quality variables in the codebook (Table 1) guided the following analysis. We investigated how each quality aspect of participant-provided SD changed over two sessions.

We hypothesized that SD qualities would (i) remain the same in the control condition and (ii) improve with comments from sighted or blind reviewers.

Although we had two groups of evaluators with distinct abilities in a vision who used a slightly different set of SD quality codes, we used the same statistical methods to analyze the data since the data format was congruent.

- **Quality Change Analysis.** To assess the changes in SD qualities over two sessions and between feedback conditions, we used a generalized linear mixed model (GLMM) with a logit link function. We used GLMM instead of a mixed-design ANOVA to incorporate the random effects (*i.e.*, participants and videos). For the main effects, we had *feedback* type (*without-feedback, sighted-feedback,* and *blind-feedback*) and session (*session 1* and *session 2*), as well as their interaction. We treated videos and participants as random intercepts. This analysis allowed us to ask, *"can an author improve SD quality with the feedback from a reviewer?"*
- ***Session 2* Outcome Comparison** The above analysis does not allow us to ask *"is the SD quality good in the treatment conditions after the authors' revision?"* Thus, we conducted a post-hoc test by only using the data from *session 2*. We used GLMM with a logit link function. We had a *feedback* type as a main effect and *participants* and videos as random intercepts.

After looking into the results from each individual evaluator, we compared the results from the two groups. We also assessed how much time the collaborative process between authors and reviewers took to produce SDs. We note that while there was imbalance in gender, we did not observe significant effect on the results.

5.1 Sighted Evaluation

To see the agreement level of the two members before evaluating all the SDs, (i) we randomly sampled 90 from all of the SDs (N=360), (ii) two members independently evaluated the 90 SDs, (iii) computed agreement between the two, and (iv) reevaluated until the agreement was sufficiently high for the 90 SDs. We randomly selected 15 SDs from each of six condition pairs (3 feedback conditions x 2 sessions). This process took three rounds. In the first round, the sighted evaluator did not assess the Interest and Clarity qualities because we thought that such an evaluation would be biased due to visual information. However, we included these qualities in evaluation later to compare the results with blind evaluation results. First-round agreement was (Descriptive, Objective, Accurate, Succinct, Learning, Referable, Sufficient, Interest, Clarity) = (49%, 70%, 46%, 72%, 57%, 76%, 74%, NA, NA). In the second and third rounds, the agreements were (68%, 73%, 77%, 64%, 70%, 76%, 74%, 56%, 69%) and (69%, 73%, 74%, 90%, 70%, 76%, 74%, 56%, 69%). Between the rounds, the two members met, discussed their disagreements, and added examples to the codebook (Appendix A) to help the next round of evaluation. The Interest agreement was low even after the third round; establishing that the agreement for this quality was difficult due to its subjective nature. Thus, the two members moved on to independently evaluating the remaining 270 anyway. The disagreements in quality assessment for all 360 SDs were resolved through discussion.

We reported the number of approvals agreed by the two evaluators in a tuple (# approvals in *without-feedback*, # approvals in *sighted-feedback*, # approvals in *blind-feedback*). We examined the main effects (*i.e., feedback* type and *session*) and the interaction between them Table 3. We use "*control*" and "*without-feedback*" interchangeably to indicate the same baseline condition.

Descriptive.—The number of approvals changed from (22, 15, 18) in *session 1* to (30, 41, 21) in *session 2*. For the *sighted-feedback* condition, we observed a significant interaction between the *feedback type* (*i.e., sighted-feedback*) and *session* (z = 2.616; p = 0.009 < 0.001). The positive z-value and visual inspection of Fig. 3 (Descriptive pane) suggest that descriptiveness improved. The post-hoc test result (Table 4) indicates possible trend of *sighted-feedback* condition ($z = 1.897$; $p = 0.0578$). The result suggests that the authors in the *sighted-feedback* group improved the SD descriptiveness.

Objective.—The number of approvals changed from (45, 36, 42) to (42, 40, 53). There was a significant interaction between *blind-feedback* and *session* ($z = 2.61$; $p = 0.009 < 0.05$). The result indicates that the participants in *blind-feedback* group improved the SD objectivity of the SD. We also observed a significant effect of the *blind-feedback* to the SD objectiveness in the post-hoc test ($z = 2.415$, $p = 0.0157 < 0.05$), suggesting the superior objective quality of SDs from this group. .

Succinct.—The number of approvals changed from (56, 58, 59) to (56, 49, 58). There is a significant interaction between *sighted-feedback* and *session* ($z = -1.962$, $p = 0.0498 < 0.05$). The negative z-value and visual inspection of Fig. 3 indicate that the participants in the *sighted-feedback* group compromised the succinctness to address the feedback. This is not surprising because to address the comments the participants attempted to provide more descriptions, making SDs longer.

Learning.—The number of approvals changed from (13, 15, 13) to (15, 45, 21). We observed a significant interaction between *sighted-feedback* and *session* ($z = 3.443$; $p < 0.001$). The post-hoc test showed that the learning quality of SDs in the *session 2* is significantly better in *sighted-feedback* condition compared to the control condition ($z = 5.218$; $p < 0.001$). This suggests that the authors with *sighted-feedback* disproportionately improved SD Learning quality compared to the *control* condition.

Sufficient.—The number of approvals changed from (5, 8,6) to (9, 28, 8). There were no significant main effects, but we observed a trend in the interaction between the *sighted-feedback* and *session* ($z = 1.703$, $p = 0.089$). The post-hoc test on *session 2*, there was a significant effect on the *sighted-feedback* condition ($z = 3.421$, $p < 0.001$). This suggests that the author in the *sighted-feedback* condition was more likely to provide a SD with sufficient information.

Interest.—The number of approvals changed from (34, 29, 25) to (32, 43, 32). We did not observe significant main effects. There was a significant interaction effect between *sighted-feedback* and *session* (z=2.477, p=0.013<0.05). The post-hoc test showed a possible trend on the *sighted-feedback* condition (z=1.927, p=0.054). While the nature of this quality

is subjective and thus less reliable, the data may suggest that the participants can improve the interest quality with *sighted feedback*.

Clarity.—The number of approvals changed from (28, 25, 21) to (26, 41, 23). We did not observe significant main effects, but we observe a significant interaction between the *sighted-feedback* and *session* ($z = 2.602$, $p = 0.009 < 0.01$). The post-hoc analysis showed a significant effect of *sighted-feedback* ($z = 2.605$, $p = 0.009 < 0.01$), suggesting the clarity improved and was higher with feedback from *sighted reviewers* (Table 4).

Accurate.—The number of approvals changed from (31, 33, 31) to (38, 42, 41). There were no significant main effects of *feedback type* and *session*, and no interaction effects. There were no significant main effects in the post-hoc test too.

Referable.—The number of approvals changed from (30, 30, 35) to (32, 46, 39). There were no significant main effects, but there was a significant interaction between *sighted-feedback* and the *session* ($z = 2.268$, $p = 0.023 < 0.05$). The post-hoc test showed a significant effect of the *sighted-feedback* to the quality ($z = 2.597$, $p = 0.009 < 0.01$). This suggests that the participants in *sighted-feedback* condition can improve SDs to reference characters and objects well.

Summary.—The results of the Quality Change Analysis suggested that Sighted feedback improved descriptiveness, learning, referability, interest, and clarity. Blind feedback improved objectiveness. This suggests that feedback from different groups complement each other in improving SD qualities. We observed high succinctness over all the conditions, probably induced by the succinctness visualization.

