frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 19 June 2025
DOI 10.3389/fcomp.2025.1575176

OPEN ACCESS

EDITED BY
Roberto Therón,
University of Salamanca, Spain

REVIEWED BY
Kostas Karpouzis,
Panteion University, Greece
Paula Igareda,
Pompeu Fabra University, Spain

*CORRESPONDENCE
Lloyd May
✉ lloyd@ccrma.stanford.edu

RECEIVED 11 February 2025
ACCEPTED 25 March 2025
PUBLISHED 19 June 2025

CITATION
May L, Clemens M, Dang K, Ohshiro K, Sridhar S, Wee P, Fuentes M, Lee S and Cartwright M (2025) "Choices? That's the dream": challenges and opportunities in non-speech information closed-captioning. *Front. Comput. Sci.* 7:1575176. doi: 10.3389/fcomp.2025.1575176

"Choices? That's the dream": challenges and opportunities in non-speech information closed-captioning

Lloyd May[1]*, Michael Clemens[2], Khang Dang[2], Keita Ohshiro[2], Sripathi Sridhar[2], Pauline Wee[3], Magdalena Fuentes[3], Sooyeon Lee[2] and Mark Cartwright[2]

[1]Center for Computer Research in Music and Acoustics, Stanford University, Stanford, CA, United States, [2]Department of Informatics, New Jersey Institute of Technology, Newark, NJ, United States, [3]Music and Audio Research Laboratory, New York University, New York, NY, United States

Introduction: Access to non-speech information (NSI) in video content is essential to creating accessible and engaging video content, particularly for D/deaf and Hard-of-Hearing (DHH) audiences. In this paper we present an overview of the current state of NSI captioning research, professional practice, and user preferences.

Methods: We utilized a comprehensive review approach that combined a systematic literature review methodology with a mixed-methods survey and interview study. 1276 papers were screened with 36 eligible for the final inductive best fit analysis. 168 DHH participants completed an online survey and 15 participated in semi-structured interviews. Additionally, 5 professional captioners participated in semi-structured interviews.

Results and discussion: We offer systematic insights into the current challenges related to NSI captioning faced by DHH users and professional captioners, trends in recent NSI captioning research, as well as opportunities for future work that enhance user agency, utilize integrated research methodologies, and broaden community involvement.

KEYWORDS

closed-captioning, subtitles, non-speech information, sound effects, music, deaf and hard-of-hearing, accessibility

1 Introduction

Closed captions are textual versions of sound. They are critical for the accessibility of audio-visual content, especially for the large worldwide d/Deaf or hard-of-hearing (DHH) population (Henry, 2022). While captions primarily contain transcriptions of speech, they may also contain non-speech information (NSI), i.e., all of the information in sound besides the spoken words. NSI includes information about non-speech sounds such as environmental sounds, sound effects, incidental sounds, and music. The broad category of NSI also includes additional narrative information and extra-speech information (ESI), which gives context to spoken or signed language such as manner of speech (e.g., "[Laughing] I love it!") or speaker label (e.g., "[Serafina] I love it!"). NSI captions can deepen emotional connection, provide additional context, and communicate information critical for understanding video content (Zdenek, 2015).

However, despite its importance and NSI captioning guidelines (Caldwell et al., 2008) that recommend captioning NSI, NSI historically has often been overlooked in favor of speech (Downey, 2008). In fact, one recent study analyzed the NSI captioning trends on YouTube and found that only 4% of the videos in their sample had NSI captions (May et al., 2024). The authors also found the prevalence of NSI captions is actually decreasing—a

trend they speculate is due to the increasing use of automatic speech recognition (ASR) to aid in 'manual' captioning pipelines (3Play Media, 2023, 2017). This finding implies there is a need for better NSI captioning tools, including ones that leverage the use of machine listening, similar to how speech captioning tools leverage ASR.

Unfortunately, we know surprisingly little about the NSI captioning needs and challenges of either DHH viewers or captioners—critical information for designing new NSI captioning tools. While NSI research is a rapidly growing area of research (see Section 4), studies that focus specifically on NSI are typically small-scale studies aimed at evaluating novel NSI display technologies (e.g., Wang et al., 2016; Vy and Fels, 2009; Mori and Fels, 2009) whereas large-scale studies on captioning have focused on speech and general captioning issues rather than NSI (Hersh, 2013; Jensema, 1998; Austin and Myers, 1984; Fitzgerald and Jensema, 1981). Furthermore, while there has been considerable progress on the machine listening task of automated audio captioning (AAC) (Xu et al., 2024; Mei et al., 2022), i.e., automatically describing non-speech sounds with natural language, such machine-learning models have not been designed based on the needs of DHH audiences nor has their application to sound accessibility been studied. Thus, it is unclear if and how such technologies should be utilized for novel NSI captioning tools.

In this work, we address this knowledge gap by developing a comprehensive understanding of the challenges, needs, and opportunities related to NSI from the perspectives of both audiences and captioners. Specifically, we aim to answer the following research questions:

- **RQ1**. Based on current literature, what are the key areas for improvement and future research in NSI captioning?
- **RQ2**. What specific aspects of NSI captioning are most important to viewers and captioners?
- **RQ3**. What contextual factors impact the preferences and concerns of DHH viewers and captioners regarding NSI captioning?

To accomplish this, we conducted a systematic literature review of NSI captioning research, a survey and interview study with DHH viewers, and an interview study with professional captioners. We aim for this work to inform the design of NSI caption authoring and display tools, including those assisted by AI. Advances in both authoring and display techniques are crucial to addressing the needs of DHH audiences and ensuring that video content is as accessible and enjoyable as possible.

2 Background

2.1 Closed-captioning

Closed captioning (i.e., captions that can be voluntarily turned on/off) debuted on network television in March 1980 (Downey, 2008). Over the next twenty years, US legislation was passed to require televisions to include caption decoders (United States Congress, 1990) and broadcasters to caption most of their content (United States Congress, 1996). In 2012, additional regulations in the USA were passed that require most previously broadcast content to be captioned when redistributed over the internet on official channels (United States Congress, 2012). Yet, the vast majority of video content on the internet (e.g., YouTube) does not meet this criteria and thus is not mandated to be captioned by law. Without such regulation, many videos remain uncaptioned due to the considerable time required to manually caption content (May et al., 2024).

Automatic speech recognition (ASR), which algorithmically transcribes speech to text, is commonly used to efficiently caption speech and was first used by YouTube for automated speech captioning at scale in 2009 (Harrenstien, 2009). However, while ASR systems have undoubtedly had positive impacts, the lack of human vetting and editing means generated captions often contain inaccuracies (Berke et al., 2017), so much so that autocaptions have been nicknamed "autocraptions" by members of the DHH community (Evans, 2019). Thus, researchers have developed other methods to efficiently create speech captions, including semi-automated methods, in which ASR is followed by human corrections (Wald, 2006), and crowdsourced speech captioning methods (Lasecki et al., 2013; Harrington and Vanderheiden, 2013; Naim et al., 2013). This use of ASR is now common practice—in a 2023 survey on captioning with over 300 respondents across a variety of industries, over 40 percent of respondents reported using post-edited ASR outputs for captions (3Play Media, 2023).

2.2 Extending closed-captioning methods

In addition to traditional closed-captions, research has explored a variety of ways to communicate sound information to users in novel ways using additional sound communication technologies (SCTs). SCTs range from on-screen solutions, such as using icons or emojis to communicate information about the sound (Alonzo et al., 2022), to visualizing properties of the sound such as stereo location and frequency information (May et al., 2023; Square Enix, 2010; McGowan et al., 2017). Other novel SCTs have attempted to reduce visual clutter by utilizing off-screen technologies, such as vibrotactile haptics (Kushalnagar et al., 2014), or allowing users to customize captions, such as the ability to select caption colors for individual characters (Gorman et al., 2021). SCTs are generally designed to communicate aspects of NSI that are often difficult or cumbersome to communicate using traditional captioning, such as temporal and contextual information.

Several models have emerged to better understand the function of captioning and SCTs through a critical or rhetorical lens. Tsaousi et al. proposed that there are four main reasons for using sound effects in video content, namely: Exegetic (reinforces meaning), Narrative (coherence with the plot or story), Contextual (additional information about surrounding events), and Emotive/Aesthetic (convey or induce emotion) (Tsaousi, 2015). They then propose a Source, Function, Adequacy model of NSI captioning based on this framework, viewing NSI captioning as a way to answer a series of sequential questions: What is making the sound? Why is it making the sound? Have we captioned the sound adequately? May et al. proposed a parallel framework to understand NSI communication more broadly, namely the Selection, Curation, and Communication model (May et al., 2023). This also proposes a series of sequential questions with a greater focus on the end user: What sounds would the user want to be communicated to them (Selection)?

What attributes of these selected sounds would the user want to be communicated to them (Curation)? How should these curated attributes of the selected sounds be communicated to the user (Communication)? These models, in addition to other rhetorical and critical analyses (Zdenek, 2015, 2018; Martin, 2024), inform and frame the development and optimization of captions and SCTs for video content.

2.3 Aural diversity

Captions provide access to sound information to a diverse range of people, inclusive of both hearing and DHH people. For example, captions may benefit people who do not have speakers, people with learning disabilities, people who are learning new languages, and many more (Cavender and Ladner, 2008). *Aural diversity* is a term used to describe the plurality of senses of hearing—it is used to emphasize the recognition and acceptance that people hear in varied ways (Drever and Hugill, 2022). For example, it recognizes that a small subset of the world population falls into the International Standards Organization's (ISO) definition of "otologically normal" (their term) (for Standardization, 2023), and that in addition to those with clinically defined "conductive hearing loss" or "sensorineural hearing loss", there are numerous other conditions that affect sound perception, e.g., autism and tinnitus.

Within those who identify as D/deaf, or hard of hearing, there is also diversity of hearing, with the lines between them often blurred—there is no one "typical" D/deaf person (Holcomb, 2013). People who identify as "**D**eaf" relate to Deaf culture, which is not only audiological in nature but also political, linguistic, and social. Therefore, a single person might identify with more than one of the Deaf, deaf, and HoH labels. A central component of Deaf culture is the use of a signed language, e.g., American Sign Language (ASL) (Holcomb, 2013), a visual language formalized in the early 19th century when the American Deaf community was initially organized (Holcomb, 2013). Deaf signers in the United States are typically bilingual to some extent in both ASL and written English, but the fluency in each may vary significantly Holcomb (2013). In contrast, the terms "deaf" and "hard-of-hearing" refer primarily to audiological hearing levels in a clinical setting, namely "profound" hearing loss (90 dB or greater) and "mild" to "severe" hearing loss, respectively (Cavender and Ladner, 2008). However, colloquial usage of the terms vary and there is no hard line between "deaf" and "hard-of-hearing". For instance, while the term "hard-of-hearing" typically refers to someone with at least some residual hearing, they may or may not associate with Deaf culture; they may or may not use ASL; and they may or may not use assistive hearing devices such as hearings aids (HAs) or cochlear implants (CIs) (Holcomb, 2013). HAs use an external microphone and speaker typically warn over the ear to amplify low-level sounds more than high-level sounds in a frequency-dependent manner (Moore, 2012). CIs, on the other hand, use direct electrical stimulation of the cochlea, an internal hearing organ, to create the sensation of sound (Moore, 2012). Both HAs and CIs give rise to a diverse array of hearing experiences and abilities and are typically programmed to privilege speech over music and other sounds (Drever and Hugill, 2022).

