Exploring Community-Driven Descriptions for Making Livestreams Accessible

Daniel Killough
The University of Texas at Austin
Department of Computer Science
Austin, Texas, USA
contact@dkillough.com

Amy Pavel
The University of Texas at Austin
Department of Computer Science
Austin, Texas, USA
apavel@cs.utexas.edu

ABSTRACT

People watch livestreams to connect with others and learn about their hobbies. Livestreams feature multiple visual streams including the main video, webcams, on-screen overlays, and chat, all of which are inaccessible to livestream viewers with visual impairments. While prior work explores creating audio descriptions for recorded videos, live videos present new challenges: authoring descriptions in real-time, describing domain-specific content, and prioritizing which complex visual information to describe. We explore inviting livestream community members who are domain experts to provide live descriptions. We first conducted a study with 18 sighted livestream community members authoring descriptions for livestreams using three different description methods: live descriptions using text, live descriptions using speech, and asynchronous descriptions using text. We then conducted a study with 9 livestream community members with visual impairments, who shared their current strategies and challenges for watching livestreams and provided feedback on the community-written descriptions. We conclude with implications for improving the accessibility of livestreams.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in accessibility**; *Computer supported cooperative work.*

KEYWORDS

Live Video Streaming, Livestreaming, Accessibility, Visual Impairments, Blind and Low Vision, Audio Descriptions

ACM Reference Format:
Daniel Killough and Amy Pavel. 2023. Exploring Community-Driven Descriptions for Making Livestreams Accessible. In *The 25th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '23), October 22–25, 2023, New York, NY, USA.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3597638.3608425

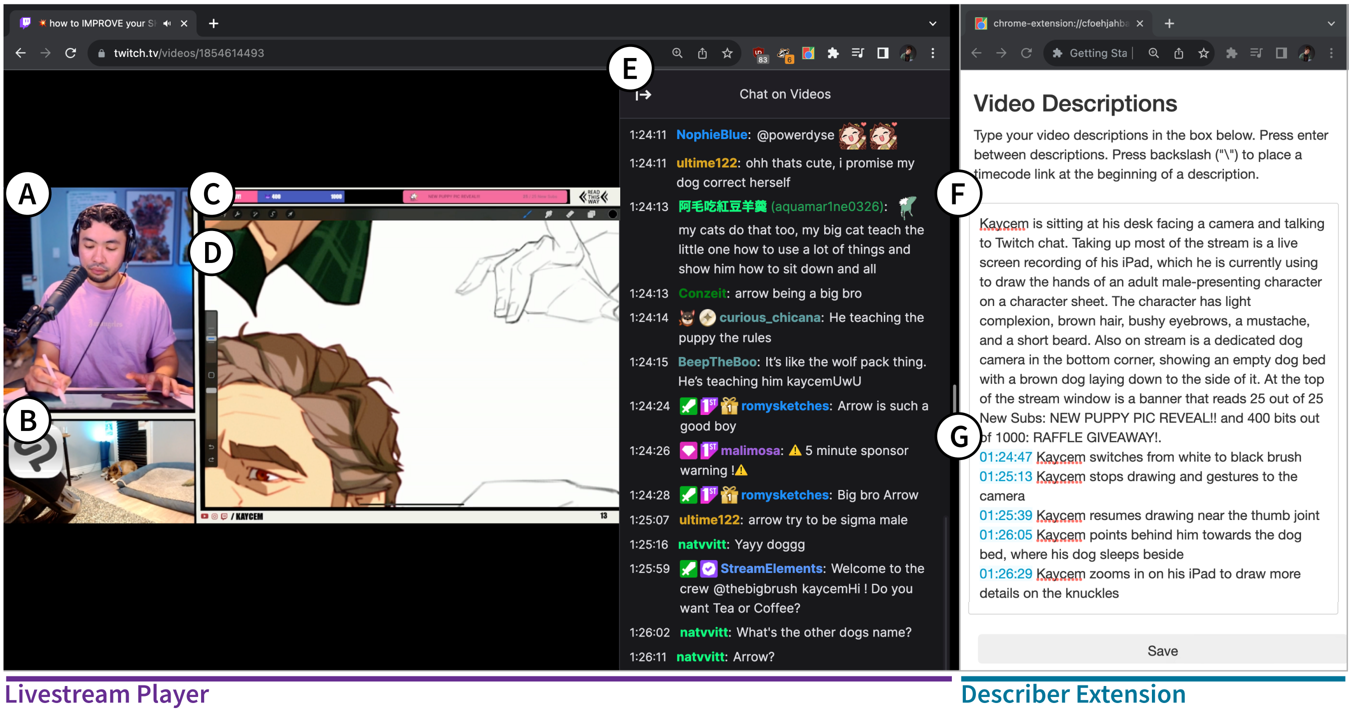
ASSETS '23, October 22–25, 2023, New York, NY, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0220-4/23/10.
https://doi.org/10.1145/3597638.3608425

1 INTRODUCTION

Live videos (i.e. "livestreams") shared on streaming services like YouTube [48], Facebook [29], and Twitch [45] are becoming increasingly popular. People authoring livestreams (i.e. "streamers") broadcast long activities in real-time, such as playing games [12], producing art [10], leading educational exercises [5], coding [9], or exploring the outdoors [28]. During a livestream, the streamer and audience synchronously interact with each other via the streamer's live video — including webcams (Figure 1A-B), on-screen overlays (Figure 1C), and main activity video (Figure 1D) — as well as with the audience live chat (Figure 1E). Real-time interaction provides viewers a chance to suggest next steps, ask for clarifications, and discuss reactions. The format of livestreams thus enables online communities to form around shared experiences [12], and even extend offline [41]. However, the rich visual content of livestreams that affords community building is not accessible to people with visual impairments.

To make recorded videos accessible, people add narration of the important visual content in the video, i.e. *audio descriptions*. Prior work explored how to create audio descriptions for recorded videos such as films [3, 33, 43], user-generated videos [18, 26, 27, 34, 46], slide presentations [35, 36] and GIFs [11] by providing computational description support [27, 34, 35, 46, 50] and proposing what to describe for specific video types (*e.g.*, GIFs [11], films [43]). Previous work has not yet explored technology to support live descriptions or description preferences for livestream-specific content (*e.g.*, long expert streams). Jun et al. investigated the accessibility of livestreaming for *streamers* with visual impairments [21]. Streamers have accessibility needs that overlap with those of viewers (*e.g.*, accessing chat), yet it remains unclear how to make live videos accessible. While online services provide on-demand live descriptions [6, 8], streams are challenging to understand for people without domain familiarity due to complex visual content (*e.g.*, multiplayer gameplay, expert software). Following the success of community-driven efforts to make videos accessible for d/Deaf and Hard of Hearing audience members via fansubbing [25] or community captions [15], and drawing on community sourcing [13, 23], we invite sighted livestream viewers familiar with the livestream content to make livestreams non-visually accessible.

We present two studies exploring the feasibility of community-driven livestream accessibility. We first invited 18 sighted livestream community members to author descriptions of livestreams in domains that they were familiar with to compare three description approaches: live description using voice, live description using text, and asynchronous description using text. We also interviewed 9 livestream community members with visual impairments to share



1:24:11 NophieBlue: @powerdyse

1:24:11 ultime122: ohh thats cute, i promise my dog correct herself

1:24:13 阿毛吃紅豆羊羹 (aquamar1ne0326): my cats do that too, my big cat teach the little one how to use a lot of things and show him how to sit down and all

1:24:13 Conzeit: arrow being a big bro

1:24:14 curious_chicana: He teaching the puppy the rules

1:24:15 BeepTheBoo: It's like the wolf pack thing. He's teaching him kaycemUwU

1:24:24 romysketches: Arrow is such a good boy

1:24:26 malimosa: ⚠ 5 minute sponsor warning !⚠

1:24:28 romysketches: Big bro Arrow

1:25:07 ultime122: arrow try to be sigma male

1:25:16 natvvitt: Yayy doggg

1:25:59 StreamElements: Welcome to the crew @thebigbrush kaycemHi ! Do you want Tea or Coffee?

1:26:02 natvvitt: What's the other dogs name?

1:26:11 natvvitt: Arrow?

Video Descriptions

Type your video descriptions in the box below. Press enter between descriptions. Press backslash ("\") to place a timecode link at the beginning of a description.