5.2 Blind Evaluation

Similar to Sighted Evaluation, we report the raw approval numbers in three-tuples ((# approvals in *without-feedback*, # approvals in *sighted-feedback*, # approvals in *blind-feedback*)) and report the significant main effects and interaction effects (Table 5).

Descriptive.—The number of approvals changed from (49, 43, 46) in *session 1* to (52, 49, 56) in *session 2*. We did not observe significant main effects or interaction effects. The post hoc test did not show significant effects either. However, we took note that descriptiveness is high across the conditions.

Objective.—The number of approvals changed from (39, 28, 25) to (34, 29, 37). The statistical analysis revealed significant main effects under the *sighted-feedback* and *blind-feedback* conditions. For the *sighted-feedback* group, we observed a significant main effect ($z = -2.198$; $p = 0.028 < 0.05$) but no interaction. For the *blind-feedback* group, we observed a significant main effect of *blind-feedback* ($z = -2.771$; $p = 0.006 < 0.01$) and a significant interaction effect ($z = 2.400$, $p = 0.016 < 0.05$). The results and visual inspection of Fig. 4 (Objective pane) suggest that SDs under the control condition were better overall compared to treatment conditions. However, *blind-feedback* had a positive impact on SDs' objectiveness. The post hoc test did not show a significant difference in objectiveness between conditions.

Succinct.—The number of approvals changed from (35, 37, 40) to (39, 42, 39). We did not observe any significant main effects or interaction effects. The post hoc test did not show significant effects.

Learning.—The number of approvals changed from (27, 26, 27) to (30, 25, 31). We did not observe any significant main effects or interaction effects. The post hoc test did not show significant effects.

Sufficient.—The number of approvals changed from (36, 28, 38) to (43, 39, 39). The statistical analysis showed a trend of the *sighted-feedback* to the sufficiency of the SDs ($z = -1.698$, $p = 0.090$). There were no significant main effects of *blind-feedback* and interaction effects. The post hoc test did not show a significant difference in sufficiency.

Interest.—The number of approvals changed from (18, 12, 26) to (20, 11, 18). We did not observe significant main effects or interaction effects. The post-hoc test did not show a significant difference either. We noticed lower approval counts across conditions compared to the other quality types.

Clarity.—The number of approvals changed from (48, 43, 42) to (50, 42, 51). We did not observe significant main effects or interaction effects. The post-hoc analysis showed a significant effect of *sighted-feedback* ($z = -2.230$, $p = 0.0257 < 0.05$), suggesting that the clarity was lower in this condition (Table 6).

Summary.—The result suggests that the blind evaluator perceived that the objectiveness of SDs improved with *blind-feedback*. SDs' descriptiveness was relatively high, but we did not see the benefit of feedback. We noticed that the Interest and Learning scores were lower compared to other qualities.

5.3 Comparison between Sighted and Blind Evaluation

To investigate the similarities and differences in how the sighted evaluators and blind evaluator perceived the SD qualities, we looked at the agreement in SD quality approvals and rejections between the two groups. We focused our analysis on the seven codes that both sighted and blind evaluators used *(Descriptive, Objective, Succinct, Learning, Sufficient, Interest, and Clarity)*. We first counted the total number of agreements across all videos between the two sighted evaluators' amalgamated assessment and the blind evaluator's assessment. For example, if both sighted assessment and blind assessment approved or rejected an SD, we counted that as an agreement. We also stratified the agreement counts by feedback conditions and video types to see if they have any effects on agreements (Fig. 5, Appendix B)

In comparison, we observed different levels of agreements across the seven codes. For the $N = 360$ SDs, the numbers of agreement were (Descriptive, Objective, Succinct, Learning, Sufficient, Interest, Clarity) = (164, 226, 224, 176, 139, 109, 208) (or (46%, 63%, 62%, 49%, 39%, 30%, 58%)). Table 7 shows agreement with the confusion matrices. We observed higher levels of agreement for *Objective* and *Succinct*, which was due to both sighted and blind evaluators approving the majority of SDs. For Interest, the sighted evaluator had a

tendency to approve more SD, whereas the blind evaluator approved more SDs for *Learning*, *Sufficient*, and *Clarity*.

Descriptive.—The sighted and blind evaluators' assessment agreed on 162 out of 360 SDs. Agreement slightly increased from 73 to 89. While the blind evaluator approved most SDs' descriptiveness (295/360), the sighted evaluator rejected the majority (213/360). But both of the evaluators agreed that more SD appeared to be descriptive in *session 2*, especially in the *sighted-feedback* condition (Fig. 5). The disagreement in descriptiveness between the blind and sighted evaluator is likely because the sighted evaluators strictly rejected the ones that missed out on the detailed information about the scene. Oh the other hand, the blind evaluators perceived the SDs to be descriptive enough when they could get a sense of the scene's content.

Objective.—The sighted and blind evaluators' assessment agreed on 226/360 SDs. Both groups approved more than half of SDs. However, Fig. 5 suggests that the agreement for the Advertisement Video was weaker. Fig. 5 also shows that both sighted and blind evaluators increased the number of approvals in the *blind-feedback* condition, which indicates the positive impact of blind review on this quality dimension.

Succinct.—The two assessments agreed for 224/360 SDs for succinctness. The sighted evaluator was more generous in assessing the succinctness, giving 336/360 approvals. The blind evaluator's approval counts were lower for Explainer and Instructional Videos. Both mostly agreed on the assessment for Advertisement Video (Fig. 5). This was somewhat surprising as we expected succinctness to be high in agreement with both evaluations due to feedback from both reviewers and succinctness visualization. The gap suggests sighted and blind people perceive succinctness differently.

Learning.—The evaluators disagreed in assessing the learning quality (agreement=176/360). We observed different patterns between video types (Fig. 5). For the Explainer Video, the sighted evaluator had shown that *sighted-feedback* improved SD Learning quality, but the blind evaluator had perceived the SD otherwise. For the Instructional Video, we noticed that the sighted evaluator was more strict. During the discussion within the research team, the sighted evaluators noted that they rejected SDs when they found instructional errors. That was when the SDs' accuracy in describing the visual content interfered Learning. The blind evaluator was more generous supposedly because of the lack of this interference. The blind evaluator and sighted evaluators disagreed on the assessment of the Advertisement Video's SDs (Fig. 5). The discrepancy may be due to the sighted evaluators watching the video and being able to relate to the contents of the SDs more easily. The blind evaluator had a harder time understanding the story and its objective (*i.e.,* branding) solely from SDs.

Sufficient.—The agreement between sighted and blind evaluators was 139/360. The most notable disagreement was seen in the Explainer and Advertisement Videos. This was because the blind evaluator accepted most of the SDs but sighted evaluators did not. The discrepancy may have been caused by the fact that the sighted evaluators could see the

content of the Instructional and Advertisement Videos and many of the descriptions did not depict the scene entirely in, which the blind evaluator was not aware of.

Interest.—The agreement between sighted and blind evaluators was 109/360. The sighted evaluator appeared to be more generous in accepting the Interest qualities, whereas the blind evaluator had rejected most of SD (Fig. 5). This disagreement was expected as the variable seems to be subjective in nature, and different people tend to perceive different contents as interesting.