3 Methods

We utilized a comprehensive review approach that combined a systematic literature review methodology with a mixed-methods survey and interview study to shed light on the current state of NSI captioning, user and captioner preferences, and offer insights into future research directions.

3.1 Systematic literature review

We conducted a systematic literature review to address RQ1. The review used a modified PRISMA (Page et al., 2021) framework, removing categories that did not apply to the HCI domain, such as "Effect measures". We selected *Google Scholar*, *ACM Digital Library*, and *IEEE Xplore* as databases for the review. Based on the scope of the project, we developed the following inclusion criteria:

- **Closed-captions:** Must relate to communication of sound information in video. Subtitles that only translate dialogue into another language and descriptions of static images are excluded.
- **NSI:** Must include detailed, nuanced, or novel representation or communication of NSI that goes beyond simple, nondescript NSI category captions or labeling (e.g. "[Music]", "[Laughter]").
- **Non-interactive:** Must relate to captioning of fixed-playback media and exclude interactive media, such as games.
- **Non-real-time:** Related primarily to pre-recorded media, and excludes contributions that focus primarily on real-time sound awareness or NSI detection/communication.
- **Contribution type:** A peer-reviewed journal or conference article, excluding theses, pre-prints (such as those on *arXiv.org*), and dissertations.
- **Date range:** Published within 2004 - 2024 (inclusive).
- **Language:** The paper must be written in English.

Based on these criteria and iterative refinement of keywords, we limited the date range to 2004–2024 and utilized the following search string: *("closed captioning" OR "subtitles") AND "non-speech" AND -site:arxiv.org*. Five authors were responsible for screening approximately 250 papers each, closely reading the title and abstract, and skimming the full paper for first-pass exclusion. 1276 papers were excluded during this first round. The 47 papers that remained were read in full by two authors who collaboratively evaluated each paper according to the inclusion criteria. Disagreements between authors were resolved through a close-reference of the inclusion criteria and clarifying understandings of concepts such as interactivity and real-time. 11 papers were excluded during this final evaluation process.

Following guidance on reporting HCI-specific systematic review (Rogers et al., 2024), we utilized an inductive *Best Fit* methodology (Carroll et al., 2013) to determine overarching categories for the 36 included papers, as categorization and group-level analysis were most appropriate for the goal of a broad understanding of the current state of NSI captioning research, user preferences, and captioner practices. The initial categories were

formed based on a scoping literature review of ten of the most recently published papers included in the analysis.

3.2 Survey and interview methods

We employed a mixed-methods approach, utilizing survey and semi-structured interview methodologies to investigate the current state of NSI captioning. Two populations of interest were identified: (1) professional *captioners* and (2) DHH *viewers*. All recruited participants were located in the USA and were comfortable reading English and were fluent in either spoken English and/or American Sign Language (ASL).

For the viewer survey, 39 participants were recruited through Deaf-centered email lists and social media groups, as well as through snowball sampling and word of mouth recruitment. These participants were asked to complete a survey to better understand their experience and preferences for NSI captioning. The participants recruited in this way were mainly DHH, although some non-DHH participants did complete the survey. This non-DHH participant data was analyzed separately and is available in the Supplemental material. We used the *Prolific* recruitment platform to recruit additional DHH viewers for the survey, utilizing the pre-screening feature to ensure that participants met the study's eligibility criteria. Specifically, we utilized Prolific's pre-screening criteria of only inviting participants on Prolific who self-identified as having "hearing loss or hearing difficulties" or "having a cochlear implant," to proceed with the study. All participants, regardless of the recruitment mechanism, completed screening and demographic questions to ensure eligibility requirements were met. A total of 150 participants were recruited through Prolific and were compensated at a rate of $4 per 15 min, leading to a total of 191 viewer survey responses, of which 168 were valid responses from DHH viewers. 15 of these participants were recruited for the interview portion of the study, as summarized in Table 1. This diverse sample allowed us to explore multiple perspectives within the captioning ecosystem.

The survey consisted of four main sections: (1) demographics and hearing status history, (2) current satisfaction with NSI captioning (3) NSI selection, curation, and communication (SCC) preferences by NSI type (music, sound effect, speaker identification, extra-speech information), and (4) initial reactions to proposed novel NSI SCTs from previous literature (Jeon et al., 2024; May et al., 2023; Alonzo et al., 2022; Gorman et al., 2021; Wang et al., 2016).

Questions in Section 3 were informed by prior literature, such as the use of the SCC framework (May et al., 2023) in organizing the questions and utilizing four specific NSI types/categories (Zdenek, 2018, 2015; May et al., 2024; de Lacerda Pataca et al., 2023). Anonymized survey data is available upon request.

Given the viewer-centered nature of this project and the challenges faced in recruiting professional captioners, as detailed in Section 6.6, we elected to recruit a cohort of 5 captioners to be interviewed, as shown in Table 2. To protect participant anonymity in the smaller pool of captioner participants, we report only the average age (32.5, $\sigma = 3.4$) and that on participant self-identified as male, and 4 self-identified as female. The insights gained from the captioner interviews were used only to add additional context to the primary results of the systematic literature review and viewer-centered methodologies.

To address RQ2, semi-structured interviews were conducted. The interview questions were rooted in the same prevalent themes identified in the literature review that guided the survey design. We elected to use this method as several previous studies employed open-ended interview techniques to identify broad themes regarding DHH users' experience with NSI communication (May et al., 2023; Alonzo et al., 2022; McDonnell et al., 2024). Therefore, we selected a semi-structured approach that was grounded in previously established themes to gain additional nuanced and detailed insights. This approach allowed us to connect themes identified across several papers in a single interview, putting DHH users' lived experiences and insights in conversation with trends in NSI captioning research. While the semi-structured nature of the interview allowed participants to introduce and explore novel themes, the literature-informed themes focused the interview, thereby limiting the time dedicated to the exploration of new ideas.

The themes used to create the semi-structured interview questions were the same as those used to structure the survey, namely the selection, curation, communication framework (May et al., 2023), and the four sub-categories of NSI (music, sound effects, speaker identification, manner of speech) when structuring questions. Most interview questions probed participants to expand on their answers in the survey and speculate on the contextual factors that impact their answers. For example, a question probing the curation of music captions for a viewer might be: "In the survey, you said that you prefer captions for music to use emotional adjectives, such as 'angry' or 'eerie'. Why are emotional adjectives important to you in music captions?" and related follow-up questions such as "Does the genre of the video or movie impact the importance of emotional adjectives?" Whereas questions for captioners focused on technical details of the captioning process and how they caption NSI in practice instead of personal preferences, such as: "You noted that you usually use emotional adjectives when captioning music. Is that generally included in a client's captioning style guide, or a choice you make yourself?"

In addition to asking questions regarding the SCC for each NSI sub-category, we asked both captioners and viewers four questions about their personal philosophical views on the role and function of closed captions, available in the Supplemental material. This was done to gain high-level insights into possible motivations that underlie captioning preferences, such as the relationship between a preference for no emotive adjectives in sound effect descriptions and a personal view that the role of captions is to provide strictly factual information.

Data were collected through interviews conducted via *Zoom* video conferencing software and participants were compensated with a $20 gift card. A licensed American Sign Language (ASL) interpreter was present on request for all interviews. To ensure privacy and data security, all sessions were recorded locally and scrubbed of any personally identifiable information before being transcribed using *Whisper* (Radford et al., 2023), an automatic speech recognition system, on a local machine. The first author (A1) manually verified all transcripts by comparing them to the recordings, ensuring data accuracy before analysis and removing

TABLE 1 Demographics of participants interviewed as part of the *viewer* group.

ID #	Hearing status	Age	Gender	Hearing assistive technology use	Sign language use
V1	Deaf	32	Female	None	Frequent
V2	HoH	30	Male	Hearing Aid on both sides for 8 years	Infrequent
V3	Deaf	45	Female	None	Frequent
V4	Deaf, deaf, HoH	50	Male	CI on both sides for 5 years	Frequent
V5	HoH	25	Male	None	Infrequent
V6	Deaf, deaf, HoH	28	Female	Hearing Aid on one side for 10+ years, and a CI on the other for 10+ years	Infrequent
V7	Deaf	32	Male	Hearing Aid on one side for 10+ years	Frequent
V8	HoH	52	Female	Hearing Aid on one side for 10+ years, and a CI on the other for 6 years	Never
V9	Deaf, deaf	54	Female	CI on one side for 10+ years	Never
V10	HoH	31	Male	None	Infrequent
V11	HoH	29	Female	None	Never
V12	Deaf	47	Male	None	Frequent
V13	Hearing, child of Deaf adult (CODA) [a]	23	Male	None	Frequent
V14	HoH	25	Male	None	Never
V15	Deaf	49	Female	None	Frequent

[a] While not DHH themselves, CODAs are an important part of the Deaf community and have rich lived-experience of the interplay between Deaf and Hearing culture Tripp (2023). Therefore, we elected to include a CODA in our study to include this perspective.

all personally identifiable information. We leveraged a combination of both inductive and deductive thematic analyses to analyze our interview data. An initial codebook was derived from the relevant literature by A1 and further refined throughout the coding process. The selection, curation, and communication (SCC) of NSI captioning analysis (May et al., 2023) was selected as part of the *a priori* codes for interview analysis and as an organizing scheme for the survey questionnaires. Additionally, four primary categories of NSI, namely (a) sound effects and ambient sounds (SFX), (b) music, (c) speaker identification, and (d) manner of speech and paralinguistic information (MoS), were identified as meaningful NSI groupings in both previous literature and the interview analysis (May et al., 2023; Jain et al., 2021; May et al., 2024). When a new code was added, each transcript was re-coded by the authors to ensure data consistency throughout the coding process. For viewer transcripts, three authors (A1, A3, and A4) independently coded four transcripts, achieving an inter-coder reliability score (Krippendorf's Alpha) of 46%. The authors then went through one transcript fully together, settling discrepancies within the codes and further refining the codebook together. After the initial meeting, the authors re-coded those four transcripts and achieved an inter-coder reliability score of 82%, before dividing and coding the remaining 12 transcripts. For captioner transcripts, the process was similar in that all three authors coded one transcript with 79% inter-coder reliability, then divided the other four for individual coding. All procedures were endorsed by Stanford University's Internal Review Board.