Kaycem is sitting at his desk facing a camera and talking to Twitch chat. Taking up most of the stream is a live screen recording of his iPad, which he is currently using to draw the hands of an adult male-presenting character on a character sheet. The character has light complexion, brown hair, bushy eyebrows, a mustache, and a short beard. Also on stream is a dedicated dog camera in the bottom corner, showing an empty dog bed with a brown dog laying down to the side of it. At the top of the stream window is a banner that reads 25 out of 25 New Subs: NEW PUPPY PIC REVEAL!! and 400 bits out of 1000: RAFFLE GIVEAWAY!.
01:24:47 Kaycem switches from white to black brush
01:25:13 Kaycem stops drawing and gestures to the camera
01:25:39 Kaycem resumes drawing near the thumb joint
01:26:05 Kaycem points behind him towards the dog bed, where his dog sleeps beside
01:26:29 Kaycem zooms in on his iPad to draw more details on the knuckles

Figure 1: The *Livestream Player* (left) features the livestream (A-D) and audience live chat (E). The livestream includes webcams for the streamer (A) and a dog (B), an overlay with status indicators (C), and the main video displaying a screenshare of a creative application (D). The *Describer Extension* (right) enables describers to input text descriptions while the livestream plays. Pressing the backslash key while a Livestream Player window is open inserts the current video timecode into the textbox (F). Clicking a timecode (G) seeks the Livestream Player video to the corresponding playback time. Source: Twitch livestream *How to IMPROVE your SKILLS QUICKLY! Character Design Bootcamp #2 Day 06/30 !bootcamp !youtube !resources* by Kaycem [22].

their current livestream viewing practices and challenges and provide feedback on the descriptions written by community members.

Overall, sighted community members generated descriptions that increased the accessibility of livestreams using all description methods. While sighted community members found it more challenging to provide live rather than asynchronous descriptions, they adapted several strategies to successfully create live descriptions including: describing during the streamer's narration, using domain-specific terms to quickly author descriptions (*e.g.*, "Up-B" to describe a character's special attack in a game), and primarily describing updates due to individual actions (i.e. play-by-plays) rather than the scene as a whole. For providing live descriptions, community members differed in their preference for text vs. voice for description input. However, community members provided significantly more descriptions and description words per video minute using voice input than using text input for live descriptions.

Community members with visual impairments reported accessibility issues with consuming livestreams due to the platform's interface and the livestream content itself. Though most community members with visual impairments interviewed use YouTube instead of Twitch to avoid platform accessibility issues, livestream content remained inaccessible. Community members reported that the streamers' speech often diverged from describing their actions (*e.g.*, telling a story while creating an art piece) and used frequent visual references to other parts of the video that were difficult to understand (*e.g.*, reacting to an unknown chat message or referring an on-camera event). Viewers found community-written descriptions to be valuable in understanding the video as they filled in gaps left by the speaker. Viewers also suggested improvements for future descriptions, such as providing adjustable preferences on the expertise level, level of detail, and amount of overlap with the audio channel. We conclude with directions for future systems aiming to make livestreams accessible.

In summary, we contribute:

- An exploratory study with livestream community members providing descriptions of live video
- Interviews with livestream viewers with visual impairments sharing current strategies and challenges for watching livestreams
- Description preferences from livestream viewers with visual impairments derived from a co-watching exercise and feedback on community-written descriptions

2 BACKGROUND

Our work builds upon prior work in video, livestreaming accessibility, and crowdsourcing for accessibility.

2.1 Video Accessibility

To make videos accessible to people with visual impairments, professionals traditionally create *audio descriptions*, or narrations of the important visual content in a scene that cannot be understood from the audio alone [38]. While audio descriptions increasingly exist for films and TV, they rarely exist for user-generated content. Prior work developed tools to make authoring audio descriptions easier by generating them automatically [46], or aiding novices in authoring audio descriptions [3, 18, 27, 31, 32, 34, 50]. For example, prior work helped novices edit their descriptions to fit into times without narration [34], identify parts of the video likely to be inaccessible [27], host their descriptions [18], gain feedback on their descriptions [32], and locate silences [3, 34]. These systems all process recorded videos rather than live videos, such that they are not suitable for livestreams. Such video accessibility work also explored generally understandable visual content rather than the domain-specific visual content present in livestreams. We explore how community member familiar with the domain may be able to provide descriptions for live rather than recorded videos.

While audio descriptions typically occur within gaps in video narration [3, 34], adequate gaps do not always occur (*e.g.*, for short videos [11], or videos with frequent speech [34, 35]). To address this time constraint, prior work used rich audio to convey video themes [11], and provided users control over how often or when to pause a video to receive additional descriptions [34, 35]. Live video presents new time constraints for describers aiming to describe content as it happens, as well as for listeners aiming to keep up with the video pace. We investigate the feasibility of producing and consuming descriptions under such time constraints.

2.2 Livestreams and Accessibility

Livestreaming, broadcasting live video over the internet, has grown over recent decades with increased internet speeds and a broad selection of platforms (*e.g.*, justin.tv now Twitch, Facebook Live, YouTube Live, TikTok LIVE). We discuss livestream features common on platforms such as Twitch and YouTube Live to reflect on implications for accessibility for viewers with visual impairments:

Long, real-time broadcasts: As livestreams are broadcast in real time, streams are often unedited and occur over long durations (*e.g.*, up to 5 hours or more [28]). Compared to edited, recorded videos, livestreams activate communities around watching the content in real time [42], engage viewers with one another for more time [12], and enable viewers to gain depth in the streamed activity (*e.g.* watching a game rather than highlights; seeing an artist work instead of explain the high level steps). While viewers may watch the livestream for a long time (*e.g.*, 5 hours [28]), they may join in the middle of the stream and need to "catch up" [47] with what occurred earlier in the broadcast. As audio describing videos typically occurs during post-processing and requires additional editing, existing methods posed by prior work for novice use are difficult to use in real time [3, 18, 27, 34, 46]. Recent work explored sonifying live tennis matches [19], but domain-specific sonification strategies do not exist for the wide variety of streamed content. A long history of radio sports broadcasts, in which experienced announcers verbally describe a game, demonstrate that describing real-time descriptions can be understandable and engaging. Livestreams of video game tournaments often feature announcers who verbally describe in-game action. Building on prior success of describing live events, we investigate the potential for audio description novices who are experts in their domain of interest to produce live descriptions.

Synchronous interactions: Livestreams have remote and synchronous interactions, as opposed to recorded videos that are remote and asynchronous [20]. Similar to prior work on watching TV with others (i.e. "social TV" [4]), community members are able to interact synchronously with each other to build interpersonal relationships [12]. Livestreamers may also interact with their audience by reading chat messages or automated on-screen notifications (*e.g.*, listing a new subscriber) and responding verbally or adapting their actions in response (*e.g.*, "Thanks for the suggestion, I will try to make the background a farm."). To encourage interactions, streamers often complement the main streamed content with webcam videos of themselves or their environment, as well as additional on-screen overlays such as subscriber, question, or chat notifications burned into the video feed using OBS Studio [37], Streamlabs [40], or StreamYard [14]. For streamers with visual impairments, it can be challenging to set up such a streaming environment [21]. For viewers, it can be difficult to access these elements as they are not screen reader accessible or not directly described by the streamer.

Conversation on and off the streaming platform: Hamilton et al. described that the use of psuedonyms and text chat can promote self-disclosure that can help people build relationships [12]. People may also carry the same psuedonyms onto shared community spaces outside of streams (*e.g.*, on Discord[1]) to continue to talk to others. We focus our study on content on the streaming platform as the precursor to other types of interactions.

2.3 Crowdsourcing Accessibility

Professional audio describers create highly polished audio descriptions for movies that involve scripting, voiceover, and editing to create the finished product. Given a limited amount of expert describers and the high cost of this process, professional description is not practical for user-generated videos. YouDescribe [18] offers an approach for people to request descriptions and for volunteer describers to provide descriptions. Prior work has also explored crowdsourcing for answering visual questions [2], providing captions and transcriptions (*e.g.*, transcription services like Rev.com), and providing on-demand visual support [8]. However, professionals and crowd workers without domain expertise alike may have difficulty describing content that is unfamiliar to them. For example, Pavel et al.'s formative work with audio describers revealed that describing a new domain can require extensive research into the domain and terminology before providing accurate descriptions [34].