Clarity.—The agreement between sighted and blind evaluators was moderate (208/360). The blind evaluator was more generous in approving the SD (Fig. 5). The Instructional video achieved the lowest Clarity approval count (Sighted=27 approvals, Blind=53 approvals). The Instructional video required the audience to be familiar with the specific terms used in the origami video and to follow the instructions carefully. Thus, the evaluators may have felt it required a higher level of clarity compared to the other two videos.

Summary.—The agreement was highest for the *Objective* quality. The blind evaluator was stricter in evaluating Objective and Succinct qualities. The sighted evaluators were more strict in evaluating *Descriptive* and *Clarity* qualities. We believe these observations would be useful in interpreting the results in future work that employs manual evaluation of SD qualities. We observed high disagreement in *Learning*, *Sufficient*, and *Interest* qualities between sighted and blind evaluators. The potential explanations for the differences are because of the video types and unconscious influence of visual contents to the sighted evaluator's quality assessment.

5.4 Task Completion Time

Using ViScene's interaction log, we calculated the time that the authors and reviewers spent to perform their tasks to inform the cost of authoring SD using ViScene. In Table 8, we reported the time it took to (i) author the first iteration of SD for the three videos, (ii) reviewing the SD, and (iii) revising the SDs. We also reported the total time taken by adding the values from the three stages.

On average, the three-stage process in the *blind-feedback* condition took the longest (Mean=166.4 min), followed by *sighted-feedback* condition (Mean = 156.5 minutes) and *without-feedback* conditions (*i.e.,* Mean = 57.3 minutes)—see Table 8. The total duration for *sighted-feedback* and *blind-feedback* being longer than the duration in the *without-feedback* condition is not surprising. This is because the additional time cost for the reviewing stage was added (Mean = 51.8 minutes for *sighted-feedback* review and Mean = 57.8 minutes for *blind-feedback* review).

The time taken to author and revise the SD had varied across three different conditions. On average, the authors completed *session 2* in the *without-feedback* condition much faster than *session 1* (*session 1*: Mean = 53.8 minutes; *session 2*: Mean = 20.5 minutes). For the authors in the *sighted-feedback* group, it took: (*session 1*: Mean = 46.7 minutes; *session 2*: Mean = 52.3 minutes). And for the authors in the *blind-feedback* group (*session 1*: Mean = 57.5 minutes; *session 2*: Mean = 50.4 minutes). The authors in the control group spent less time

in *session 2* because they did not get any reviewer feedback and only self-reflected on the SD they wrote in *session 1*.

Using the total time taken for each video, we calculated the monetary cost of generating SDs using ViScene through the process that involved *sighted-feedback* or *blind-feedback*—see Table 9. For each video, we multiplied the total time taken to author SDs by the US federal minimum wage (*i.e.*, US$ 7.25 per hour) [36] to compute the monetary cost. Because the three videos' lengths varied, we divided the cost by the video length to obtain cost per video minute (pvm). The monetary cost ranged between US$0.85pvm to US$9.65pvm (Mean = US$4.54pvm, SD=US$2.01pvm). The *blind-feedback* condition's costs per video minute are higher than *sighted-feedback* conditions for all types of the videos (Table 9). However, in both conditions across video types, the results suggest that authoring SDs via ViScene can be more cost effective than recruiting the experts. The cost per video minute for the treatment conditions varied from US$2.81pvm to US$5.48pvm. In comparison, the cost of professional service that would range from US$12pvm to US$75pvm [45]. This shows that with ViScene, the cost of generating SD for video can be reduced by 54% (=1 -US$5.48pvm / US$12pvm x 100%) to 96 (=1 -US$2.81 pvm/ US$75pvm x 100%)

6 DISCUSSION

We designed and developed ViScene, a tool that allows novice sighted authors to work with sighted or blind reviewers to generate SDs collaboratively. Based on the multifaceted quality assessment of SDs authored by our participants, we believe that the novice authors can write *Descriptive, Objective, Referable*, and *Clear* (*i.e.*, denoting high Clarity) SDs with the feedback. However, we found some disagreement between sighted and blind evaluations in, for example, Succinctness for the Instructional Video. We observed that the feedback from reviewers enabled the authors to improve SD qualities. Most importantly, the SD objectiveness, a quality that is valued highly by blind people [12], was found to have improved when the novices received feedback from blind and sighted reviewers. This is promising.

Although, blind evaluator had indicated that the Interest quality across all the videos, as well as some for *Objective*, *Succinct*, *Learning*, and *Sufficient* qualities remained low. For instance, the blind evaluator approved only a handful of SD for *Objective*, *Learning*, and *Interest* qualities for the Advertisement Video. We believe that this is partially explained by inherent limitations in the videos *i.e.*, the video's message relied heavily on the visuals to learn the content, requiring the audience to combine the visual message and audio description to comprehend the intended idea of the content creators. As Braun notes, "audio describing a film is therefore not simply a matter of substituting visual images with verbal descriptions" [8]. Our results encourage not only SD authors but also content creators to be more empathetic towards blind audiences when conveying the intended takeaway message. This is mainly so that blind audiences would be able to grasp the video contents.

Moreover, we also observed that some of the SD prioritized different goals which had introduced a conflict between qualities. For example, we noticed a trade-off between Succinct, Descriptive, and Learning. In Fig. 5, the blind evaluator perceived SDs as objective

but less succinct for Instruction Video, but vice versa for Advertisement Video. This suggests that the authors can generate SD taht excel in one quality, but would negatively affect other variables due to the respective qualities' innate conflicting nature. Future work should investigate the trade-off relationships between the variables in more depth and explore how we could help the reviewers and authors know how to prioritize and balance the crucial qualities. Perhaps, this is altogether an opportunity for building novel interfaces that enable blind users to control the verbosity of audio descriptions. Verbosity controls are common in screen readers and are explored in many assistive applications for the blind *e.g.,* in mobility and orientation [23] and real-time pedestrian detection [29].

6.1 Mixed-Ability Collaborative SD Authoring

Our findings provide rich insights on the opportunities and current limitations of deploying collaborative SD authoring as online work. ViScene introduces a new way for co-creating audio descriptions to increase video accessibility. More so, it hints at the potential for online mixed-ability collaborations between sighted authors and blind reviewers. Our results showed strong improvement in SD's *Objective* quality, which highlights the importance of involving blind people in the process to improve overall SD quality. Although when comparing sighted to blind evaluator, we find that they are often not aligned with blind evaluator, sometimes displaying more generosity in approving some of the qualities, but more critical on others. Thus, it is essential to consider the blind review and evaluation, especially in areas of disagreement (*e.g.,* Learning), so as to ensure the generated SDs are favored by the blind audience.

While we observed that novices can author good SDs with reviews, recruiting and onboarding good sighted and blind reviewers remain an open question. Perhaps as suggested in YouDescribe [47] this could be people already within or familiar with the blind community, content creators, or music/TV show fans. We also see an opportunity in recruiting good authors to be reviewers and onboarding them by designing practice activities based on our codebook. For instance, we envision engineering accessible onboarding activities using sample videos and a variety of SDs depicting good and bad examples across all quality dimensions.