Of the 191 total survey participants, 168 identified as DHH and were included in the study. The data of the 21 non-DHH participants with traditional hearing who completed the survey were analyzed separately and included in the Supplemental material. Of the 168 DHH participants, 83 (49%) self-identified as male, 78 (46%) as female, and 7 (4%) as non-binary or genderfluid. 97 (58%) self-identified as White/Caucasian, 52 (31%) as Black and/or African American, 6 as Asian (3%), and 10 (5%) as multi- or bi-racial. The mean age of DHH participants was 40.27 ($\sigma = 16.2$). 51% of DHH participants never use a signed language, while 16% sign frequently, 38% used only hearing aids (HAs), 24% only used cochlear implants (CIs), 8% used both HAs and CIs, and 30% did not regularly use any hearing assistive technologies.

TABLE 2 Demographics of participants interviewed as part of the *captioner* group.

ID #	Captioning experience
C1	Real-time captioning in an educational context, captioning video content as a freelancer
C2	Real-time captioning in an educational context
C3	Professional captioner at a large captioning firm with 10+ years of experience
C4	Professional captioner at a large captioning firm with 10+ years of experience
C5	Filmmaker and captioner

While the cultural distinctions between HoH and Deaf are nuanced, complex, and overlapping, it is important to try understand the approximate relative proportion of these identities to investigate possible bias in the participant sampling. According to the 2021 American Community Survey, approximately 11 million individuals consider themselves D/deaf or have "serious difficulty hearing", with the Hearing Loss Association of America

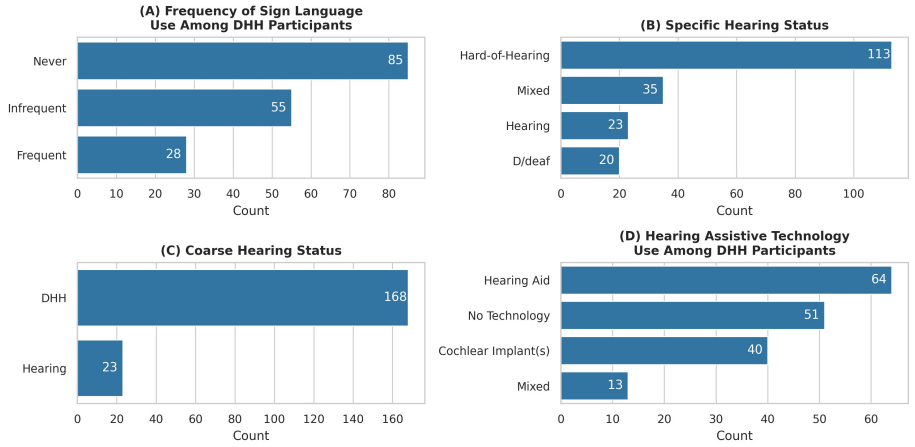

FIGURE 1
DHH survey participants by **(A)** frequency of sign language use, **(B)** specific hearing status, **(C)** coarse hearing status, and **(D)** hearing assistive technology use among DHH participants. Mixed hearing status means a person identified as more than one status (e.g. D/deaf and Hard-of-Hearing), and mixed assistive technology use means the participant uses more than one type of technology (e.g. Hearing aids and cochlear implant).

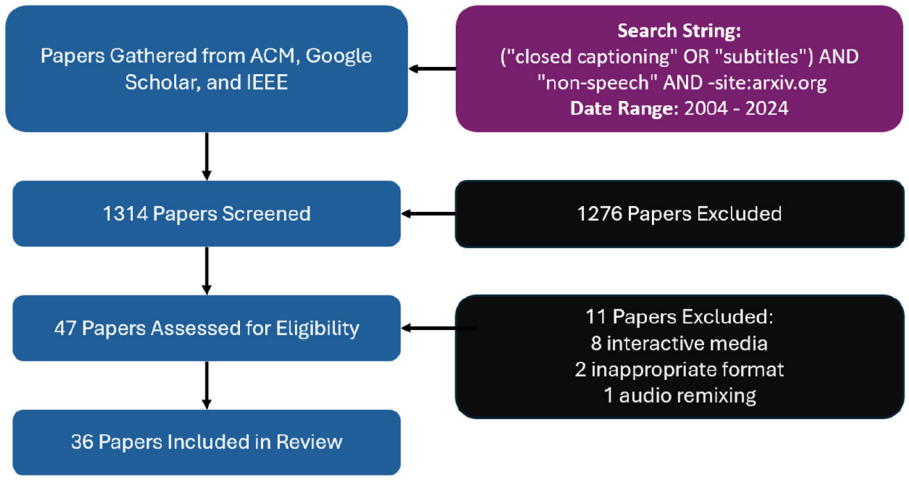

FIGURE 2
Details of the PRISMA paper collection process from three databases, screening of the 1,314 papers, detailed eligibility assessment of 47 papers, and inclusion of 36 papers.

estimates that 48 million Americans have some degree of hearing loss (United States Census Bureau, 2021; of America, 2023). Based on this, it can be approximated that ~25% of the DHH community would identify as D/deaf and ~75% as HoH. As summarized in Figure 1, 67% of DHH participants self-identified as HoH, 12% as D/deaf, and 21% as both D/deaf and HoH. Therefore, our recruited sample may slightly over-represent D/deaf participants, but this is difficult to accurately estimate given the nuances and complexities of these identities, as highlighted by the 21% of participants who identified as both D/deaf and HoH.

4 Systematic literature review results

To address RQ1, 36 papers consisting of 59 studies (as some papers contain multiple studies) were selected and analyzed, as shown in Figure 2 and described in Section 3.1. The results of the analysis are summarized in Tables 3–6. The paper metadata showed a clear trend of recent activity on NSI captions with 19 papers published between 2020 and 2024, while previous years had far fewer (2005–2009: eight papers; 2010–2014: seven papers; 2015–2019: two papers). Papers were published in a variety of venues, with 14 coming from ACM venues (9 from CHI[1] proceedings and 3 from ASSETS,[2] proceedings) 7 from IEEE venues, and 15 from other journals and conferences (3 from the *Computers Helping People with Special Needs* conference and 2

1 The ACM (Association of Computing Machinery) CHI conference on Human Factors in Computing Systems.

2 The ACM SIGACCESS (ACM Special Interest Group on Accessibility and Computing) Conference on Computers and Accessibility.

from the *Telecommunications Journal of Australia*). Notably, three papers from the non-ACM or IEEE venues were explicitly from humanities-centered disciplines, namely media studies, rhetorical studies, and disability studies. 13 of the 19 papers published between 2020 and 2024 were part of the proceedings of an ACM conference, showing a clear trend that the majority of NSI captioning research is recent and that the majority of this recent research is published at an ACM venue.

The majority of papers (81%, N = 29) involved research with participants in at least one study, while nearly one-fifth (19%, N = 7) consisted solely of research without participants, such as critical analysis and dataset creation. As shown in Table 3, NSI captioning research with participants primarily focused on user studies (67%, N = 24), where researchers developed and tested prototypes. Methods for gaining insights into users' experiences, e.g., interviews (14%, N = 5) and surveys (8%, N = 3), were less commonly used compared to user testing.

Of the 29 papers with participants, the majority conducted research with participants as viewers of NSI captioning: 28 papers (97%) included viewers in at least one study, while only one paper (3%) focused on content creators. Among the 29 papers, a total of 49 studies were conducted. Of these, 44 studies (90%) involved viewers: 42 studies (86%) were solely with viewers, while two studies (4%) included both viewers and other participants such as authors or content creators. Five studies (10%) involved participants like content creators, VR developers, technologists, and captioners instead of viewers.

As shown in Table 4, of the 59 studies, the majority (68%, N=40) included DHH participants. No study attempted to distinguish between D/deaf and Hard-of-Hearing participants when recruiting. 32 studies (54%) involved only DHH participants, while 8 studies (14%) included both DHH and hearing participants. 5 studies (8%) included only hearing participants. Additionally, four studies (7%) did not specify the participants' hearing status, leaving it unclear whether the participants were DHH or hearing.

Additionally, among the 40 studies across 24 papers with DHH participants, it was fairly common not to specify the hearing identity of DHH people (whether they were d/Deaf and/or HoH): 25% of the papers (N = 6) did not specify participants' DHH identity, covering more than 40% of the studies (N = 17). On the other hand, five studies (20%) across four papers (16%) clearly distinguished between deaf and Deaf participants.

Table 5 shows the median and range of participants for the methods that were used in two or more studies, with user studies being the most prevalent method, being used in over 70% of included studies. The number of interview participants was relatively small, with a median of nine for both total and DHH participants. In contrast, user studies had a larger median number of participants: 19 in total and 16 for DHH participants. Additionally, although the sample size is limited, we found that all methods primarily focused on DHH participants, utilizing a median number of 53.5 participants for survey-based studies.

The results of the best-fit categorization analysis of the systematic literature review are shown in Table 6, with relevant literature grouped by their research questions and methodologies. G1-3 in Table 6 summarized research on by NSI type (sound effects/music, MoS, and speaker identification respectively). G4 highlights the theoretical contributions of rhetorical and media studies analysis of captions, which inform high-level conceptions of the goals and purpose of captions. Finally, G5 illustrates themes explored by papers investigating current captioning practices on various platforms and by captioners of varying levels.

TABLE 3 The frequency of methods used across all analyzed studies.

Method	Papers	Studies
User study[a]	24 (67%)	37 (63%)
Critical analysis	7 (19%)	7 (12%)
Interview	5 (14%)	5 (8%)
Survey	3 (8%)	3 (5%)
Focus group	2 (6%)	2 (3%)
Participatory design, dataset creation, DL model training, proposing a new workflow, workshop	1 (3%)	1 (2%)

[a]Some research involving user studies includes pre- and post-surveys or interviews. However, we treated these as part of the overall user study rather than as separate methods, as their main purpose was to gather information or feedback related to the user testing.

TABLE 4 Participants' hearing identity across all analyzed studies.

Participants and hearing identity	Study count
DHH involved	40 (68%)
- Among DHH involved, no hearing people participated.	32 (54%)
- Among DHH involved, both DHH and hearing people participated.	8 (14%)
d/Deaf people only	0 (0%)
HoH people only	0 (0%)
Hearing people only	5 (8%)
Participants hearing identity not specified	4 (7%)
No participants involved	10 (17%)

The analysis highlighted that a plurality of papers (36%) investigated manner of speech and paralinguistic information (MoS), with 19% focusing on sound effects and ambient sounds, 17% on speaker identification and labeling, 13% on understanding current practices, 11% using critical theory or rhetorical analysis to analyze the role of closed-captioning, 8% on music communication, and 6% on using vibrotactile haptics. 8% of papers investigated more than one of the previously mentioned categories, such as Alonzo et al. (2022) investigating speaker labeling, sound effects, and music. Direct extension or elaboration of previous research was found among 11% of the papers, with no papers experimentally comparing their methods with previous methods and no papers using longitudinal methodologies such as diary studies. Additionally, all papers that employed user studies used short clips (<10 min).

Table 6 also highlights several additional insights that spanned across papers. Various communication paradigms were explored across NSI types, with kinetic or modified typography[3] being

3 Kinetic typography refers to changing the visual characteristics of text, such as color or size, over time. Modified typography refers to changing visual characteristics of text in a static way, such as using font weight or color to communicate importance.

TABLE 5 Median and range of participants across methods in the studies.