Instead of crowdsourcing, prior work in community sourcing [13] and learner sourcing [23] explore drawing from a pool of workers that might have expertise or vested interest in the relevant domain. This approach has had prior success in creating captions. For example, YouTube Community Captions provided community members the chance to add captions to recorded YouTube videos. Prior research also invited domain experts (student learners) who were not experts in providing captions to provide captions that accurately reflected the domain in real-time [24]. We explore

[1] https://discord.com

community-sourcing for providing descriptions for livestreams — a task that requires domain expertise to complete.

3 DESCRIBER STUDY

Prior work has explored current challenges and approaches to authoring descriptions for visual media including slide presentations [35, 36] and recorded videos [31, 34, 46]. Livestreams necessitate live description (i.e. written synchronously), rather than asynchronous description of recorded videos. Livestreams often also feature a wide variety of content that requires domain expertise to describe (*e.g.,* complex multiplayer gameplay), a breadth and depth of content that expert describers may not be familiar with. To explore the opportunities and challenges of live, community-driven descriptions, we invited 18 livestream viewers with domain expertise to describe livestreams in their domain of interest.

3.1 Methods

We conducted a remote within-subjects study with 18 participants describing videos in their domain of expertise across 7 categories. To determine the optimal method of recording live descriptions, we used three description approaches: two synchronous description input methods (one via text and one via speech) and one asynchronous description method (via text). Each participant participated in an individual, 1 hour long, remote study via Zoom (n=4) or Discord (n=14) voice call, and we compensated participants $20.

3.1.1 Participants. We recruited 18 sighted participants (P1-P18) from Discord servers and Reddit. All participants were between the ages of 19 and 30 (median=21). Participants ranged from watching 30 minutes to 30 hours of livestreams per week. The participants with the two highest watchtimes per week, 28 and 30 hours, were streamers themselves or frequently watched streams while they performed other tasks. Participants reported their genders as: 11 male, 5 female, and 2 N/A or Non-Conforming. All participants were self-expressed experts in the video category they described and had not previously authored audio or text descriptions for videos.

3.1.2 Videos. To explore a variety of content, we first selected 7 popular livestream categories from Twitch, a popular livestreaming platform for viewers with visual impairments [21] and participatory communities [12]. The videos selected spanned video games (League of Legends, Smash Bros, Valorant, The Legend of Zelda: Breath of the Wild (BOTW)), board games (Chess), and creative work (Digital Art, Makeup). As video games represent the most common type of livestream, we selected a variety of video games: a multiplayer online battle arena game (League of Legends), a first-person shooter game (Valorant), a third-person fighting game (Smash Bros), and a single-player adventure game (The Legend of Zelda: Breath of the Wild). For each video category, we selected three livestreams from three different streamers for a total of 21 videos to represent a variety of livestream styles (Table 4). We selected a 5 minute clip from each video for the study. For five of the livestream categories, we recruited three participants with expertise in the category (Chess, Digital Art, League of Legends, Super Smash Bros., Valorant); for one of the livestream categories we recruited two participants (Makeup); and for one of the livestream categories we recruited one participant (Breath of the Wild). We downloaded the videos from Twitch for analysis.

3.1.3 Description Approaches. During the study we asked participants to use three description approaches: synchronous text description, asynchronous text description, and synchronous audio description. Our description interfaces built on prior systems for creating audio descriptions that enabled description authors to script and edit descriptions using text [27, 31, 32, 50], and record spoken descriptions using audio [34]. Our live text interface (i.e. *synchronous text description*) let describers write text descriptions. It did not enable describers to record their text descriptions using audio or edit descriptions they had already written as such actions are not possible in real-time. To investigate the impact of providing extra time on describer preference and rate, we allowed 2x video time (10 minutes) and enabled text editing along with video navigation in our *asynchronous text description* condition. Finally, we accounted for slower typing speeds by adding *synchronous audio description* to let participants dictate rather than type their descriptions.

3.1.4 Describer Extension. We implemented our description approaches as a Google Chrome Extension that can be used alongside Twitch to enable real-time description authoring (Figure 1, right). The extension enables describers to watch the video while writing descriptions. Describers designate a new description by pressing 'Enter' to start a new line and optionally pressing the backslash key ('\') to insert a time code of the segment they are about to describe. Describers can then write their description. To review their descriptions, describers click on the time code to jump to the corresponding point in the livestream, then read back their text descriptions while rewatching the video.

3.1.5 Procedure. We first asked participants to answer a series of demographic and background questions about their experience watching livestreams and audio describing videos. To help participants craft useful descriptions, we shared existing audio description guidelines from YouDescribe [17] and the Audio Description Project [33], and showed participants example expert descriptions of the Disney's *The Incredibles (2004)* [44]. The guidelines gave participants instruction on what to describe (e.g. speakers, lighting, facial expressions, on-screen text) and how to describe it (e.g. use present tense, be objective, avoid technical terminology when possible). Participants then installed our Google Chrome extension and completed practice descriptions for one livestream using all three description methods. After the practice session, participants completed the study task, describing three 5-minute video clips within their area of expertise, each using a different description approach, provided in a random order. We counterbalanced conditions so that each video was described using each condition only once across all participants. Post-task we conducted a semi-structured interview to collect participant feedback on their strategies for describing the video, challenges they experienced in describing the video, and preferences among description approaches.

3.1.6 Analysis. We recorded the studies using Zoom Cloud Recording for Zoom interviews and OBS Studio [37] for Discord interviews, then automatically transcribed the videos using Descript [7]. We downloaded text descriptions from our server and segmented them into individual descriptions by new lines. For audio descriptions, we transcribed the description recordings and segmented them into individual descriptions by pauses in speech. We marked the beginning of each description as the time code that the description would appear. We analyzed the interviews using affinity diagramming to group quotes into higher level themes: description strategies (e.g., priorities, challenges, commentating), modality (text, audio), timing (sync, async), and future use (e.g., motivation, scenarios, alternate uses). We analyzed the descriptions by randomly selecting a subset of 300 descriptions from the whole set of 1183 total descriptions produced by participants, then performing open coding to derive 4 higher level themes and 24 subthemes (Table 2).

3.2 Results

Overall, participants wrote 1183 descriptions over 54 total video description sessions with an average of 21.9 descriptions per video (σ = 11.5 descriptions) and 210.3 words per video (σ = 143.7 words).

Livestream description strategies. Over a random subset of 300 descriptions, participants primarily described the main content of the stream (266 descriptions), and occasionally described additional visual content including cameras (34 descriptions) and game-specific actions performed by characters (70 descriptions). To describe the main content of livestreams, participants shared information about the high-level context of the stream (i.e. *state* descriptions) and low-level updates as the stream continued (i.e. *play-by-play* descriptions). State descriptions provided context for understanding play-by-play descriptions, and participants would add a new state description whenever a notable update to the entire stream state occurred. For example, P9 provided a state update for a new League of Legends game starting: *"Doublelift is in champ select. His team bans Yuumi, Poppy, Jax, Taliyah, and Pyke. The enemy team bans Master Yi, Katarina, Akali, Lulu, and Fiddlesticks. Doublelift is support and his ADC is hovering Zeri."* (V11). Participants provided more play-by-play descriptions (215 descriptions) than state descriptions (56 descriptions).

To fit play-by-play descriptions within limited time, participants often used domain-specific terminology to provide real-time updates (*e.g.*, *"Sage plants spike"* -P16). All participants used domain-specific lingo for at least one description. For example, P15 mentioned they used several shorthand terms that refer to controller inputs including *"dair"* for *"down air"* (a type of attack performed by holding down on the controller's left joystick and pressing the A button while the player's character is not grounded), and *"Up B"* (a type of attack performed by inputting a joystick angle and button combination on the player's controller). While such descriptions helped participants fit additional information about the game, participants expressed concern about the use of technical terminology. For example, P6 questioned if viewers would understand the word *"chibi"* they used to describe a Japanese art style where characters are drawn with exaggerated features. While participants had domain specific terminology for some in-game actions, participants also mentioned that they occasionally did not know how to describe actions they were seeing (*e.g.*, complex action sequences that used glitches or exploits (P1), character poses (P7), or streamer's facial expressions (P12)), or may not be able to understand complicated action sequences that they were seeing (P4). On the other hand, participants noted it was easiest to describe objects and actions that were not domain-specific. For example, human body parts in a drawing (P7), common actions like running, swimming and shooting a bow in a game (P1), reading on-screen text verbatim (P4), or describing simple visuals (*e.g.*, a single person on screen).