6.2 Cost Effectiveness of ViScene

We believe that ViScene is a cost-effective and more readily available option for creating SDs. The average cost of employing people to author and review SDs is between US$2.81 pvm to US$5.48 pvm, which is much cheaper in comparison to the cost of professional service that ranges from US$12 pvm to US$75 pvm [45], though, the quality insurance may not be comparable. These figures should encourage video content creators to provide more accessible contents for their viewers by including SD. Having said that, in an era where 500 hours of videos are uploaded every minute on YouTube alone [20], it is improbable that ViScene would be able to generate SDs for every single video. In our study, the average time to create SD for videos, where each video is around one minute long, is 50–56 minutes. By extrapolation, it would take almost a day to generate an SD for a video that lasts 24 minutes, which renders ViScene to be not be scalable, *e.g.,* for long videos. Thus, as we describe next, automating some of the processes in SD authoring is critical. Also, we need to put extra

care in extrapolating as the cost may not increase linearly and the cost calculation result might not be applicable for longer videos. But online videos without ADs that are about few minutes long, similar length to the videos we used in our study, are prevalent. We believe such videos could benefit from more cost-effective ADs created with ViScene.

6.3 Potentials for Automation

The authoring and reviewing processes employed in this version of ViScene were all manual. The SD authors did not receive commentary feedback in real-time, which can be disruptive for the overall process. We are currently investigating the efficacy of automating some of the feedback and we see a lot of room for innovation in this area. We believe less subjective dimensions like Descriptive are easier to comment on automatically. For example, we can train natural language processing models to estimate the SD descriptiveness and provide automated feedback and encourage novice authors for more information. Of course, another fertile research topic would be automating the authoring of SDs itself. We are excited to see a few efforts this past year in this direction (*e.g.,* [49, 51, 52]), though there is much to be done for achieving high-quality audio descriptions.

7 LIMITATIONS

The full array of different types of videos was not explored due to the design of the user study. The video durations were short and the types of videos were also limited to optimize the output of quality SDs from our participants. Investigating the effect of the video duration and other video types on the SD authoring process remains a future work. ViScene reviewers and evaluators were members of the research team. We opted for this methodological design to ensure that the reviewers are fully dedicated to giving good feedback (typically, in a real-world setting reviewers could choose videos that they might be more vested in) and evaluators have a good understanding of the codebook. Having a fixed set of reviewers could have negatively affected the feedback quality, but we have mitigated the potential effects with counterbalancing. This was done by incorporating our codebook into the ViScene interface. Our next step is to explore the efficacy of employing reviewers in a real-world setting *e.g.,* by partnering with YouDescribe [50] and recruiting evaluators outside our research group.

A potential limitation is the fact that evaluations are not completely objective even through the evaluators followed the codebook that is based on professional guidelines.

To assess this, we measured inter-coder agreement across evaluations and observed that we had a high degree of agreement for quality dimensions such as *Succinct*, 90%, *Referable* (76%), *Accurate* (74%), *Sufficient* (74%), *Objective* (73%), *Learning* (70%), *Descriptive* (69%) but not so much for *Clarity* (69%), and *Interest* (56%).

8 CONCLUSION

We designed and developed ViScene, a system that enables sighted novices to author scene description—textual descriptions of scenes in a video that are converted into audio descriptions through text-to-speech with commentary feedback coming from sighted or

blind reviewers. Through a study with novice sighted scene description authors (N=60), we explored whether the sighted or blind review could improve the quality of scene descriptions. We found that those receiving feedback were able to improve qualities like *Descriptive*, *Objective*, *Referable*, and *Clarity* in their scene descriptions, which is promising. However, this was not the trend for other qualities such as *Learning* and *Sufficient*, which remained low. We discussed potential ways forward for improving the collaborative process and blind users experience with audio descriptions, as well as the role of automation in future research directions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under its Strategic Capabilities Research Centres Funding Initiative and a Lee Kong Chian Fellowship. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. Hernisa Kacorri is partially supported by the National Institute on Disability, Independent Living, and Rehabilitation Research (NIDILRR), ACL, HHS (#90REGE0008). We would to thank the Rikki Donachie and his partner, Yern Har Tse who helped us in generating the Instructional video ground truth.

REFERENCES

[1]. 3PlayMedia. 2020. Beginner's Guid to Audio Description. https://go.3playmedia.com/hubfs/WP%20PDFs/Beginners-Guide-to-Audio-Description.pdf. Accessed: 2021-01-13.

[2]. Amazon. 2020. Amazon Polly. https://aws.amazon.com/polly/. Accessed: 2020-06-01.

[3]. Anderson Myriah. 2020. The 13 most popular types of videos on YouTube [Infographic]. https://www.impactplus.com/blog/most-popular-types-of-videos-on-youtube-infographic. Accessed: 2020-11-6.

[4]. Balcazar Fabricio, Hopkins Bill L, and Suarez Yolanda. 1985. A critical, objective review of performance feedback. Journal of Organizational Behavior Management 7, 3–4 (1985), 65–89.

[5]. Branham Stacy M and Kane Shaun K. 2015. Collaborative accessibility: How blind and sighted companions co-create accessible home spaces. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2373–2382.

[6]. Branham Stacy M and Kane Shaun K. 2015. The invisible work of accessibility: how blind employees manage accessibility in mixed-ability workplaces. In Proceedings of the 17th international acm sigaccess conference on computers & accessibility. 163–171.

[7]. Branje Carmen J and Fels Deborah I. 2012. Livedescribe: can amateur describers create high-quality audio description? Journal of Visual Impairment & Blindness 106, 3 (2012), 154–165.

[8]. Braun Sabine. 2011. Creating coherence in audio description. Meta: Journal des traducteurs/Meta: Translators' Journal 56, 3 (2011), 645–662.

[9]. Caldwell Ben, Cooper Michael, Loretta Guarino Reid Gregg Vanderheiden, Chisholm Wendy, Slatin John, and White Jason. 2008. Web content accessibility guidelines (WCAG) 2.0. WWW Consortium (W3C) (2008).

[10]. Campos Virginia P, Araújo Tiago MU de, Souza Filho Guido L de, and Gonçalves Luiz MG. 2020. CineAD: a system for automated audio description script generation for the visually impaired. Universal Access in the Information Society 19, 1 (2020), 99–111.

[11]. Clement J. 2019. Hours of video uploaded to YouTube every minute as of May 2019. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute#statisticContainer. Accessed: 2020-11-5.

[12]. Audio Description Coalition. 2009. Standards for Audio Description and Code of Professional Conduct for Describers. https://audiodescriptionsolutions.com/wp-content/uploads/2016/06/adc_standards_090615.pdf. Accessed: 2020-11-6.

[13]. Comcast. 2020. Comcast 2020 Network Report. https://update.comcast.com/wp-content/uploads/sites/33/dlm_uploads/2021/02/network-report-2020.pdf.

[14]. Federal Communications Commission. 2020. 21st Century Communications and Video Accessibility Act (CVAA). https://www.fcc.gov/consumers/guides/21st-century-communications-and-video-accessibility-act-cvaa. Accessed: 2020-11-6.

[15]. Described and Captioned Media Program. 2020. Described and Captioned Media Program (DCMP). http://www.descriptionkey.org/quality_description.html. Accessed: 2019-03-19.

[16]. Donachie Rikkie. 2013. It's Just a Bit of Paper (Animal, Birds & Cards, Volume 1). CreateSpace Independent Publisihing Platform.

[17]. Fryer Louise. 2016. An introduction to audio description: A practical guide. Routledge.

[18]. Gagnon Langis, Chapdelaine Claude, Byrns David, Foucher Samuel, Heritier Maguelonne, and Gupta Vishwa. 2010. A computer-vision-assisted system for video description scripting. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 41–48.

[19]. GBH. 2020. CADET - Caption and Descriptive Editing Tool. https://www.wgbh.org/foundation/what-we-do/ncam/cadet. Accessed: 2020-11-6.