Method	Median participant # (min-max)	d/Deaf	HoH	DHH	Hearing
User study ($N = 24$)	19 (2–314)	7 (0–28)	5 (0–16)	16 (0–44)	0[a] (0–117)
Interview ($N = 5$)	9 (6–13)	5 (4–9)	3 (2–6)	9 (6–13)	0 (N/A)
Survey ($N = 3$)	83 (62–102)	N/A[b]	N/A[b]	53.5 (24–83)	39 (0–78)
Focus group ($N = 2$)	34.5 (24–45)	N/A[b]	N/A[b]	29.5 (24–35)	5 (0–10)

[a]More than half of the studies did not include hearing participants. The median number of hearing participants in the studies that did involve them is 10.
[b]An insufficient number of studies specified the separate number of d/Deaf and HoH participants.

the most popular. The techniques explored in previous research included altering text color, font, weight, and the position of individual letters in a word. Emojis or icons were also used in several papers, with more abstract sound communication strategies, such as sound-reactive animations, explored by only a handful of papers. Additionally, when papers explored mood, they generally used a model of discrete basic emotions (such as happiness, disgust, etc.), with several models being used ranging from four to six moods. Fewer papers opted to use a continual valence-arousal model.

The broader context of caption production and use was a theme addressed by several papers (Table 6, G5), such as the way excerpts from TV shows and movies, when posted on social media, have captioning that may differ substantially in style or quality from the original content (May et al., 2024). Only one study directly looked at the captioning production process (Barbero et al., 2010) and suggested that additional research may be needed in (1) empowering captioners to be more involved in the entire production process and (2) streamlining and simplifying caption file management and handling in large production houses. The literature suggests that improvement in these two areas of captioning could greatly impact the quality and availability of high-quality captions (that are more likely to contain NSI) in online streaming services.

The findings in this section highlight several key trends relevant to NSI captioning research. The focus on viewers as participants emphasizes the user-centric nature of this field. The predominance of user studies underscores the importance of direct evaluation of NSI captioning methods with DHH users. The analysis of participant numbers across methods provides insights for future research design and highlights opportunities for future work, such as larger longitudinal or comparative studies, as well as more studies that incorporate caption creation and distribution.

Examining users' interpretations of the exact purpose and function of closed-captioning provides additional insight into the specific issues, affordances, and opportunities of NSI captioning. Captions facilitate access to audiovisual content and foster a more inclusive media landscape, but they are, however, not a neutral, perfect, or exact translation (Zdenek, 2018). Particularly regarding NSI, many curatorial and aesthetic decisions are made by captioners, such as which sounds are worth captioning and what adjective best describes a particular piece of music in context (Zdenek, 2015, 2011).

5 Results of mixed methods analysis

The following section presents the results of the survey and interview studies. DHH survey participants were asked about the frequencies of certain issues that may occur while watching video content with insufficient NSI captions. Figure 3A shows that none of the provided issues were significantly more pronounced than others but that DHH viewers are experiencing these issues somewhat frequently. DHH viewer's most pertinent issue with NSI captioning was a lack of NSI captions in general, as shown in Figure 3B. A lack of correct information in the NSI captions, such as containing spoilers or a lack of detail, as well as timing issues with NSI captions, were generally more of a concern to DHH viewers than NSI being over-captioned.

5.1 Selection, curation, and communication of NSI

5.1.1 Selection: what NSI users want to be communicated to them

Figure 4A illustrates that 57% of the 168 DHH survey participants wanted both narratively important SFX and those that establish tone/mood, while 9% opted for exclusively non-obvious SFX, and 14% wanted every SFX communicated. Selection preference patterns were similar in music, as shown in Figure 4B, with a near majority (48%) preference for music that establishes tone/mood or is narratively important, with the proportion of participants wanting either all music (8%) or exclusively narratively important music (26%) being comparable to SFX. Both speaker identification Figure 4C and MoS Figure 4D had majority consensus with 53% of participants only wanting speakers to be identified when it is unclear, and 54% selecting MoS to be communicated only when it is either important or not visually obvious.

5.1.1.1 Narrative importance

The importance or relevance of a specific piece of NSI to the narrative, plot, or comprehensibility of a video was highlighted as a relevant factor in NSI selection for both viewers and captioners. When discussing what sound effects they would like captioned, V7 highlighted that *"just the important sounds because we don't really have to have all of the sounds"*. However, opinions varied among participants, ranging from V15 *"want[ing] to know everything about all the sounds"*, to V2's reflection on the importance of other

TABLE 6 Overview of categories identified in NSI closed-captioning literature from 2004 to 2024.

NSI closed-captions literature		
User experiences and challenges	**Selection, Curation, and Communication of NSI**	**Contextual factors and solutions**
G1. Sound effects, music, and ambient sounds		
(1) Exact timing, duration, and loudness Kushalnagar et al. (2014); Climent et al. (2021) (2) Source location or identification Wang et al. (2016); Climent et al. (2021); Jain et al. (2021); May et al. (2023) (3) Musical features such as timbre and pitch May et al. (2023); Choi et al. (2024)	(1) Related to sound effects (Kushalnagar et al., 2014; Wang et al., 2016; Climent et al., 2021; Barbero et al., 2010; Alonzo et al., 2022; de Lacerda Pataca and Costa, 2023) (2) Related to music (May et al., 2023; Choi et al., 2024) (3) Utilizing vibro-tactile haptics (Kushalnagar et al., 2014; Jain et al., 2021; Choi et al., 2024) (4) Utilizing kinetic typography (Wang et al., 2016) (5) Using emojis/icons (Climent et al., 2021; de Lacerda Pataca and Costa, 2023; Alonzo et al., 2022) (6) Using other visualization strategies (Jain et al., 2021; May et al., 2023; Choi et al., 2024) (7) Customizing linguistic content (Barbero et al., 2010)	(1) Visual noise (Wang et al., 2016; Alonzo et al., 2022) (2) Narrative importance (Climent et al., 2021; May et al., 2023; Alonzo et al., 2022)
G2. Manner of speech and paralinguistic information		
(1) Identifying paralinguistic cues Fourney and Fels (2008); Jeon et al. (2024) (2) Paralinguistic cues misidentification leading to negative experiences	(1) Four basic emotions (Vy et al., 2008; Mori and Fels, 2009) (2) Five basic emotions (Rashid et al., 2006, 2008; Jeon et al., 2024) (3) Six basic emotions (Fels et al., 2005; Lee et al., 2007; Kim et al., 2023b) (4) Emphasis (Kim et al., 2023b) (5) Valence/arousal model (de Lacerda Pataca et al., 2023) (6) Explored prosody (de Lacerda Pataca and Costa, 2023; de Lacerda Pataca et al., 2023) (7) Single emotion (Fourney and Fels, 2008) (8) Emoji or icon (Fels et al., 2005; Lee et al., 2007; Mendis et al., 2022; Alonzo et al., 2022) (9) Color of words (Rashid et al., 2008; Fourney and Fels, 2008; Vy et al., 2008; Mori and Fels, 2009; de Lacerda Pataca et al., 2023, 2024; Jeon et al., 2024) (10) Color of additional elements such as background or border (Fels et al., 2005; Lee et al., 2007; de Lacerda Pataca et al., 2024) (11) Animation (Fels et al., 2005; Lee et al., 2007) (12) Kinetic typography (Rashid et al., 2006; Fourney and Fels, 2008; Vy et al., 2008; Mori and Fels, 2009; de Lacerda Pataca and Costa, 2023; de Lacerda Pataca et al., 2023, 2024; Rashid et al., 2008)	(1) Distraction from video content Lee et al. (2007); de Lacerda Pataca et al. (2023); Fourney and Fels (2008) (2) Confusion over mapping de Lacerda Pataca et al. (2023); Mori and Fels (2009) (3) Increased immersion or connection de Lacerda Pataca et al. (2024); Lee et al. (2007)
G3. Speaker identification		
(1) Identify speakers	(1) Color of text (Gorman et al., 2021) (2) Using a speech bubble (de Lacerda Pataca and Costa, 2023) (3) Speaker avatar (Vy and Fels, 2009, 2010, 2011) (4) Color of caption border (Vy and Fels, 2009)	(1) Number of speakers in scene (Vy and Fels, 2009; Gorman et al., 2021) (2) Speaker intelligibility (Gorman et al., 2021) (3) Differences between hearing status (Vy and Fels, 2009, 2011) (4) Customization (Vy and Fels, 2011; Gorman et al., 2021) (5) Genre (de Lacerda Pataca and Costa, 2023; Gorman et al., 2021)
G4. Critical inquiry and theory		
	(1) Theory of NSI selection and evaluation (Zdenek, 2011; Tsaousi, 2015) (2) Translation and interpretation tensions (Martin, 2024; Zdenek, 2011) (3) Connection to existing design frameworks (Udo and Fels, 2010) (4) Lack of integration in production process (Udo and Fels, 2010)	
G5. Current practices and standards		
		(1) Platform: YouTube (Li et al., 2022; May et al., 2024) and TikTok (McDonnell et al., 2024) (2) User-generated captions (Li et al., 2022; McDonnell et al., 2024; May et al., 2024) (3) Professional captions (Liu et al., 2022; Kim et al., 2023a; May et al., 2024) (4) Quality assessment of NSI (Liu et al., 2022; Kim et al., 2023a)

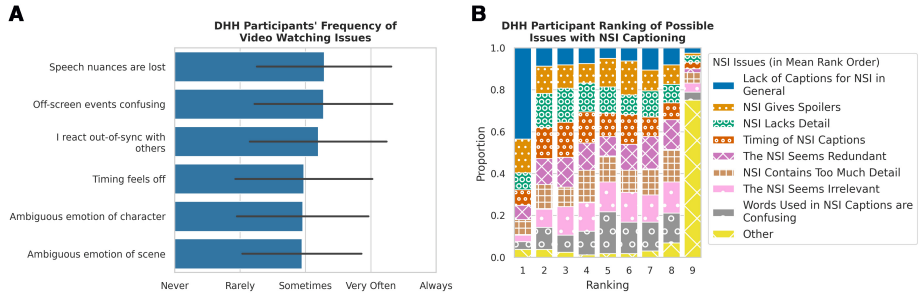

FIGURE 3
Frequency of possible issues faced by DHH viewers by **(A)** mean frequency and **(B)** ranked frequency.

elements in video content, *"If you just want [to give me] all the information, just send me the screenplay, and I'll just read it. But like, no. I want to experience it right? There's emotion here. There's pacing"*.

Also referred to as "plot pertinence" by captioners, narrative importance appeared as a guiding selection principle for captioners in deciding what NSI to caption. Both C3 and C4 highlighted that selecting NSI to caption was heavily informed by narrative importance, with C3 summarizing their guiding principle in NSI selection, *"do they need this sound to follow the video?"* In an educational setting, captioners mentioned that incidental sounds, such as sneezing or a dropped water bottle, were only deemed necessary to caption if it was reacted to by others.