Participants identified that low level, play-by-play descriptions were not always the best strategy to describe fast-paced streams or to capture important visual information. Participants responded by changing the level of granularity. For example, the pace of the chess stream on puzzles (V6) was too fast to type or speak each piece movement, so P2 described the stream by mentioning the number of puzzles completed and the number of mistakes the streamer had made. When describing art content, P6 noted that they changed their description strategy from low-level stroke-by-stroke descriptions to higher-level descriptions of what was being drawn: *"Just saying it's being drawn isn't really that helpful. Towards the end, I was trying to say like, the wings are open as if imposing, so that they can sort of imagine it's this big, otherworldly-type figure."*. Trying to add context for low-level moves in a Valorant game, P18 added commentary that could describe streamer intentions for using certain abilities or aiming certain locations. P14 mentioned that that providing descriptions felt similar to esports commentating. While commentators may provide inspiration for the style and content of the descriptions, P15 higlighted that commentating and describing serve different purposes: *"Commentating is just supplementing what people can see on the screen."*

While participants all prioritized describing the main content, they included information about other visual streams as possible, when relevant, or in reaction to unidentified sounds. P5 described: *"I'd focus mainly on [...] what they were drawing, then second priority their face cam, and third priority anything else."*. 15 of 54 sessions started with descriptions of the stream's environment in addition to the main content, but most participants only described parts of the livestream other than the main content when relevant. For example, P12 mentioned that when describing a makeup video, they did not describe the background of the streamer until the streamer directly referenced background objects or walked off-screen. Other participants highlighted that they described on-screen overlays and chat only when mentioned by the streamer or when overlays prompted an unidentified noise. However, when reflecting on their performance, P5 noted that it may have been easier to follow their description if they had described the status of the stream as a whole before starting: *"I would've said, in the top left there's the face cam, below that is the dog face cam, and to the right side of the screen is just the drawing."* P5 and P11 noted that balancing the streams was difficult due to not knowing what to prioritize (P11) or needing to pay attention to multiple screens (P5).

Comparing livestream description methods. Overall, participants ranking the description methods from 1 (most preferred) to 3 (least preferred) ranked asynchronous text descriptions as the most preferred input method ($\mu = 1.5$, $\sigma = 0.62$) followed by synchronous audio ($\mu = 2$, $\sigma = 0.91$) and synchronous text ($\mu = 2.33$, $\sigma = 0.69$). A Friedman test[2] indicated a significant difference in preference between description methods ($\chi^2(2) = 6.12$, $p < 0.05$), with a post

[2] We used Friedman and Wilcoxon tests due to ordinal data (preferences), and non-normal distributions (description count and description words per video minute).

hoc Wilcoxon test with Bonforroni correction indicating a significant difference only between asynchronous and synchronous text descriptions ($p < 0.01$). Participants also produced more descriptions per video minute with synchronous audio ($\mu = 6.28$, $\sigma = 2.92$) and asynchronous text ($\mu = 4.22$, $\sigma = 2.36$) than they could with synchronous text ($\mu = 3.20$, $\sigma = 1.29$). Similarly, participants produced more description words per video minute with synchronous audio ($\mu = 60.97$, $\sigma = 35.52$) and asynchronous text ($\mu = 43.70$, $\sigma = 24.37$) than they could with synchronous text ($\mu = 26.90$, $\sigma = 10.18$). Friedman tests indicated significant differences in description counts ($\chi^2(2) = 18.77$, $p < 0.001$) and description words ($\chi^2(2) = 17.44$, $p < 0.001$) between description methods. Post hoc Wilcoxon tests with Bonforroni correction indicated significant differences ($p < 0.05$) between all pairs of methods for both description counts and description words per video minute.

Text vs. audio descriptions. 11 participants preferred synchronous audio to synchronous text. 6 participants expressed that speed was the key limitation for text-based methods, and P14 mentioned that their typing was error-prone. To keep up with synchronous text streams, 8 participants reported that they used hotkeys and shorthand. As P5 described, *"If you know their subscriber effects, you can write it once, and then you can just copy-paste it."*. 5 participants expressed that attempting to avoid talking at the same time as the streamer was the key challenge of dictating audio descriptions. As P12 described: *"The audio was just so difficult. [...] I felt like I was butting into a conversation."*. On the other hand, when P12 was using text without looking for gaps, *"I felt like I was much more descriptive and tackling more of the things that I'm supposed to be describing rather than just like, this is what's happening."*. Participants also expressed the challenge of unpredictability of the length of the gap between speech: *'There were moments where I would have a rather long thought about how I would describe [the stream], but I would have to stop because the streamer would start talking"* (P6). P13 noted that they would describe while the streamer focused on the game, but they didn't know when the streamer's focus would break and they start talking again (V17).

Synchronous vs. asynchronous text descriptions. 12 participants preferred asynchronous text over synchronous text, and 1 participant rated them equally. Participants preferred asynchronous text as it let them focus on important parts of the stream (P10), pause the video (P5, P6, P7), and not have to describe the video perfectly the first time (P6). 3 participants did not pause more than 3 times during their asynchronous text video, including P2, who preferred *synchronous* text as it felt "more accurate" to what they wanted to say. As participants had to budget their own time for asynchronous text, 1 participant ran out of time and only described 3.5 minutes of the 5 minute clip.

8 participants reported that synchronous text descriptions added time pressure to write something down in the moment before there was something else to describe. P16 reported that *"I was gonna type some stuff, but then 40 other things also happened and like we already moved on and I was like, no, I'm just not gonna talk about this anymore."*. As P15 described: *"I almost feel bad. I feel like there were details that would be nice to know that I just wasn't able to say."*

Future description. Participants reported that composing descriptions was challenging and that they would be willing to describe videos again in the future depending on how interested they are in the video. While most participants preferred to describe videos they would watch anyway, P18 reported that they would prefer to describe videos they are *not* as interested in so that they can focus on enjoying their streams of interest. 7 participants reported that they would provide descriptions if compensated (*e.g.* by the streamer), while 11 participants would volunteer to write descriptions. P3 compared writing descriptions to chat moderation, which is often a volunteer task. As describing is challenging, several participants mentioned that they would want to describe in smaller blocks of time, from around 15 minutes (P1) up to an hour at once (P5, P7, P16) for synchronous text. 5 participants suggested alternate use cases for using written descriptions as sighted people, including watching a stream in the background or on another monitor (3 participants), walking outside without their phone out (1 participant), driving (1 participant), or getting ready for an event (1 participant).

4 AUDIENCE STUDY

We conducted a study with livestream viewers with visual impairments to learn about current livestream viewing practices and challenges and surface description preferences.

4.1 Methods

We conducted a 1 hour remote study via Zoom with 9 participants with visual impairments who used screen readers to access their device. Participants were recruited through Reddit discussion boards [39] and email lists, and all participants had used Zoom in the past. We compensated participants $25 for their time.

4.1.1 Participants. Participants U1 through U9 ranged from ages 27 to 57 (6 male and 3 female) (Table 3). All participants reported that YouTube was their primary video streaming platform, with one participant watching Twitch an equal amount. Participants spent on average of 0.25 hours to 10 hours per week watching live video.

4.1.2 Procedure. We first asked participants demographic questions and background questions about their current livestream watching practices, platform and content accessibility challenges, and strategies for gaining more information. To demonstrate current practice, participants then searched for, selected, and watched 5 minutes of any one livestream on their preferred livestream viewing platform. We invited participants to ask questions about the visual content in the video and to rate their perceived accessibility of the video from 1 (very inaccessible) to 7 (very accessible), similar to Liu et al. [26]. To provide feedback on sample descriptions, participants selected one topic from the 7 livestream categories in Section 3.1.2 and watched 3 different five-minute clips on that topic. We paired each of these 3 streams with a description from a different description approach (synchronous text, synchronous audio, asynchronous text) produced during the Describer Study. All participants selecting the same category were served the same video-description approach pairs in a random order. Participants accessed descriptions via links to a webpage displaying a recording of the video and a description box (Figure 2). Before each video, the researcher provided an overview on the video context, including a brief description of the streamer and the main content of the clip. Once participants began watching each clip, descriptions were read back automatically by the participant's screen reader as the corresponding timestamp in the video overlapped with a stored description's time code. To control for audio quality and noise, all descriptions created with the synchronous audio description method were transcribed and played back as if they were written via text. After each stream, we invited participants to ask questions about the visual content in the scene, rate their perceived accessibility of the video with and without descriptions from 1 (very inaccessible) to 7 (very accessible), and provide feedback on what they liked, disliked, and wished to improve about the descriptions. Finally, we asked participants closing questions about their overall livestream description comparisons and preferences.