[20]. Hale James. 2019. More Than 500 Hours Of Content Are Now Being Uploaded To YouTube Every Minute. https://www.tubefilter.com/2019/05/07/number-hours-video-uploaded-to-youtube-per-minute/. Accessed: 2020-11-5.

[21]. Hattie John and Timperley Helen. 2007. The power of feedback. Review of educational research 77, 1 (2007), 81–112.

[22]. World-Wide Web COnsortium Web Accessibility Initiative. 2016. Making the Web-Accessible. https://www.w3.org/WAI/. Accessed: 2020-11-6.

[23]. Kacorri Hernisa, Mascetti Sergio, Gerino Andrea, Ahmetovic Dragan, Alampi Valeria, Takagi Hironobu, and Asakawa Chieko. 2018. Insights on Assistive Orientation and Mobility of People with Visual Impairment Based on Large-Scale Longitudinal Data. ACM Trans. Access. Comput 11, 1, Article 5 (March 2018), 28 pages. 10.1145/3178853

[24]. Kluger Avraham N and DeNisi Angelo. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological bulletin 119, 2 (1996), 254.

[25]. Kobayashi Masatomo, Fukuda Kentarou, Takagi Hironobu, and Asakawa Chieko. 2009. Providing synthesized audio description for online videos. In Proceedings of the 11th international ACM SIGACCESS conference on Computers and accessibility. 249–250.

[26]. Kobayashi Masatomo, Trisha O'Connell Bryan Gould, Takagi Hironobu, and Asakawa Chieko. 2010. Are synthesized video descriptions acceptable?. In Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility. 163–170.

[27]. Kulkarni Chinmay E, Bernstein Michael S, and Klemmer Scott R. 2015. Peer-Studio: rapid peer feedback emphasizes revision and improves performance. In Proceedings of the second (2015) ACM conference on learning@ scale. 75–84.

[28]. Lakritz James and Salway Andrew. 2006. The semi-automatic generation of audio description from screenplays. Dept. of Computing Technical Report CS-06–05, University of Surrey (2006).

[29]. Lee Kyungjun, Sato Daisuke, Asakawa Saki, Kacorri Hernisa, and Asakawa Chieko. 2020. Pedestrian Detection with Wearable Cameras for the Blind: A Two-Way Perspective (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. 10.1145/3313831.3376398

[30]. Leung Hoi Ching Dawning. 2018. Audio description of audiovisual programmes for the visually impaired in Hong Kong. Ph.D. Dissertation. UCL (University College London).

[31]. 3Play Media. 2020. 3Play Plugin. https://www.3playmedia.com/services/features/plugins/3play-plugin/. Accessed: 2020-11-6.

[32]. Mediakix. 2019. The Most Popular Types of YouTube Video. https://mediakix.com/blog/most-popular-youtube-videos/. Accessed: 2020-11-6.

[33]. Morris Meredith Ringel, Johnson Jazette, Bennett Cynthia L, and Cutrell Edward. 2018. Rich representations of visual content for screen reader users. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. 1–11.

[34]. Natalie Rosiana, Jarjue Ebrima, Kacorri Hernisa, and Hara Kotaro. 2020. ViScene: A Collaborative Authoring Tool for Scene Descriptions in Videos. In The 22nd International ACM SIGACCESS Conference on Computers and Accessibility. 1–4.

[35]. Netflix. 2020. Audio Description Style Guide v2.1. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215510667-Audio-Description-Style-Guide-v2-1. Accessed: 2020-11-6.

[36]. Department of Labor US 2021. Minimum Wage. https://www.dol.gov/general/topic/wages/minimumwage. Accessed: 2021-1-13.

[37]. American Council of the Blind. 2020. Sample of Audio Description. https://www.acb.org/adp/samples.html. Accessed: 2020-11-6.

[38]. American Council of the Blind. 2021. Audio Description using the Web Speech API. https://acb.org/adp/education.html. Accessed: 2020-01-13.

[39]. Packer Jaclyn, Vizenor Katie, and Joshua A Miele. 2015. An overview of video description: history, benefits, and guidelines. Journal of Visual Impairment & Blindness 109, 2 (2015), 83–93.

[40]. Pauls Jamie. 2016. Audio Description Comes to Netflix. https://www.afb.org/aw/16/7/15436. Accessed: 2021-1-13.

[41]. Pauls Jamie. 2016. Netflix Audio Description: What a Difference a Year Makes. https://www.afb.org/aw/17/7/15312. Accessed: 2021-1-13.

[42]. <j/>Pavel Amy, Reyes Gabriel, and Bigham Jeffrey P. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology. 747–759.

[43]. Player Able. 2020. Able Player: Fuly Accessible cross-browser HTML Media Player. https://www.3playmedia.com/services/features/plugins/3play-plugin/. Accessed: 2020-11-6.

[44]. Slatin John M. 2001. The art of ALT: toward a more accessible Web. Computers and Composition 18, 1 (2001), 73–81.

[45]. Thompson Terril. 2017. My Audio Description Talk @ CSUN). https://terrillthompson.com/813. Accessed: 2020-11-6.

[46]. Thompson Terril. 2019. Audio Description using the Web Speech API. https://terrillthompson.com/1173. Accessed: 2020-11-6.

[47]. Veroniiiica. 2019. How to Create Audio Description for YouTube with YouDescribe. https://www.perkinselearning.org/technology/blog/how-create-audio-description-youtube-youdescribe. Accessed: 2021-7-2.

[48]. Walczak Agnieszka and Fryer Louise. 2018. Vocal delivery of audio description by genre: measuring users' presence. Perspectives 26, 1 (2018), 69–83.

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

[50]. YouDescribe. 2020. YouDescribe. https://youdescribe.org/support/tutorial. Accessed: 2020-11-6.

[51]. Beste F Yuksel Pooyan Fazli, Mathur Umang, Bisht Vaishali, Soo Jung Kim Joshua Junhee Lee, Seung Jung Jin Yue-Ting Siu, Miele Joshua A, and Yoon Ilmi. 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.

[52]. Beste F Yuksel Soo Jung Kim, Seung Jung Jin Joshua Junhee Lee, Fazli Pooyan, Mathur Umang, Bisht Vaishali, Yoon Ilmi, Siu Yue-Ting, and Miele Joshua A. 2020. Increasing Video Accessibility for Visually Impaired Users with Human-in-the-Loop Machine Learning.

In Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. 1–9.