5.1.1.2 Visual redundancy

The presence of other visual indications of NSI was a parameter deemed relevant by many viewers and captioners. For example, if a glass bottle can clearly be seen shattering on screen, then the NSI of glass shattering has a high degree of visual redundancy. If the sound-causing or supporting action occurs off-screen or is visually obscured, then that NSI would have a low degree of visual redundancy. V1 noted that *"if [captions are] adding more detailed information, like 'rapid gunfire', that's helpful. But if it just says 'gunfire' or 'gunshot', and you can see [a gun firing] on the screen, that's not that helpful"*. Visual redundancy and narrative importance were often considered together, such as in V12's insight: *"Is that important to the storyline? I don't know if it's just like a regular gunshot in a regular movie. Obviously, I can see that person is, you know, being shot at or whatever. I wouldn't need that. But if it's off-screen, like in the background, then yes, that would be good to know"*.

5.1.1.3 Balancing NSI: clutter vs. clarity

Capturing every sound effect in a video can lead to cluttered and overwhelming captions, detracting from the main content, particularly in dialogue-heavy scenes. Participants expressed concerns about this balance, with V12 noting, *"My concern would be it captioning every sound and every thing, it kind of clutters what's happening"*. Similarly, V8 reflected on the potential for overkill in NSI captioning, *"[if they label every single thing,] I could imagine it could be a little overkill or distracting"*. The pacing and relative density of narratively important NSI are crucial factors captioners consider when selecting NSI to caption, with C1 clarifying that *"captioning one thing [often] means I can't caption something else"*. It is often not feasible to communicate all NSI at every moment, and captioners prioritize and filter NSI, in a contextually aware manner, *"What's going on? Can I squeeze [an NSI caption] in?"* (C3).

5.1.1.4 Omissions in NSI captioning

One of the most significant challenges with NSI captions is the omission or misrepresentation of important sounds, which can lead to a lack of context and understanding for viewers. For instance, sound effects that are essential to the narrative or atmosphere are sometimes not captured. As V12 described a moment in a video, *"[The character] was like, 'Oh, do you hear that? Wow, that's loud.' But that part wasn't captioned."* Previous research (May et al., 2024) has highlighted the variable presence and density of NSI captioning across YouTube channels, regardless of whether an independent content producer or a large production studio publishes the content. Figure 3B illustrates this as "lack of NSI captioning in general" was the NSI captioning issue experienced most frequently by survey participants.

5.1.2 Curation: what about the selected NSI do users want to know?

5.1.2.1 Objective vs. interpretive information

A common point of discussion was the tension between using factual, more objective words to describe NSI (e.g. "[Loud piano music]" or "[Rapid gunfire]") as opposed to words that require additional subjective interpretation (e.g. "[Joyful victory music]" or "[Frustrated gunfire]"). Captions that were viewed as overly interpreting information, such as describing music with emotional or affective language (e.g., "[Melancholic melody plays]"), were viewed as *"infantilizing"* (V8) or *"spoon feeding"* (V2) by some participants. These participants often preferred factual descriptions of the NSI's sonic properties, such as "[Slow violin music]", as this afforded them increased agency in their video viewing. V2 noted, *"Don't tell me it's 'spooky alien music', tell me about the music and I'll decide if it's spooky in context"*. V6 commented that information parity with a hearing audience member is most important, wanting to know *"how would most people interpret [the sound]"*, and if the sound is more ambiguous *"just describe it sonically"*. Overall, participants indicated a preference for objective

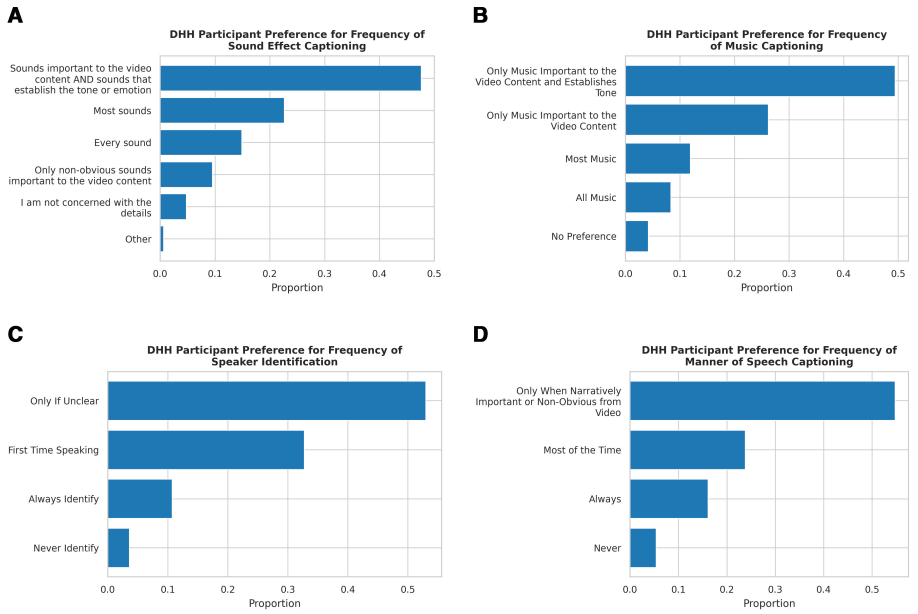

FIGURE 4
Participant preferences for frequency of NSI captioning by type: **(A)** sound effect, **(B)** music, **(C)** speaker identification, and **(D)** manner of speech.

features, such as the type and exact timing of the sound in SFX curation (Figure 5A), with volume and location being secondary considerations. For manner of speech, as shown in Figure 5C, participants found intentional deviations from normal speech to be the most important aspect of MoS to be captioned, followed by emotion/affect and then volume/loudness.

However, many other participants preferred mainly interpretive information, with V8 commenting *"tell me what the music **does**. I don't care if it's violins or whatever"*. This difference in preference is illustrated in Figure 5B, where more objective (style/genre) and interpretive (mood/emotion) attributes of music were rated nearly identically. A similar pattern was found in SFX curation, with some participants highlighting that describing the sound itself or the action that caused the sound was an important factor for them. For example, a caption of "[Camera Beeps]" describes the sound ('beeps') and the origin of the sound ('camera'), whereas "[Camera Turns Off]" describes the action that resulted in the beep rather than the beep itself. V4 highlighted that equity of information access was important to them in the sound vs. action in SFX captioning, asking, *"If a hearing person were to listen to this, would they know that it's the sound of the camera turning off?"*

5.1.2.2 Cultural information

For some viewers, the cultural context plays an important role in deciding what information about NSI they would like communciated to them. For example, even if knowing the meta-data of music playing in a video, such as the title and artist of a song, is not immediately helpful in the context of viewing the video, it can be helpful in gaining additional cultural insights. V12 summarized this by saying: "*It's not like I live under a rock. I mean, I'm Deaf, but like, I do go out... I want to know if it's Lady Gaga. [My friends] might talk about it later.*" However, an over-reliance on assumed cultural knowledge could also lead to confusion, such as V14 commenting on using cultural information to describe MoS, such as "[Singing like Dolly Parton]," V14 commented:*"If you don't know who Dolly Parton is, that's not gonna be helpful."*

An additional point of tension in culturally aware information communication is in balancing the efficiency of descriptions without assuming information, such as the gender or race of a person. For example, it might be more efficient to caption gender markers as a means of speaker identification, such as "[Man]", as opposed to other visual identifiers, such as "[Person in Blue Shirt]". C1 highlighted this tension: *"the [balancing act] of trying to be helpful and identifying who's speaking without totally applying things to the person that you don't know."*

5.1.3 Communication: how should the curated information about the selected NSI be communicated?

5.1.3.1 Capturing musical elements

Music in captions presents unique challenges. While music is a crucial component of many media experiences, conveying its mood and emotional impact through captions is difficult. As V4 mentioned: *"You only have a limited amount of time and information that you can show... my problem generally with captioning music is that you don't really get the mood across about what's really happening in the music."* Communicating temporal aspects of music, such as changes in volume or affect, can be challenging. This is due to the relatively slow rate at which captions can be appropriately shown compared to the rate at which temporal aspects of music change. For example, conveying the rising tension of violins morphing from creaking, to faint notes, to audible notes, to screeching in a horror movie.

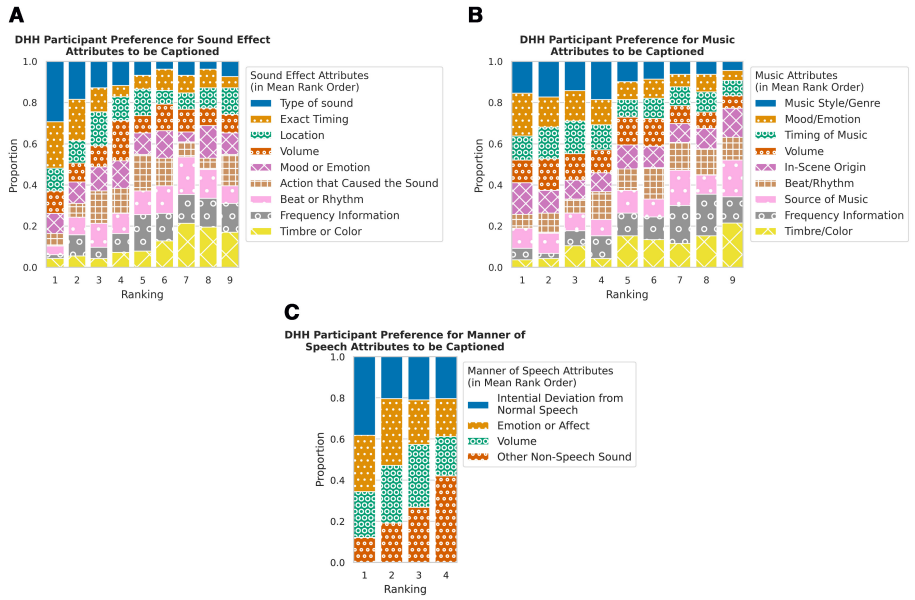

FIGURE 5
Surveyed DHH participants' ranking of attributes pertinent to **(A)** sound effects, **(B)** music, and **(C)** manner of speech. Speaker label was omitted as there are fewer relevant attributes. Bar segments are colored and patterned in mean rank order.