4.1.3 Analysis. We asked participants to screen share with sound using Zoom, recorded the studies using Zoom Cloud Recording, then automatically transcribed the videos using Microsoft Office Word 365 [30] and Adobe Premiere Pro CC [1]. We grouped participant responses according to our questions (e.g., current practice, strategies, challenges, and description preferences), then iteratively identified concepts using open coding.

4.2 Results

Overall, participants rated the accessibility of their preferred streaming platform as 4.7 ($\sigma = 1.2$) out of 7 and similarly rated the live content on these platforms as 4.2 ($\sigma = 1.5$) out of 7. Participants rated the videos they hand-selected during the co-watching study from streamers they were familiar with as 5.78 ($\sigma = 1.39$). For our example descriptions to probe for feedback, 5 participants chose to watch Breath of the Wild (BOTW), 2 participants chose digital art, 1 participant chose chess, and 1 participant chose Super Smash Bros. Participants rated the videos they watched as 2.3 ($\sigma = 1.2$) without descriptions and 5.2 ($\sigma = 1.4$) with descriptions.

4.2.1 Current livestream viewing practices. Participants primarily watched livestreams to gain information (U1, U2, U3, U4, U5, U6, U8, U9) or for entertainment (U1, U2, U3, U4, U6, U7, U8). Live videos for gaining information included live news (U6, U9), travel (U1, U3), online conferencing (U5), learning guitar (U3), cooking (U8), household repair (U8), learning game strategy (U9), and other personal interests and hobbies (U2, U3, U4, U5). Domains for live videos in entertainment included gaming (U1, U4, U6, U7, U9), music (U1, U2, U3, U8), podcasts (U2, U6), Q&A's (U1, U3), and general commentary (U3, U7). Participants reported that they watched livestreams in particular as they appreciated the ability to interact in real-time with the presenter and other viewers, including the ability to ask questions and receive information at the same time as recorded and alongside everyone else (U5, U6, U7, U8, U9). Participants reported that livestreams included "more honest reactions" (U1) from streamers compared to typical edited content (U4). Participants also appreciated learning more about other people's experiences (U1, U2), including culture (U1), travel (U1, U2), or catching up with their friends (U2).

To find livestreams to watch, only U3 and U4 reported using recommendation feeds to find live videos of interest, unlike prior work for recorded videos in which most people used their recommendations [26]. Instead, participants watched streams shared by their friends (U5, U7), users on other social media (U2, U7), or news sites (U9). Many participants also monitored notifications from channels they follow, tuning in when they go live (U1, U2, U4, U6, U9). Otherwise, participants would search for their hobbies or specific topics they're interested in and pick one of the top results (U2, U5, U6, U8). During the co-watching portion, 5 participants used YouTube search to look for specific topics or streamers they regularly watch; U4 selected a recommendation from a subscribed channel on the front page of YouTube; U5 used a recommendation from an email mailing list; U7 checked their Twitch following list, but no one was online, so they used Twitch search for specific streamers they are familiar with; and U8 used Google search and appended "YouTube" to their query.

4.2.2 Current accessibility of livestream content & platforms. Participants reported that livestreams were most accessible when they had good audio quality, clear voices, a lack of background music (U5, U6), extensive narration from the streamer (U3, U8), and lack of fast-paced action (U3, U5, U8). U6 and U9 both picked accessible audio game streams with no visuals and presented by a visually impaired streamer— these streams were completely accessible to them. Some reasons that livestreams were inaccessible were similar to prior work exploring the accessibility of recorded video [26], including: on-screen text burned into the video but not described, unclear visual references (*e.g.*, "this", "there"), unidentified sounds, and lack of description of the main visual content. However, livestreams posed additional challenges: First, unexplained sounds were frequent due to sound-producing overlays added to the video (U5, U6) (*e.g.*, a subscriber notification). Additionally, as livestreams are long and unedited compared to recorded videos, streamers often left long silences as they took a break from talking (U3, U4, U5, U6)— or they would break from talking about the game to talk about miscellaneous topics, such as telling a story or responding to chat, that could make it difficult to follow the main content (U2, U3, U7). While watching a stream, U2 commented: *"I'm not sure if he's showing anything or if he's just talking or I have no idea here."* Most participants noted that chat messages were particularly hard to read due to factors like the speed of the chat and custom emotes (i.e. emoji-like images specific to the stream), such that when streamers responded to chat without describing it (*e.g.*, *"Yeah, I agree with that, let's try it."*), participants were unable to understand the context for the response. Participants reported that they also wanted more information about the streamer, including facial expressions and body language (U3), as well as what they look like (U8). U7 mentioned that when streamers were playing games, they wanted more background information about the game status (*e.g.*, the place on the map, the damage updates) that were typically not included in streamer narrations: *"I can hear that they're taking damage, but without a low health indicator noise you don't know how low they are."* U9 noted that when it was a game that they were not familiar with, it was difficult to learn what was going on.

Finally, the livestream platforms themselves were not fully accessible. 4 participants who used YouTube to watch livestreams mentioned wanting to watch Twitch streams outside of the study, but found the platform difficult to use due to poor labeling of interface elements and difficulty navigating using a screen reader.

4.2.3 Strategies for gaining information about inaccessible streams. Participants mentioned strategies for handling inaccessible streams including: moving on to find another stream, asking the streamer or audience for additional information (on Discord or chat), prompting the streamer to change their narration style, and asking friends or family. When participants were not invested in a particular stream, participants indicated they would move on to find other streams (U2, U5, U6, U8): *"It's simple. I don't watch it. I mean, if it's gonna frustrate me, so some people might get mad and rant about it. I'm like, OK, I can't watch it"* (U2). Participants U2, U4, U5, U6, and U9 reported reaching out to the streamer directly in chat or via email to provide more complete narrations for their actions. U4 mentioned that *"I have been known to reach out to the video provider to the video upload and say 'hey, I'm a blind individual consuming your content. Tell me what you're doing. Tell me what you're seeing."*. U4 reported that most of the streamers they follow on Twitch are good about trying to cue their viewers into what they're doing, and that U4 would remind them when the streamer forgets. Participants also suggested looking in chat or asking other viewers questions via chat (U1, U3, U5, U6, U7), though the chat itself was difficult to navigate. U7 described gaining additional information via live chat: *"One time I was on LilyPichu's stream and I asked why there's a tomato emoji after the name of the stream that day. And I was like, I'm blind and I'm just curious. And a few people said it's because she dyed her hair red and oh, okay. But it was hard to find that in the massive stream of faces with tears of joy."* U7 also mentioned they look at the chat when joining a stream as people often comment on what is going on in the stream, but neither U7 (nor U6) send a message themselves unless they are particularly curious about something, as they see it as bothersome. Participants also asked sighted friends and family members to answer visual questions (U3, U6). Finally, participants used external online sources such as web search (U5, U7, U9), Twitter (U5), and wikis (U7) to learn more about the context for the stream.

4.2.4 Livestream description preferences.

Participants all reported that they wanted descriptions to prioritize describing the main content. Additional descriptions of visual content should be described as relevant, including: reading out pop up overlays with text (U1), the appearance of the streamer (U2), and streamer reactions from facial expressions (U5). This preference order reflected community members' stated priorities in creating their descriptions.