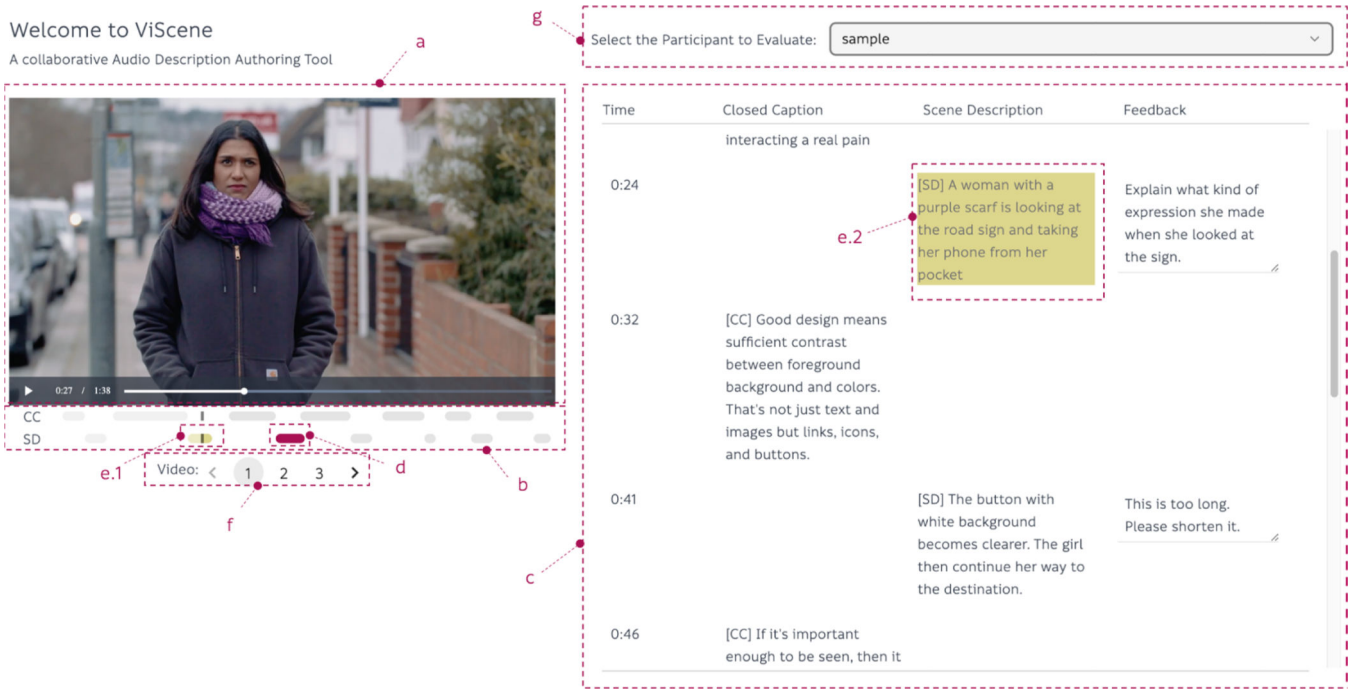

Figure 1:

ViScene's interface. (a) the video pane; (b) closed captions (CC) and scene descriptions (SD) bars; (c) a table with Time, CC, SD, and Feedback columns; (d) SD succinctness feedback, (e) CC/SD text-segment correspondence visualization, (f) video selector, and (g) author dropdown selector (for reviewers).

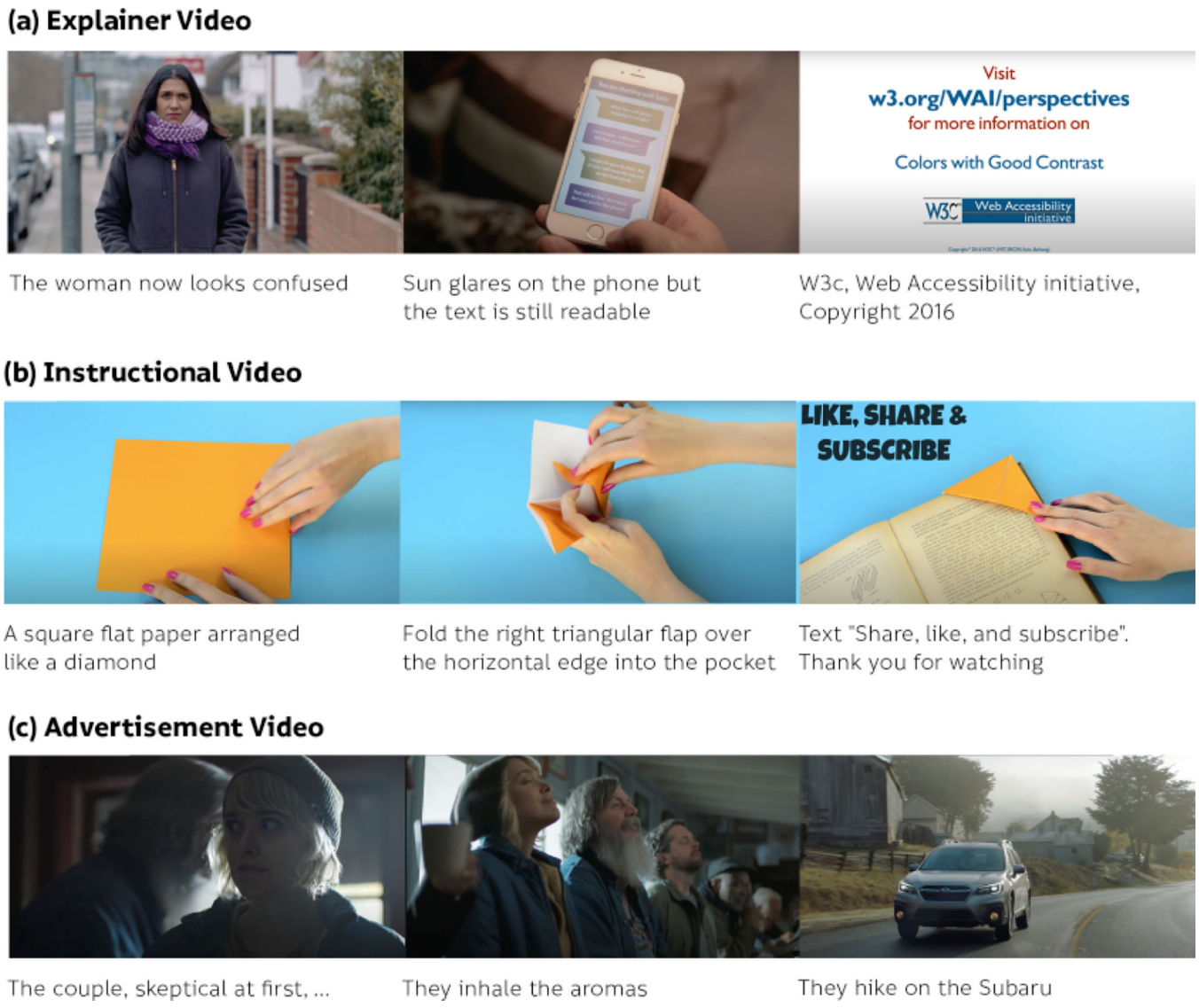

Figure 2:

The screenshots of scenes in Explainer, Instructional, and Advertisement Videos. Each row represents one of the videos. Texts below the rows are the ground truth SDs for their respective screenshots.

Figure 3:

The Sighted Evaluation approval counts for each SD quality variable across feedback conditions and sessions. Two sighted evaluator independently assessed the quality of 360 SDs (60 participants conditions x 2 sessions x 3 videos),. The final approvals were decided through discussion and resolving disagreements between the two sighted evaluators. A line from a hole to a filled circle shows the change in approvals.