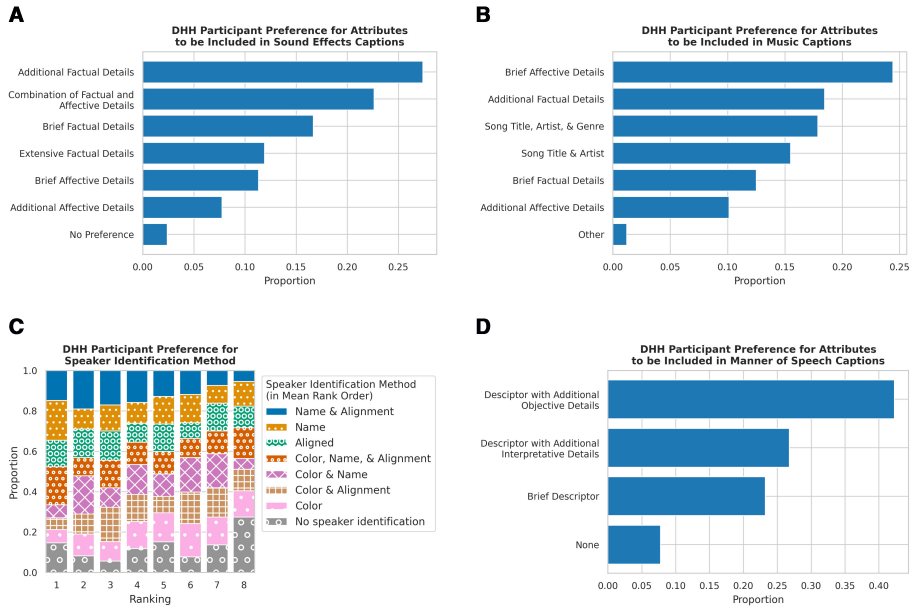

FIGURE 6
Surveyed DHH participants' preferred approach for captioning NSI by NSI type: **(A)** sound effects, **(B)** music, **(C)** speaker label, **(D)** manner of speech.

5.1.3.2 Confusion in speaker identification

A significant challenge in NSI captioning is the accurate identification of speakers, particularly in scenes involving multiple characters. Mislabeling or a lack of clear distinction between speakers can lead to confusion and disrupt the viewer's understanding of the dialogue. V12 highlighted an example of this issue: *"It looked like it was one consistent statement, and we couldn't figure out who was talking. I was like, I don't know what's happening in this situation."* This confusion is further compounded when captions introduce character names that the viewer may not yet recognize, as V12 pointed out: *"There was one time where the captions gave the name of a person, and the audience really obviously wasn't supposed to know, but I knew because of the captioning."* Subtle choices in speaker identification, such as revealing a character's name in captions before they've been named in the show's dialogue, can greatly impact dramatic suspense, as

V12 noted, *"if it's captioned as '[Waitress]', we might think, 'okay, maybe she's like a minor character'. If she's captioned as '[Sarah]', and we don't know her name yet, then we know, 'oh, her name's Sarah, she's an important character', but that hasn't been explained to us yet. It sort of spoils things"*.

5.1.3.3 Distraction

A theme present in much of the existing research (Fels et al., 2005; Rashid et al., 2008; Fourney and Fels, 2008; Vy et al., 2008; Mori and Fels, 2009; Kim et al., 2023b), and raised by several viewers is the potential for captions and other sound communication technologies to distract or overwhelm viewers. V9 gave an example of their experience watching season four of Netflix's *Stanger Things*: *"I did not think it was possible, but they proved it's possible to over-caption... It was 'the caption show.' As the captions were being talked about, it's not a good thing because captions should be easy, simple, quick."*, additionally summarizing their stance that *"good captions are scanned, bad captions are read."*

When describing a recent trend of *poetic* or *creative* captions, which can use extended metaphors and many adjectives to describe NSI, C5 noted how these captions can *"actually, just confuse me and take me out of the video because I'm sort of so lost. So in a way, they felt like a weapon... Like virtue signaling"* V8's experience was more "*situational... I think this kind of poetic description or going into the thoughts of the speaker, viewing away from functional users, I think becomes a different genre or format. I think it's really a creative format. I really like it because it gives you a different way of looking at these really functional words. But there's a time and place for everything.*"

Many viewers highlighted that communication techniques that require additional attention, such as poetic captions, or add significant novel visual information that the user is not accustomed to, such as a sound visualizer, are particularly prone to distracting viewers from the video content itself. However, acclimation and previous experience were an important factor in the probability of sound communication technologies distracting a viewer. For example, many caption users in Europe and the UK prefer speaker labels to be differentiated using text color (Gorman et al., 2021), while this practice is far less common in North America, with V7 noting *"one color [for caption text] would be better because otherwise it's too disorientating."* Differences between hearing status groups were also found to be a modulating factor in reports of distraction in previous literature (Fels et al., 2005; Rashid et al., 2008; Lee et al., 2007), and were expressed by V2: "I can hear most of the sound so just help me out, don't overwhelm me with things I don't need."

5.2 Contextual factors in NSI captioning

Viewer preferences for NSI captioning are not fixed and are modulated by several contextual factors.

5.2.1 Video context

5.2.1.1 Genre dependence

The content's genre plays a significant role in shaping user expectations and preferences for captions. Different genres evoke distinct emotions, pacing, and complexities that impact the type and level of captioning desired. For instance, action/sci-fi viewers may prefer more contextual information, while those watching comedy or light-hearted content might prioritize following the natural flow of the dialogue. V1 noted that for sci-fi genres, *"[I] would probably like more contextual information. You know, a lot of [off-screen] hearing what's happening that you can't see, the actions, who's doing something. I think that would impact the storyline"*.

For news and informational content, accuracy and clarity of captions may be more of a priority than an emotional connection with the media. V1 commented on the timing element of captions regarding watching news, *"If I'm watching the news and just trying to learn and just see what's going on, then maybe pausing [so that the ESI captions would be in sync] would be [a better fit]"*. In horror or suspense genres, capturing sound effects and music in captions is crucial for building atmosphere. There are many elements of horror that are particularly challenging to caption. V15 noted the discongruity between visual indicators and the captions, *"In the whole movie [Hannibal Lecter], his voice is very just monotone, it's just very chilly and all that, and I only figured that out from the facial expressions and the lighting on him"*. V4 recalled issues that dealt with captioning the mood of the music prior to the scene unfolding, *"If you simply put it in the captions, '[creepy music]', then it completely ruins it. It's not the same"*. V2 also commented on the inadequacy of captioning for music within this context, *"if you're going to tell me that there's 'spooky alien music' or something, that doesn't do anything for me. That's not the purpose... The purpose of the music is to create an affect and to create an atmosphere"*.

Certain genres of video content afford different solutions; for example, the genre needs of a comedy panel show, where the number of speakers is generally fixed for the duration of an episode, are different from content that has a larger or dynamic set of speakers (Gorman et al., 2021; de Lacerda Pataca and Costa, 2023). In the survey, participants were shown stills of four different current captioning technologies, namely kinetic typography to indicate emotion, adaptive caption to select caption colors for individual speakers, emojis to indicate emotion, and comic-style text to communicate SFX. Participants indicated during interviews and in open-response survey questions that genre plays an important role in the possible use of these technologies, indicating that, for example, kinetic typography might be most appropriate for dramatic movies, whereas comic-style text might function best for action and animated movies. V2 mentioned how genre conventions normalized open captions in *"...these Japanese game shows, [the open captions] look like stylized, comic-book sound effects. Cute things are round, awkward things more jagged... It's part of the show"*. V8 commented how matching the intended visual aesthetic of the video would be important for them, noting, *"you can't just put an emoji [on screen], it breaks the whole visual style"*.

5.2.1.2 The interplay of audio quality and captioning needs

The audio quality of the viewing environment significantly impacts how viewers rely on and prefer captions. V2 observed, *"I think that one kind of nuance here is that if you're watching on your phone, you're getting a lot less auditory information than you are if you're watching it in a cinema."*. In situations with

poorer audio quality, such as on mobile devices or in noisy environments, clear and accessible captions become even more crucial for comprehension and accessibility. V2 continued, *"The other problem with a movie theater is that you don't have control over anything"*. Audio fidelity, hearing ability, listening environment, and hearing assistive technology state all impact the viewer's access to audio, which in turn may impact NSI captioning preferences. This dynamic interplay between audio access and NSI captioning needs highlights the importance of considering the audio context when designing captioning solutions.

5.2.2 Viewing context

5.2.2.1 Impact of viewing location on caption preferences

The physical environment in which media is consumed significantly impacts preferences for caption styles and features. Users may favor condensed captions due to limited space on smaller screens like those found on phones or tablets. V1 mentioned, *"If I am on a train or any sort of transportation and I'm using my phone, I would probably prefer [that] the captions are a little more condensed, just because the phone has a smaller screen"*. Conversely, in more environments that are more private or have fewer distractions, such as at home, viewers may desire more detailed captions that enhance their immersion in the content. V1 explained, *"At home, I'm usually chilling, just laying on the couch, just kind of consuming the content. And so more information is just easily something I'd want"*. The degree of control over the viewing experience also plays a role. Caption preferences may differ in settings like movie theaters, where viewers have limited control over volume or screen size. The lack of adaptability in such environments can be frustrating for those who rely on captions. V11 recalls feeling left out due to the inadequate accommodations for movie viewers, *"I don't use the captioning machines at the movie theater. So I purely go based on my ears, which sometimes I miss like jokes and stuff"*. V3, who mentioned not liking multiple colors in captions, also mentioned another issue with captioning in theaters, *"I do remember watching a movie in the theater and they did have colors"*.

5.2.2.2 Augmenting the viewing experience: additional sensory elements

Users are open to integrating non-visual sensory elements like lighting, projection, and haptics to enrich their media experience, especially for accessibility. However, such additions must be contextually relevant and intentional to avoid distraction or confusion. V6 noted, *"I think too much sensory input, again, can be confusing or make it harder to process what you're trying to process"*. The use of these elements should serve a clear purpose, such as conveying off-screen information enhancing emotional impact, or supplementing captions to add additional temporal and dynamic NSI (Kushalnagar et al., 2014). V6 also suggested, *"I personally would be more inclined to [want] off-screen information, than like this is a sad scene, so we're going to put a blue backlight"*. The success of such enhancements relies on robust technological integration to create a meaningful multi-sensory experience. V6 continued *"When we're talking about these as accessibility tools, then I'm on the side of making it as accessible [as possible] and [the] intent as clear as possible"*.

5.2.3 Tailoring captions for diverse audiences

5.2.3.1 Adapting captions for audiences of diverse abilities

The effectiveness of captions often hinges on how well they cater to the diverse needs of viewers. For instance, while the pacing of captions may be manageable for those with normal eyesight, it can pose challenges for individuals with visual impairments. V3 highlighted this issue, *"For me, I just have normal eyesight. It's fine with me. But for the DeafBlind, they struggle with that"*. In addition to pacing, the use of colors in captions can be both helpful and problematic, depending on the viewer's visual abilities. V13 pointed out the potential distraction colors might cause, *"I think for people who are colorblind or maybe like low vision, that might be distracting, but I personally don't mind colors [being] used to differentiate [speakers]"*. While the lines between HoH and D/deaf communities can be blurry, a noticeable difference in the frequency of caption use was found. Figure 7A shows that people who self-identify as exclusively D/deaf use closed captions more frequently than those who self-identify as exclusively HoH.

Participants also expressed the importance of considering neurodiversity and aural diversity in captioning preference research. V9 highlighted a tension they observed between the needs of non-DHH people with autism and non-autistic DHH caption users: *"People who are very autistic... tend to be super detailed because it's art to them, but captions are not art, captions are science. They're just communicating facts and reflecting what is said and heard"*. Additionally, a participant[4] who identified "loosely as hard-of-hearing" explained that they have an auditory processing disorder which made it difficult to understand certain spoken words, but that other sounds were fully intelligible. They expressed frustration that this aural diversity is not always accommodated in discussions of DHH culture and related topics.