Description content. From community member-provided descriptions, all participants reported that the descriptions provided useful information for understanding the stream. As a result, most participants reported that the accessibility of the video improved after descriptions. U1 summarized: *"Most of the time you're watching the video when you're [...] totally blind and you have no idea what's going on; having any form of description is helpful."* U2 highlighted that such descriptions were particularly useful for inaccessible moments: *"For streams where the streamer is talking aimlessly or about something independent of the main activity they're performing, it makes a stream someone would otherwise click off to something interesting."*. Participants mentioned that they liked when the describer included information about the player's strategy instead of only raw visual content— contrary to audio description guidelines. Participants also found the contextual information (*e.g.*, global descriptions of the video context before the start of the video) to be helpful in understanding the content of the video. Participants offered additional information they may want to know about for state descriptions including how fast the chat is moving (U8) and the goals of the gameplay (U5, U7). At times, participants disagreed with a describer's choice to include additional information (e.g., describing a subscriber notification in the middle of a drawing stream). Other participants reported that they wanted additional information about background about the game (U5) and additional detail about the minutiae of the game (U3, U8, U6).

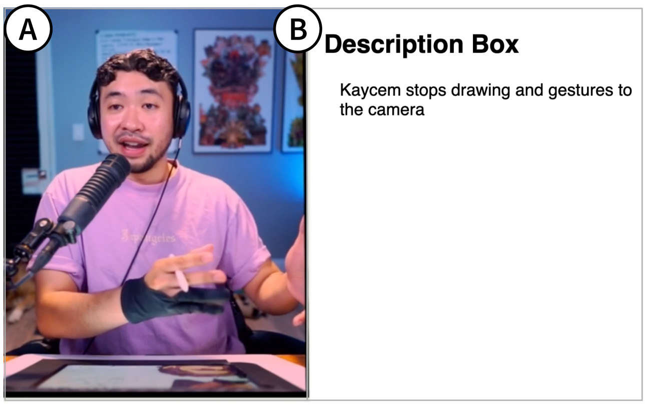
Expertise of describers and terminology. Most participants reported that they wanted descriptions from people with familiarity with the visual content as *"otherwise you can't learn from the stream"* (U2). U1 reported that they had heard descriptions with errors from people who lack experience in the past, and U7 mentioned that they wanted describers to be fans of the media: *"The more that the person's passionate about it, the better they do tend to be."* Participants also clarified that while many domains require expertise (*e.g.*, chess, art), expertise may not be necessary for some types of content (*e.g.*, a travel video).

U5 said that when audience members are coming to the video *"with zero background knowledge"*, the descriptions should be as *"beginner mode"* as possible; however, the describers in our study frequently used domain-specific terminology. While most participants found the terminology to be understandable from prior knowledge or context cues, other participants reported that they would like the ability to gain more information about terms used (*e.g.*, via a wiki link -U5). To make the video understandable to people with different skill levels, U9 suggested that descriptions contain no more than around 30% terminology so that there are enough context cues to understand the terms used.

Description format. Participants appreciated the text format of the descriptions as they could use their existing screen reader settings and reread descriptions when they were confused. However, participants said that they would like the ability to easily refer to past descriptions and a mode where they can pause the video to read descriptions in isolation. At times, text to speech pronounced words incorrectly, especially if the word was domain-specific or contained a spelling error brought on by the describer. U1 clarified that it would take them extra time to understand mispronunciations, but that they ultimately understood the content.

As we presented both forms of description provided by community members (text and speech) as screen reader-accessible text for consistency (*e.g.*, screen reader speed, audio quality, and background noise), the most critical description feedback was for the voice-generated descriptions. In particular, participants found these descriptions to contain awkward pauses, long gaps in description, incomplete sentences, and filler words (um, uh) that made them more difficult to understand.

Description timing. Participants all wanted the ability to be able to keep up with the stream in real time. Participants reported that they would tolerate delays from "within 15 seconds" (U4) to up to "1-2 minutes" (U7). U5 mentioned that they wanted the option to manually adjust the video delay to catch up with descriptions. U3 summarized that synchronized descriptions had the effect of: *"I'm right here. I'm participating in the moment. This is what's happening. [...] it's really cool to feel like I'm an equal participant in that moment" (U3).* U1 mentioned *"I think I'm just looking forward to something like this becoming mainstream. With more accessibility and primarily*



Kaycem stops drawing and gestures to the camera

Figure 2: Description playback application featuring a section of the source video (A) and the description box (B). The description box updates as the playback time of the video matches a timecode for a description. If the text within the description box is updated, a sound effect plays and the new description is automatically read out loud by the participant's screen reader. Source video: *how to IMPROVE your SKILLS QUICKLY + NEW SUB GOAL?!? !bootcamp !youtube* by Kaycem [22]

more description for videos either in a text format or human narrated description.".

5 DISCUSSION

Livestream viewers with visual impairments reported that they strategically used social support from the livestream community and their network to make livestreams accessible, but that they valued additional information provided by community member descriptions. Our studies point to additional ways to make livestreams more accessible to viewers with visual impairments.

5.1 Improving Live Descriptions In Practice

Livestream viewers found community member descriptions useful for understanding the visual content in a stream. Our studies indicate additional ways to help community members craft high-quality descriptions in practice:

5.1.1 Synchronous Listener-Describer Communication. Our study imitated live description and viewing experiences (*e.g.*, inability to skip ahead, timing), and in future work we will explore fully synchronous description. Audience members' current synchronous use of chat and Discord to gather information and provide feedback suggests that future two-way listener-describer communication may be beneficial to improve descriptions. For example, listeners could discuss their expertise and preferences with describers apriori, provide live description feedback, or ask live visual questions.

5.1.2 Maintaining Description Consistency. Community members reported they could describe 15 minutes to 1 hour at a time, such that multiple describers will be required to produce descriptions for streams that are many hours long. On the other hand, our audience study indicated that *description consistency* is important, as new terminology takes time to learn. To maintain description consistency between describers, future work will explore creating per-domain hotkey libraries of common characters and actions (similar to copy-paste strategies used by describers in our study) or providing describers a warm-up period where they observe the prior describers' descriptions.

5.1.3 Sticky State Descriptions. Community members used *state* descriptions to set the scene of the stream and subsequent *play-by-play* descriptions to deliver updates. However, *play-by-play* descriptions were often not understandable without corresponding state descriptions, such that viewers joining mid-stream may be confused. Future automated or human-authored descriptions should include "sticky" state descriptions that persist until new state descriptions are authored. When livestream viewers join a stream, they could first listen to the sticky state description before listening to play-by-play updates. Providing state descriptions alone, rather than full live descriptions, may also provide an easier task for first-time describers.

5.1.4 Tailoring Descriptions. Livestream viewers reported a variety of preferences in the level of detail they wanted from descriptions and the level of expertise they wanted the describer to have. In the future, we will explore automatically classifying descriptions (*e.g.*, based on our codes) and either provide viewers the ability to selectively toggle certain types of descriptions (*e.g.*, streamer appearance, state vs. play-by-play descriptions), or learn from viewer preferences indicated via lightweight feedback (*e.g.*, skipping the description, or thumbs up/thumbs down).

5.1.5 Transcription Accuracy. While several describers preferred audio input, livestream viewers noted that descriptions transcribed from audio were less clear than descriptions written via text. To improve the delivery of descriptions produced via live audio description, future work will explore automatically removing filler words and repetitions from audio input using a large language model. Future work will also explore improving text transcription errors for uncommon words by using a library of common domain actions and characters to inform transcription.

5.2 Beyond Community Descriptions

While livestream viewers found descriptions useful, they suggested additional approaches to make livestreams more accessible, including filtering chat messages and prompting streamers to describe visual content ahead of time. To highlight useful chat messages, future work may surface visual information in the livestream chat to augment descriptions [16] or identify important chat messages [47]. To help streamers remember to describe their stream, future work may consider automated approaches to prompt streamers to describe missing visual information [27, 36]. In the future, we will also explore how to augment community descriptions with multiple modalities (*e.g.* sonification [19], haptics [49]), and help livestream viewers surface accessible livestreams [26]. For example, we will explore modifying Liu et al.'s [26] approach for surfacing accessible YouTube videos to provide the accessibility score of a stream in progress, or for the streamer's prior streams, before a viewer decides to join.