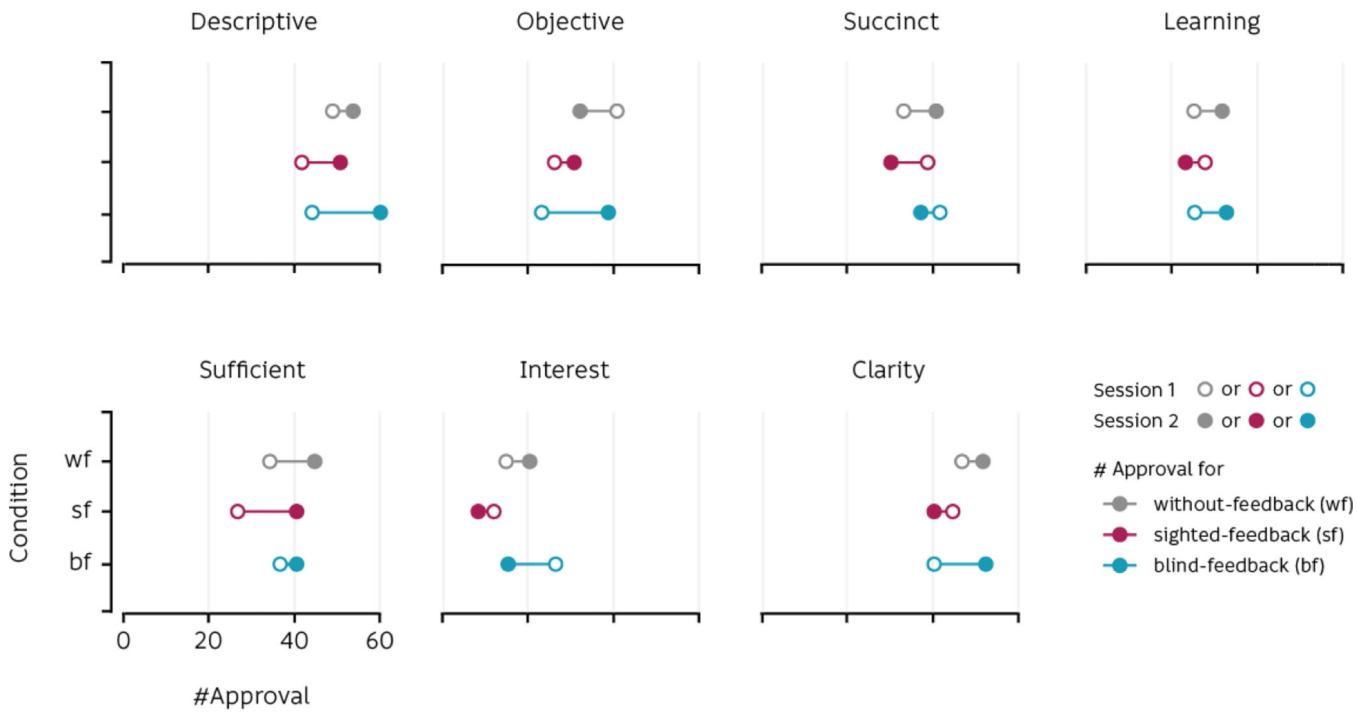

Figure 4:

The Blind Evaluation approval counts for each SD quality variable. Like Fig. 3, each line from a hole to a filled circle shows a change in approval counts. A blind evaluator assessed the quality of 360 SDs (60 participants conditions ×2 sessions × 3 videos) along each variable.

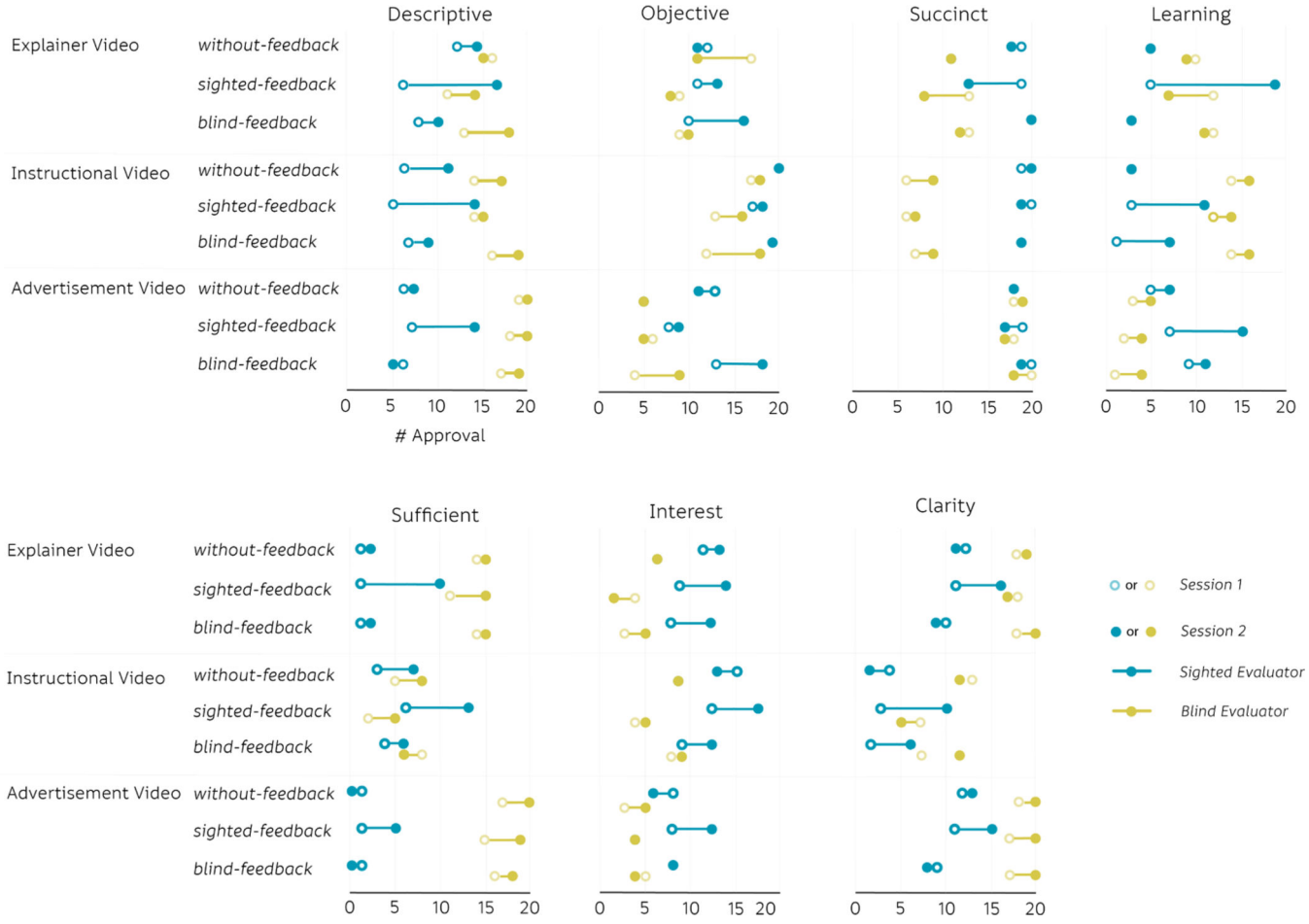

Figure 5:

Approval count for the codes that were evaluated both by sighted and blind evaluators. The x-axis indicates the count of approval and the y-axis represents the type of the videos and the conditions.

Table 1:

Scene description quality codebook.