5.2.3.2 Considering audience age and preferences

Captions should also be tailored to the target audience's age and preferences. For example, children might benefit from more colorful and detailed captions to help them engage with and understand the content. As V5 explained, *"For kids, you need to put a lot of captions because they need to see [captions to] understand what they see. So when you have captions, [children] really get involved in the action and in the movie"*. Additionally, there is a preference among adults to have control over the content they view, particularly in deciding whether or not to censor explicit language in captions if the audio and video are not censored. V9 voiced this sentiment clearly, *"I should not rely on [external] judgment whether I see the bad words or not. I am not a child. I can decide for myself"*.

5.2.3.3 Hearing assistive technology and sign language

The survey responses showed that the type of hearing assistive technology used by DHH folks and their sign language use influences NSI and captioning preferences. For example, Figure 7B shows that 67% of people who sign frequently always use captions when they're available compared to 37% of infrequent signers.

4 The participant requested that specific statements of their hearing status not be connected to other information about them.

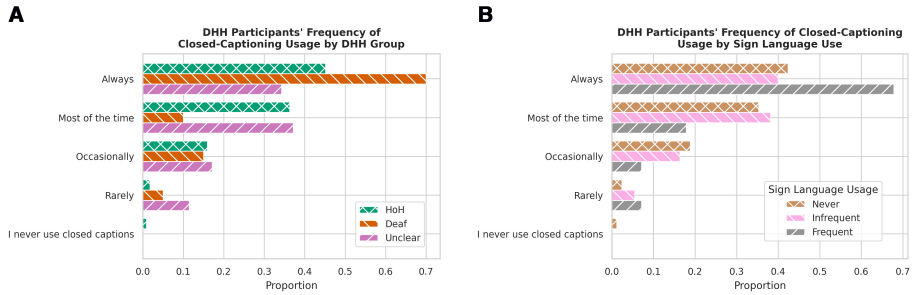

FIGURE 7
Frequency of closed-caption use when closed-captions are available. **(A)** Frequency of closed-caption use by specific hearing status. **(B)** Frequency of closed-caption use by frequency of sign language use.

Similarly, 28% of CI users would prefer every sound effect be captioned compared to 11% of HA users, as shown in Figure 8A. Additionally, CI users were more likely to want speakers identified more frequently than HA users. V13 noted that the effect of hearing assistive technology use itself can modulate their captioning use, *"If [my CI] is behaving, I rely on captions a little less"*.

5.2.3.4 Clustering participant preferences

Analyzing the selection and communication preferences across all four NSI types, as shown in Figures 4, 6, each question has an answer that a plurality of DHH participants selected. Figures 6A, D show that a plurality of participants prefer objective details and descriptions for sound effect and MoS captioning. However, for music, a plurality of survey responders preferred affective details (Figure 6B). For speaker identification, name and alignment were the most highly ranked communication strategies while color ranked consistently low (Figure 6C). Each of these modal answers were selected by at least 7% more participants than the next most selected answers. However, if we were to construct a captioning best practice guide from these modal answers, not a single participant would have all of their preferences met and more than 50% of participants would have less than half of their preferences met, as shown in Figure 9. Therefore, a modal best practice solution would produce a one-size-fits-**none** solution.

Additionally, a Gaussian Mixture Model (GMM) clustering analysis was performed to investigate preference clustering, with the answers to all selection and curation questions used as input. The analysis was performed with cluster numbers ranging from two to ten. The best clustering used four clusters and produced a silhouette score of 0.09, indicating that selection and communication preferences do not form robust, independent clusters. This can be seen, for example, in how participant selection preferences are not consistent across NSI types as 6% of participants would like all NSI types to be captioned always/frequently, 5% would prefer NSI be captioned rarely/never, 32% prefer moderately often, while 57% of participants' NSI selection preferences vary depending on NSI type, as shown in Figure 4. V1 highlighted the importance of offering choices tailored to specific needs, envisioning a system where *"a Deaf person could click on the type of captions they could prefer based on the information that they want... To have choices? That's the dream!"*

5.2.4 Captioning production process

A lack of connection between professional captioners and the creative teams producing video content appeared to negatively impact the ability of captioners to communicate nuanced NSI. Not having access to the authorial intent behind the video work, captioners often rely on guesswork and assumptions, as C4 noted: *"We will beg and plead for the sales rep to try and get a script from the client. It's usually pretty difficult. People, for some reason, can't get it together to give us one."* They expanded on the many stakeholders involved in the captioning process of large productions that can lead to captions being "an afterthought" (C4). The technical complexity of modern media production, with the same video content possibly airing on different TV channels and streaming services, and having smaller clips pulled from the main show for social media, can lead to technical errors and loss of caption files.

5.3 The role of captions

5.3.1 Ensuring equitable access

The importance of high-quality, accurate closed captions in providing full access to video content was communicated by many participants. V10 stated, *"[Captions are] a moral obligation."* This sentiment underscores the ethical responsibility of content providers to make their content accessible to all, viewing captioning not as an optional extra but as a fundamental element of their service. V10 continued, *"People with disabilities are also their viewers,"* highlighting the importance of recognizing and catering to the needs of this audience. The notion of captions as a *contract* between content providers and viewers emphasizes the expectation of reliable and comprehensive access to audiovisual information. However, not all participants agreed with this statement. V11 noted, *"[Captioning companies are] responsible for accurately transcribing the speech that's there, but in terms of like all the sounds in the content, I think that might be a bit much."*

5.3.2 Facilitating understanding and engagement

Beyond providing access to speech, participants expressed that captions are pivotal in facilitating understanding and enhancing their engagement. However, participants also expressed that the line between facilitating understanding and overwhelming views

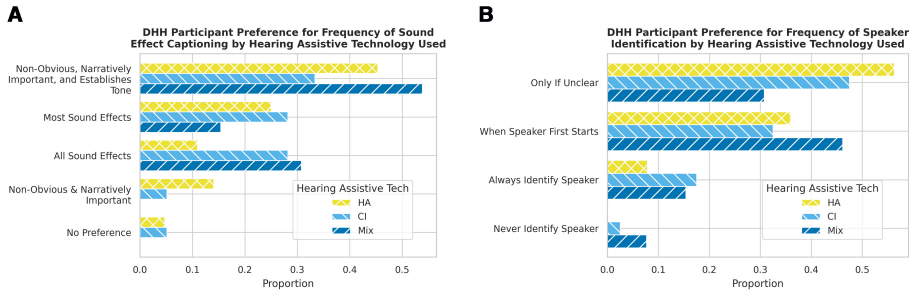

FIGURE 8
Curation preferences of surveyed DHH participants by hearing assistive technology use. **(A)** Sound effect and ambient sound curation preferences by hearing assistive technology use. **(B)** Speaker identification preferences by hearing assistive technology use.

can be quite thin for some viewers. V6 noted, *"[captions are] meant to help you follow the story and not be distracting."* To some viewers, captions must be clear, concise, and seamlessly integrated into the visual narrative, ensuring they aid comprehension without becoming intrusive. V12 highlighted the need for accuracy in captions, stating, *"I'm expecting it to be accurate. Like we're not doing this half-[heartedly], right? It's either nothing, or we're all in because halfway doesn't help."* This statement underscores the significance of reliable and trustworthy NSI captions, contributing to a meaningful and immersive viewing experience. Similarly, information parity was a guiding concern, with C3 noting: *"like they told us in training, translate the hearing experience for the non-hearing viewer... I'd rather give less information than incorrect information."*

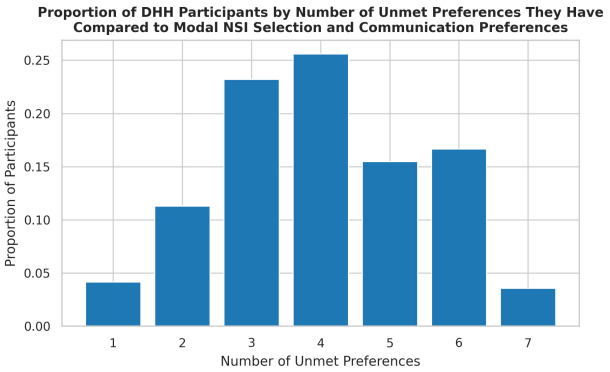

FIGURE 9
Number of preferences of surveyed DHH participants that would go unmet if NSI captions were to be constructed based on the most common answer to all selection and communication questions across the four types of NSI.

5.3.3 Empowering individual preferences

The role of captions extends beyond a one-size-fits-all approach (Arroyo Chavez et al., 2024). Captions should empower viewers to tailor their captioning experience to their needs and preferences. Whether it's adjusting font size and color, controlling the level of detail in NSI, or even incorporating elements like color-coded speaker identification, captions should be customizable to suit a diverse range of viewers. V10 stated, *"I don't need to be spoon-fed,"* highlighting the desire for agency and control over the captioning experience. V1 expressed a wish for features like sound effects to be optional, emphasizing the importance of catering to individual preferences and sensitivities.

Viewers were divided on the factual vs. subjective nature of contextual information communication, with some viewers agreeing that the role of a captioner is akin to a reporter, communicating only factual information in a neutral way. However, others viewed captions as part of the artistic and creative process and noted that they viewed captions as serving the role of co-storyteller, facilitating narrative engagement, and that making reasonable interpretations was appropriate in facilitating this. This formed a factual to subjective spectrum, with many participants falling somewhere in between the two ends. For example, V3 stated that a guiding principle for them was "full access," meaning that they "have access to the same information that an average [hearing] person would." V3 nuanced this by clarifying that they would only want subjective interpretations if an "average" person would agree with the interpretation. If, for example, music was emotionally ambiguous, V3 would prefer only factual descriptions of the music.

5.3.4 Insights into general captioning

Several issues that affect speech captioning, as identified in previous literature, were also found to impact NSI captioning in this study. Participants noted that there was significant variability in the quality of captions across content distribution systems, with real-time content (e.g., live sport) and social video sharing (e.g., *YouTube*), being particularly error prone (May et al., 2024; Xu et al., 2024; McDonnell et al., 2024). Specific issues around captions appearing early or late relative to the video, as well as a lack of clear contrast between the captions and the background, were also noted as factors that negatively impact both NSI and speech captioning. These points are discussed in detail in the Supplementary material.

6 Discussion

Our investigation into non-speech information (NSI) captioning reveals a complex landscape where user needs and professional and research practices intersect. The findings of this

work include the nuanced preferences users express for different types of NSI, the desire among users to control their captioning experience, as well as the patterns in methodologies that have been used to study NSI captioning to date. We highlight opportunities to address this methodological homogeneity, such as the use longitudinal methodologies with more naturalistic stimuli to better evaluate NSI captioning technologies after sufficient participant acclimation. Several themes found across the systematic literature review, online survey, and interview study were identified, such as concerns about cognitive load and information processing in the design and evaluation of NSI captioning systems. In the following section, we expand these themes to provide a comprehensive understanding of the current state of NSI captioning and conclude with recommendations for future research and development to develop NSI captioning systems and tools that may better serve DHH communities.