5.3 Study Limitations

While our descriptions are usable by all screen reader users, livestream viewers in our study were all blind (n=5) or blind with light perception (n=4). Future work should investigate livestream viewing practices and challenges of low vision viewers, and explore approaches to provide support for low vision viewers. For example, community members could link their descriptions to a spatial region such that low vision viewers may play back descriptions relevant to their current zoomed-in view. Our studies featured video categories related to the expertise of sighted describers. While such descriptions produced useful description feedback in the audience study, our future work will explore studying synchronous descriptions with matched interests. Our studies explored the feasibility of community-driven live descriptions in the context of short, 1 hour study sessions. Our work reveals an opportunity for longer-term studies to study the impact of fatigue and expertise gain while using and providing descriptions over longer periods of time.

6 CONCLUSION

We conducted two studies exploring community-driven descriptions to improve the accessibility of livestreams. Our work reveals that livestream community members with visual impairments already use social support from the livestream community to gain additional information. However, community-driven descriptions would provide a valuable, additional channel for livestream viewers to gain access to livestream information, as well as the community building afforded by live participation. As our work is the first to investigate describing livestreams in real-time, we identify areas for future work in improving community-driven descriptions and general livestream accessibility. As we communicate via new mediums such as livestreams, we must assure that they are accessible to everyone. We hope that our work will catalyze future research and systems to improve livestream accessibility.

REFERENCES

[1] Adobe. 2022 (accessed Dec 13, 2022). Premiere Pro. https://www.adobe.com/products/premiere.html

[2] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. 2010. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. 333–342.

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

[4] Pablo Cesar and David Geerts. 2011. Past, present, and future of social TV: A categorization. In *2011 IEEE consumer communications and networking conference (CCNC)*. IEEE, 347–351.

[5] Xinyue Chen, Si Chen, Xu Wang, and Yun Huang. 2021. "I was afraid, but now I enjoy being a streamer!" Understanding the Challenges and Prospects of Using Live Streaming for Online Education. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW3 (2021), 1–32.

[6] Aira Tech Corp. 2023 (accessed May 2023). Aira. https://aira.io.

[7] Descript. 2022 (accessed Sep 6, 2022). Descript. https://www.descript.com/

[8] Be My Eyes. 2023 (accessed May 2023). Be My Eyes. https://www.bemyeyes.com.

[9] Travis Faas, Lynn Dombrowski, Alyson Young, and Andrew D Miller. 2018. Watch me code: Programming mentorship communities on twitch. tv. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.

[10] C Ailie Fraser, Joy O Kim, Alison Thornsberry, Scott Klemmer, and Mira Dontcheva. 2019. Sharing the studio: How creative livestreaming can inspire, educate, and engage. In *Proceedings of the 2019 on Creativity and Cognition*. 144–155.

[11] Cole Gleason, Amy Pavel, Himalini Gururaj, Kris Kitani, and Jeffrey Bigham. 2020. Making GIFs Accessible. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–10.

[12] William A Hamilton, Oliver Garretson, and Andruid Kerne. 2014. Streaming on twitch: fostering participatory communities of play within live mixed media. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1315–1324.

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

[14] Hopin. 2023 (accessed May 2023). StreamYard. https://streamyard.com.

[15] Yun Huang, Yifeng Huang, Na Xue, and Jeffrey P Bigham. 2017. Leveraging complementary contributions of different workers for efficient crowdsourcing of video captions. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 4617–4626.

[16] Mina Huh, YunJung Lee, Dasom Choi, Haesoo Kim, Uran Oh, and Juho Kim. 2022. Cocomix: Utilizing Comments to Improve Non-Visual Webtoon Accessibility. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

[17] The Smith-Kettlewell Eye Research Institute. 2019. YouDescribe FAQ for describers. https://youdescribe.org/support/describers.

[18] The Smith-Kettlewell Eye Research Institute. 2019. YouDescribe.com. https://youdescribe.org/.

[19] Gaurav Jain, Basel Hindi, Connor Courtien, Xin Yi Therese Xu, Conrad Wyrick, Michael Malcolm, and Brian A. Smith. 2023. Towards Accessible Sports Broadcasts for Blind and Low-Vision Viewers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–7.

[20] Robert Johansen. 1988. *Groupware: Computer support for business teams.* The Free Press.

[21] Joonyoung Jun, Woosuk Seo, Jihyeon Park, Subin Park, and Hyunggu Jung. 2021. Exploring the experiences of streamers with visual impairments. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–23.

[22] Kaycem. 2023 (accessed July 2023). how to IMPROVE your SKILLS QUICKLY + NEW SUB GOAL?!? !bootcamp !youtube. https://www.twitch.tv/videos/1854614493.

[23] Juho Kim et al. 2015. *Learnersourcing: improving learning with collective learner activity.* Ph. D. Dissertation. Massachusetts Institute of Technology.

[24] Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. 23–34.

[25] Hye-Kyung Lee. 2011. Participatory media fandom: A case study of anime fansubbing. *Media, culture & society* 33, 8 (2011), 1131–1147.

[26] Xingyu Liu, Patrick Carrington, Xiang 'Anthony' Chen, and Amy Pavel. 2021. What Makes Videos Accessible to Blind and Visually Impaired People?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–4.

[27] Xingyu Liu, Ruolin Wang, Dingzeyu Li, Xiang'Anthony' Chen, and Amy Pavel. UIST 2022. CrossA11y: Identifying Video Accessibility Issues via Cross-modal Grounding.

[28] Zhicong Lu, Michelle Annett, and Daniel Wigdor. 2019. Vicariously experiencing it all without going outside: A study of outdoor livestreaming in China. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–28.

[29] Meta. 2023 (accessed April 2023). Facebook Live. https://www.facebook.com.

[30] Microsoft. 2023. Word for the web. https://www.microsoft365.com/launch/word

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

[32] Rosiana Natalie, Joshua Tseng, Hernisa Kacorri, and Kotaro Hara. 2023. Supporting Novices Author Audio Descriptions via Automatic Feedback. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[33] American Council of the Blind. 2003. The Audio Description Project. https://adp.acb.org/guidelines.html.

[34] Amy Pavel, Gabriel Reyes, and Jeffrey P. Bigham. 2020. Rescribe: Authoring and Automatically Editing Audio Descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 747–759. https://doi.org/10.1145/3379337.3415864

[35] 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.

[36] 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.

[37] OBS Project. 2023 (accessed May 2023). OBS: Open Broadcaster Software. https://obsproject.com/.

[38] The Audio Description Project. 2019. adp.acb.org. https://adp.acb.org/guidelines.html.

[39] Reddit. 2023 (accessed April 2023). r/BlindSurveys. https://reddit.com/r/blindsurveys

[40] Logitech Services S.A. 2023 (accessed May 2023). Streamlabs. https://streamlabs.com.

[41] Jeff T Sheng and Sanjay R Kairam. 2020. From virtual strangers to irl friends: relationship development in livestreaming communities on twitch. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–34.

[42] Thomas Smith, Marianna Obrist, and Peter Wright. 2013. Live-streaming changes the (video) game. In *Proceedings of the 11th european conference on Interactive TV and video*. 131–138.

[43] Joel Snyder. 2005. Audio description: The visual made verbal. In *International Congress Series*, Vol. 1282. Elsevier, 935–939.

[44] Pixar Animation Studios. 2004 (accessed August 2022). The Incredibles: Am I Fired Scene with Audio Description. https://www.youtube.com/watch?t=128&v=2zhzVGmyjtg.

[45] Twitch. 2023 (accessed April 2023). Twitch. https://www.twitch.tv.

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

[47] Saelyne Yang, Jisu Yim, Juho Kim, and Hijung Valentina Shin. 2022. CatchLive: Real-Time Summarization of Live Streams with Stream Content and Interaction Data. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 500, 20 pages. https://doi.org/10.1145/3491102.3517461

[48] YouTube. 2023 (accessed April 2023). YouTube Live. https://www.youtube.com/@live

[49] Bei Yuan and Eelke Folmer. 2008. Blind hero: enabling guitar hero for the visually impaired. In *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*. 169–176.

[50] Beste F Yuksel, Pooyan Fazli, Umang Mathur, Vaishali Bisht, Soo Jung Kim, Joshua Junhee Lee, Seung Jung Jin, Yue-Ting Siu, Joshua A Miele, and Ilmi Yoon. 2020. Human-in-the-loop machine learning to increase video accessibility for visually impaired and blind users. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 47–60.