Variable	Description
Descriptive	SD provides a pictorial and kinetic description of objects, people, and settings and explains how people act in the scene. SD audience should be able to imagine the visual properties of objects and how people are moving and acting in the scene, which allows for a basic understanding of the video's narrative. For example, SD should explain if a character in the scene is old, what color is the cloth that a character is wearing, that a character is interacting with a phone, and so on. The SD should not lead the audience into needing further clarification of having misinterpreted the scene. For example, "a blue street sign" indicates that the street sign has a color of blue and it is located on the street. [1, 15, 44]
Objective	SD illustrates objects, people, or relationships between them in an unbiased manner using objective language. SD should not include a speculative description. SD should also avoid making their own inference about what is occurring in the scene without proper evidence. SD should also avoid using adjectives to create a speculative description which could lead to different personal interpretation among the audience. For example, if female and male characters are standing in a scene, SD should describe characters as "female and male characters" without speculating they are a couple. SD should not use languages like "the text is hard to read" but instead objectively describe colors, text size, and other features that contributed to hard reading. [1, 12, 15, 44]
Succinct	An audio generated from textual SD should fit in a gap without a dialogue or a natural pause in the video. [15, 44]
Learning	SD should convey the video's intended message to the audience. For example, in Explainer and Instructional videos, the audience should learn the intended material described in the video. And, for videos like Advertisement video, the audience should be able to learn the story of the video by listening to audio description generated from SD. [15]
Sufficient	SD should depict all the scenes and provide sufficient information for the audience to comprehend the content of a video while not being overly descriptive. SD should capture important overviews of the scene to help the audience comprehend what is happening in the scene, but it should not describe insignificant details that are less relevant in comprehending the scene.[33]
Accurate (^)	SD should not provide the incorrect information of what is shown in the scene *(e.g.,* "green street sign" instead of "blue street sign"). SD should also be grammatically correct. However, minor grammatical errors that have no impact on the conveyance of the scene should not be penalized in an evaluation. Assessment of accuracy requires visual inspection of scenes. Thus, the code is used only by sighted reviewers and sighted evaluators. [15, 44]
Referable (^)	SD should use language that is accessible to everyone with different disabilities. The use of demonstrative pronouns like "this", "there", "that" is not Referable as it is not understandable for people with visual impairments because these pronouns need to be complemented with visual help. [15]
Interest	SD should make the video be interesting for the audience by writing a cohesive narrative. The tone of the description should reflect the tone of the video. For example, if the video shows a fun story, the SD should use vocabulary that would convey that tone [17, 35, 48]
Clarity	SD should communicate descriptive information in a language and manner that are easy to follow for people with visual impairments. The audience should know the presence of characters or objects in a scene before SD describes their properties, relationship with other characters/objects, their action and motion, and so on. [48]

Note:

^ = these variables are assessed only by the sighted evaluator.

ASSETS. Author manuscript; available in PMC 2022 February 18.

Table 2:

The sample of SD and the comments from the reviewers.

Scene Description from Participants	Source of Comments	Reviewer Comments
Road sign with dull colours, woman looks confused at the sign	Sighted reviewer	• Do describe the sign in relation to the topic of contrast e.g. sign had low or high contrast • Do also mention what was the woman's action after she looks confused. It is good to describe the scene completely.
The lady was able to find the place using her phone	Blind reviewer	• Confusing phrasing.Where this applies: The use of the word "the place" is not specific, and makes the listener wonder what this refers to. • Suggestion: Be more specific about where the woman is. e.g. Use phrases like "her destination" or describe the name of the location and the fact that the woman was trying to reach this location.

Table 3:

The Sighted Evaluation main effects and interaction effects of *feedback type* and session from the GLMM-based analysis to assess changes in SD qualities. SF: *sighted-feedback*, BF: *blind-feedback*.

		Descriptive	Objective	Succinct	Learning	Sufficient	Interest	Clarity	Accurate	Referable
Main	SF		.							
	BF									
	session									
Interaction	SF × session	**		*	***	.	*	**		*
	BF × session		**							

. p <0.1

* p <0.05

** p <0.01

*** p <0.001

ASSETS. Author manuscript; available in PMC 2022 February 18.

Table 4:

The main effects in the post-hoc test result for sighted evaluator. SF: *sighted-feedback*, BF: *blind-feedback*.

Factor	Descriptive	Objective	Succinct	Learning	Sufficient	Interest	Clarity	Accurate	Referable
SF	.			***	***	.	**		**
BF		*							

. $p < 0.1$

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

Table 5:

The blind evaluation main effects and interaction effects of *feedback type* and *session* from the GLMM-based analysis to assess changes in SD qualities. SF: *sighted-feedback*, BF: *blind-feedback*

		Descriptive	Objective	Succinct	Learning	Sufficient	Interest	Clarity
Main	SF		*			.		
	BF		**					
	session							
Interaction	SF × session							
	BF × session		*					

. $p < 0.1$

* $p < 0.05$

** $p < 0.01$

*** $p < 0.001$

Table 6:

Post hoc test result for blind evaluator. SF: *sighted-feedback*, BF: *blind-feedback*

Factor	Descriptive	Objective	Succinct	Learning	Sufficient	Interest	Clarity
sighted-feedback							*
blind-feedback							

* $p < 0.05$

ASSETS. Author manuscript; available in PMC 2022 February 18.

Table 7:

A confusion matrices of agreement counts between the blind and sighted evaluators. The count value in the table is the sum from the two sessions. (A: Approve, R: Reject)

		Descriptive		Objective		Succinct		Learning		Sufficient		Interest		Clarity	
		Sighted Eval.		Sighted Eval.		Sighted Eval.		Sighted Eval.		Sighted Eval.		Sighted Eval.		Sighted Eval	
		A	R	A	R	A	R	A	R	A	R	A	R	A	R
Blind Eval.	Approve	122	173	158	34	211	11	52	114	33	190	90	5	144	132
	Reject	25	40	100	68	125	13	70	124	31	106	246	19	20	64

ASSETS. Author manuscript; available in PMC 2022 February 18.

Table 8:

The completion time (in minutes) to author or review SDs for three videos in different sessions and *feedback type*. We also present the total duration, which is the sum of *session 1*, review, and *session 2* durations. WF: *without-feedback*, SF: *sighted-feedback*, BF: *blind-feedback*.

	Session 1 **(min)**			**Review (min)**			***Session 2*** **(min)**			**Total (min)**		
	Med.	**Mean**	**SD**	**Med.**	**Mean**	**SD**	**Med.**	**Mean**	**SD**	**Med.**	**Mean**	**SD**
WF	46	53.8	22	N/A	N/A	N/A	6.6	20.5	26.3	57.3	74.3	38.9
SF	43.0	46.7	11.1	48.2	51.8	11.2	47	52.3	11.4	156.5	150.8	22.3
BF	51.5	57.5	11.3	57.8	61.7	10.9	52.1	50.4	15.2	166.4	169.8	27.7

Table 9:

The estimated cost to generate SD for different *feedback types* and videos. We obtain the value of cost (in US$) by multiplying the total time duration with the US federal minimum wage, which is US$ 7.25 per hour [36]. We further compute the cost per minute for each video (presented in pvm). Video durations are 1 min 38 seconds, 1 min 42 seconds, and 1 min for the Explainer, Instructional, and Advertisement videos. SF: *sighted-feedback*, BF: *blind-feedback*.

		Total (s)			Cost (US$)			Cost per minute (pvm)		
	Video	**Med.**	**Mean**	**SD**	**Med.**	**Mean**	**SD**	**Med.**	**Mean**	**SD**
SF	Explainer	2279	2701	1394	4.59	5.44	2.81	2.81	3.33	1.72
	Instructional	3885	3856	1164	7.82	7.77	2.34	4.60	4.57	1.38
	Advertisement	2308	2488	1067	4.65	5.01	2.15	4.65	5.01	2.15
BF	Explainer	2742	3043	1652	5.52	6.13	3.33	3.38	3.75	2.04
	Instructional	4627	4610	1472	9.32	9.28	2.96	5.48	5.45	1.74
	Advertisement	2490	2537	1109	5.01	5.11	2.23	5.01	5.11	2.23