A DESCRIBER PARTICIPANT DEMOGRAPHICS

PID	Category	Age	Gender	Hours/Week	Expertise	Audio	Text
P1	Breath of the Wild	30	Male	15	8	1	5
P2	Chess	19	Male	8	9	2	2
P3	Chess	21	Male	3-4	7	3	3
P4	Chess	21	Male	1	3	1	1
P5	Digital Art	21	Female	1-2	9	1	2
P6	Digital Art	23	Male	2	3	1	1
P7	Digital Art	23	Female	0.5	8	1	3
P8	League of Legends	22	Male	0-0.5	4	1	1
P9	League of Legends	19	Male	<1	9	1	1
P10	League of Legends	21	Female	~2	7	1	2
P11	Makeup	21	Non-conforming	30	9	3	3
P12	Makeup	21	Female	0-2	2	1	1
P13	Smash Bros	21	Male	6	10	1	1
P14	Smash Bros	21	Male	10-12	9	1	1
P15	Smash Bros	22	Male	1-2	9	3	6
P16	Valorant	21	Female	6	8	1	1
P17	Valorant	21	Male	5	6	2	8
P18	Valorant	21	N/A	28	9	1	1

Table 1: Demographics of describer study participants. "Hours/Week" refers to the number of hours per week the participant typically watches live video. "Expertise" refers to a self-reported, 1-10 scale of how familiar each participant felt with their chosen category. "Audio" and "Text" refer to a self-reported, 1-10 scale of how familiar each participant felt with writing audio or text descriptions before the study.

B DESCRIPTION CODEBOOK

Code	Description	Quantity	Percentage	Example
Main:Play-by-Play	"Minor" updates as they happen	215	71.67%	"She brushes setting powder under her eye left to right"
Main:State	"Major" state updates (e.g. score updates, round changes, tempo shifts)	56	18.67%	"The new round starts"
Action:Character	Character-specific action names	44	14.67%	"Mythra Photon Edge"
Action:Verbiage	Game-specific verbiage	32	10.67%	"Sage plants spike"
Action:Controller	Character-specific actions referenced using its controller input	5	1.67%	"Pyra whiffs an up-B"
Camera:Streamer	Focused on the streamer themselves	30	10%	"Frederic takes off his headphones"
Camera:Background	Focused behind the streamer or in their environment	3	1%	"There are plant shelves in the background."
Camera:Misc	From any additional cameras (e.g. dedicated dog camera)	1	0.33%	"His dog looks up at him"
Audio:Correction	Correcting a previously recorded description (Audio only)	11	3.67%	"No, not the Nexus. The Nexus towers."
Audio:Unintelligible	Some part of the description is not understandable (Audio only)	2	0.67%	"Use another A ha[?] To catch uh Mew2King."
Characters	Character names	105	35%	"Before picking Bard, we see his team has locked in Zeri ADC, Qiyana Jungle, Renekton top, Cassiopeia Mid."
Locations	Use of words referring to locations specific to the media	77	25.67%	"Timmy holds snowman and yellow from default on B site"
Lingo	Use of words specific to the media or category of media	32	10.67%	"Mew2King gets him with a landing neutral air forward tilt kill confirm."
Object	Use of objects, items, and/or utility	23	7.67%	"A prowler goes out as he nades mid"
Tools	Use or switching between of specific items to perform main content	18	6%	"She uses a brush to blend in a setting powder"
Uninformative	Not able to be understood without additional context	15	5%	"k"
Text	On-screen text read audibly	11	3.67%	"His time ends with 27 correct and 3 incorrect"
Menu	Menus and inventory navigation; gamemode selection	9	3%	"She goes into the inventory and switches the suit to the climbing gear."
Commentary	Side commentary from describer irrelevant to the stream content	8	2.67%	"ez takes out yasuo wannabe"
Context	Adds context about visual information to existing audio in the stream	5	1.67%	"White plays a move to support a pawn push which Alexandra doesn't think is necessary"
Overlays	Content on streamer's overlay	4	1.33%	"Subscriber emotes still bouncing across the screen"
Redundant	Description that could have been discerned from audio alone	3	1%	"He wants to do some vision training" (After the streamer says they want to do some vision training)
Chat	Describing messages from chat; giving context on streamer's behalf	1	0.33%	"There are a good amount of Fs in the chat."
Miscellaneous	Anything not marked in previous codes (with rationale)	3	1%	"Fox pointed ears, long hair" (ambiguous)

Table 2: List of codes applied to descriptions. Quantity and percentage metrics are out of 300 descriptions coded. Percentages add up to more than 100% as multiple codes may be assigned for each description. In these descriptions, we have identified 4 higher-level themes: "Main", referring to the main content on stream, whether that be a game, the streamer's webcam, the streamer's drawing tablet; "Action", referring to different types of actions performed by characters in-game; "Camera", referring to descriptions of content based around one or more cameras in the scene; and "Audio", issues with descriptions as a result of the audio input method. Codes have been sorted in order of most prevalent to least prevalent, with higher-level themes grouped at the top, followed by individual lower-level themes, and miscellaneous at the bottom.

C AUDIENCE PARTICIPANT DEMOGRAPHICS

PID	Age	Gender	Screenreader	Level of Vision	#1 Platform	Hours/Week	Most-watched categories	Co-watched category	Chosen category
U1	35	M	VoiceOver	Blind with light perception	YouTube	0.5-1	Travel, Lifestyle	Travel	BOTW
U2	30	F	VoiceOver	Totally Blind	YouTube	5-10	Religious, Educational	Lecture	Art
U3	29	F	**NVDA**, JAWS	Blind with light perception	YouTube	0.25	Music, Commentary	Travel	Chess
U4	36	M	NVDA	Blind with light perception	**YouTube**, Twitch	6 - 8	Gaming	Aviation	BOTW
U5	57	M	NVDA	Totally Blind	YouTube	1.5	Commentary	Lecture	Smash
U6	43	F	VoiceOver	Totally Blind	YouTube	1	Gaming, Commentary	Gaming	BOTW
U7	27	M	**NVDA**, JAWS	Blind	YouTube	1	Gaming	Gaming	BOTW
U8	56	M	JAWS	Blind with light perception	YouTube	5 - 6	Music, Cooking, DIY	Cooking	Art
U9	45	M	NVDA	Totally Blind	YouTube	0.75	News, Gaming	Gaming	BOTW

Table 3: Demographics of user study participants. U3 and U7 cited familiarity with both NVDA and JAWS but opted to use NVDA for the study. U4 used YouTube and Twitch equally but opted to use YouTube for the co-watching portion of the study.

D VIDEO REFERENCE

Video ID	Category	Streamer	Intent	On-screen chat	Cameras	Overlays
V1	Breath of the Wild	PointCrow	Modified game	x	1	x
V2	Breath of the Wild	LimCube	Speedrunning	x	1	x
V3	Breath of the Wild	DapperDame	Let's Play	x	1	x
V4	Chess	BotezLive	Blitz	x	1	x
V5	Chess	PrincepsComitatus	Classical		0	
V6	Chess	Keithonsky	Puzzles	x	1	x
V7	Digital Art	39daph	Stylized emojis	x	1	x
V8	Digital Art	kaycem	Educational	x	2	x
V9	Digital Art	sezza	Magical realism	x	1	x
V10	League of Legends	loltyler1	Mid-to-late game	x	1	
V11	League of Legends	Doublelift	Champion Select	x	1	x
V12	League of Legends	LHS Esports	Tournament Camera		0	
V13	Makeup	ThatGaymingAsian	Creating a "look"	x	1	x
V14	Makeup	CodeMiko	Preparing for an event	x	2	
V15	Makeup	MeeshAmerica	Preparing for work		1	
V16	Smash Bros	Hungrybox	Elite Smash	x	1	x
V17	Smash Bros	Mew2King	Elite Smash	x	1	x
V18	Smash Bros	Plup	Elite Smash	x	1	x
V19	Valorant	Masayoshi	Ranked	x	1	x
V20	Valorant	iiTzTimmy	Deathmatch	x	2	x
V21	Valorant	itzbelac	Tournament Camera		0	

Table 4: Information about each video described. "Intent" refers to how a particular video is differentiated from other videos of the same category. "On-screen chat" contains an X if chat appeared on screen during the stream. "Cameras" refers to a number of webcams on-screen. "Overlays" contains an X if donations, subscription alerts, or on-screen text appears during the stream.