6.1 Users' selection and communication preferences differ by NSI type

Addressing RQ3, across NSI types, approximately 50% of participants want important, non-obvious, tone/mood-relevant NSI to be communicated, 10% want all NSI communicated, and 10% select only narratively important NSI. However, as shown in Figure 9, if captions are created using the modal choice in each selection and communication question across all four NSI types, then the resultant caption would not meet the needs of nearly any DHH user, with over 50% of users only having half of their NSI selection and communication preferences met. Approximately 10% of users desiring either far more or far less NSI is a non-insignificant amount given the vast number of closed-caption users. Therefore, even though distributions of preferences appear roughly constant across NSI types, specific user preferences are not, indicating a clear need for captions to be customizable with regard to each NSI type.

This finding has additional implications for automated audio captioning, as machine listening systems could be designed to appropriately tag NSI captions by NSI type. This would allow for customization frameworks to leverage these tags and display those NSI captions in whatever way the user requires. Additionally, accuracy of NSI captioning is a consideration in automated systems. As many viewers expressed and C3 summarized, "[captioners would] rather give less information than incorrect information." Therefore, machine listening models prediction confidence or similar metrics could also be leveraged. Examples of this could include directly communicating confidence to the user, or relying on less-detailed captions when confidence is low. "[Music]," while a sub-optimal caption, is perhaps a better caption for a model to produce rather than incorrectly captioning the music's genre. This may be of particular importance as it can be difficult to evaluate the quality of NSI captioning, as exsisting closed caption evaluation frameworks highlight (Liu et al., 2022; Kim et al., 2023a).

We found a variety of factors throughout the analysis that contribute to NSI captioning preferences. Several of these were identified in both the literature review and in the survey/interview analysis, such as the importance of genre considerations in speaker identification (Table 6, G3). They are summarized as follows:

- Social viewing situation (who else is present while watching.)
- Screen size.
- Narrative importance of the NSI.
- Visual redundancy of the NSI.
- Relative information density while the NSI is occurring (i.e. how 'busy' the video is).
- Access to audio (including hearing ability, hearing assistive technology use, and audio fidelity).
- Familiarity with the NSI communication method or technology.

Additionally, several characteristics of NSI captioning were also identified as possible points of modulation, namely:

- The type of descriptive language used (subjective/affective vs factual/objective).
- The length or verbosity of the caption.
- In the case of sound effects, whether the caption is describing the sound itself or the action that caused the sound.

6.2 Desire for increased user agency

The overwhelming desire for greater customization and control over captions underscores the need for a more flexible and user-centric approach to captioning systems. This desire was seen in the survey and interview analysis as well as in previous work (Table 6 G3, Speaker identification—(4) Customization or user options). Participants expressed a strong preference for personalized captions, frustrated by many aspects of the current one-size-fits-none NSI captioning system.

This desire for personalization may stem from fundamentally different interpretations of the role of captioning, with those viewing captioning akin to a journalistic, neutral reporting feeling infantilized or confused by the presence of captions with subjective interpretation. However, those who view captioning as co-storytellers may similarly be confused by the lack of contextual information in factual-only captioning. For example, those who view captioning as a *reporter* might prefer descriptions of music that highlight sonic characteristics, such as "[quiet violin melody]" whereas those who video captioning as a *co-storyteller* may desire additional contextual/functional information, such as "[gentle romantic music plays]." It is important to note that participants did not always strictly fall into only one of these two frameworks but rather existed on a spectrum and noted that factors such as the genre and artistic intention of the video content were important considerations. This demand for customization reflects the diverse needs and preferences of caption users. Factors such as reading speed, visual acuity, and personal preferences can significantly influence the ideal caption format for each individual in addition to their beliefs and preferences regarding the role of captions.

Expanding the Selection, Curation, and Communication framework to create a *"toggle list"* (V12) of desired captioning feature would prove a great step toward greater user agency. Narrative importance and visual redundancy are two aspects of NSI selection that appear crucial to consider, such as allowing viewers

to, for example, select that only narratively important, visually non-obvious sound effects be captioned. By empowering users to control their captioning experience, we can ensure a more inclusive and accessible media landscape. The demand for greater flexibility and customization in caption settings across various platforms and content types remained consistent among our participant population. Embracing user empowerment and personalization in captioning can lead to increased user satisfaction, improved accessibility, and a more tailored and enjoyable viewing experience for everyone.

6.3 Trends in current research

6.3.1 Methodological homogeneity and opportunities

Previous research has primarily utilized user study methodologies, as shown in Table 3, with few using co-design or participatory methodologies and none using longitudinal methodologies. There were no studies that used more naturalistic viewing scenarios such as full movies or episodes. While several studies followed on from previous research (de Lacerda Pataca et al., 2023, 2024; Lee et al., 2007; Vy et al., 2008), we did not find any studies employing direct comparison to previous communication methodologies employed in previous research against their new methods. There is therefore clear methodological opportunities to use methods that are longitudinal, use more naturalistic stimuli (such as full episodes or clips), and use comparative methods.

6.3.2 Common questions, varied solutions

Clear connections of research questions are seen in studies that attempt to communicate the same aspect of the same NSI type, such as emotion in the manner of speech (Fels et al., 2005; Rashid et al., 2006; Lee et al., 2007; Rashid et al., 2008; de Lacerda Pataca et al., 2023). Several communication methods have been well explored by previous literature, such as kinetic typography (de Lacerda Pataca et al., 2024; Kim et al., 2023b). However, while NSI type and communication methods have been widely explored, potential communities that might benefit from these technologies have remained relatively unexplored. For example, a participant we interviewed (who preferred not to have their participant number associated with this statement) disclosed that while they identify as HOH, they have typically functioning hearing organs but have find perceiving some speech to be challenging due to an auditory processing disorder. While aural diversity, an umbrella term referring to all non-normative hearing experiences (Drever and Hugill, 2022), has seen use in other research fields, this more expansive approach has not been actively utilized in the NSI captioning research field. Captioning research has the potential to impact many communities in addition to the DHH community, including the neurodivergent community, the cognitively Disabled community, and English language learners. An example of such involvement could be investigating how people with several of these identities intersecting, such as DHH English language learners, utilize captions.

Therefore, based on this analysis of previous literature, several opportunities for future work are clear:

- Employ participatory, co-design, and show-not-tell methodologies to further include the DHH and other communities of interest.
- Utilize longitudinal methodologies with longer video stimuli to better understand novel NSI communication strategies in more natural viewing situations. This would also allow for greater control of novelty bias and long adaptation times that are possible confounds in much of the existing research (de Lacerda Pataca et al., 2023; May et al., 2023).
- Form stronger comparative connections between previous methods by directly comparing new NSI communication strategies to previous research in user studies.
- Expand research to include other communities of interest while ensuring the needs of the DHH community are always centered.

6.4 Distraction, novelty, and information overload

Reports of users feeling confused, overwhelmed, or distracted from video context by novel SCTs were themes among many papers (Fels et al., 2005; Rashid et al., 2008; Fourney and Fels, 2008; Vy et al., 2008; Mori and Fels, 2009; Kim et al., 2023b). Additionally, visual noise or information density was found to be contextual factor in sound effects and ambient sound captioning (Table 6, G1). In previous work, these feelings were impacted by demographic factors such as hearing status, with DHH users generally expressing more feelings of distraction or information overload (Fels et al., 2005; Rashid et al., 2008; Lee et al., 2007). Our study provided additional evidence to the previous work, with many participants highlighting that communication techniques that require additional attention, such as poetic captions, or add significant novel visual information that the user is not accustomed to, such as a sound visualizer, are particularly prone to be distracting. However, all studies that evaluated novel SCTs used non-longitudinal methods and short clips as stimuli, both factors that could increase feelings of confusion by potentially not allowing enough time for users to acclimate to the novel SCT. Acclimation and previous experience appear to greatly modulate this as many caption users in Europe and the UK prefer speaker labels to be differentiated using text color (Gorman et al., 2021), while this practice is far less common in North America.

Another factor that may, counterintuitively, contribute to distract is the overuse of novel SCTs in evaluations. Many participants indicated preferences only for narratively important NSI to be captioned using as few words as possible. Therefore, this might suggest that the incorporation of parameters such as importance of NSI, obtrusiveness of SCT communication, and degree of novelty would benefit the design and evaluation of novel SCTs. Additionally, the idea that certain NSI captions can themselves be distracting to some viewers further emphasizes the need for flexibility and customization. This flexibility would allow individuals to, for example, choose NSI captioning frequency and

levels of details that best suit their needs and viewing habits, thereby reducing the likelihood of NSI captioning being percieved as overwhelming.

6.5 Design and methodological recommendations

Based on our analysis of the systematic literature review, viewer survey study, and viewer and captioner interviews, we summarize our findings by proposing the following design recommendations for NSI communication systems and research:

1. **Enhance user agency through customizable NSI captions:** Given the diverse range of NSI captioning preferences that were impacted by NSI type, genre, and viewing situation, we recommend that future work investigate the design of a captioning system that can be customized and personalized to better meet a user's specific preferences for NSI captioning for specific content and viewing scenarios. For example, a captioning system could prompt the user to select their preference for the adjectives used when describing music in NSI captioning (e.g., genre information, affective adjectives, etc.).

   The role of automation and AI should be carefully considered in context as these systems are developed. Automated systems developed to "raise the floor," such as systems to automatically generate captions, should be designed and evaluated in ways that reflect the inteded context of their use (e.g., for un-captioned user-generated content on YouTube). Whereas technologies aimed to "raise the ceiling," such as a caption personalization system integrated with a streaming service, should be evaluated within that context. Additionally, particular attention should be paid to differences in user preferences regarding AI authorship vs. AI customization of human-authored content.
2. **Utilize known NSI captioning characteristics for customization:** These new customization options could leverage existing NSI captioning frameworks and include selection features most relevant to the viewer, such as narrative importance and visual redundancy, and communication options, such as choices spanning the spectrum between reporter-style factual details to co-storyteller interpretive information, as meaningful options. These systems should be developed in a way that would accommodate a variety of user interactions ranging from a simple selection of a pre-set group of options based on identified user-clustering, to allowing users to input granular preferences into settings.
3. **Naturalistic, integrated research methodologies:** The field of NSI communication research appears to be in a position to greatly benefit from utilizing study methodologies that better approximate real-world viewing conditions by utilizing longer, more naturalistic video stimuli, longitudinal experimental designs, and direct comparison to relevant previous SCTs. For example, researchers could develop a web extension plug-in that would allow users to use an automated SCT while watching online videos of their choice. This would create an opportunity for more ecologically valid insights into the new SCT.
4. **Broaden and deepen community involvement:** Community-based research methods, such as co-design and participatory design, have shown promise in the field of accessibility research more broadly and may provide fruitful paths forward in a research field inundated with an expansive design space. Additionally, the inclusion of insights from more diverse communities and stakeholders such as second-language learners and those with auditory processing differences/disabilities, in addition to the DHH community, seems likely to increase the impact of future work.

6.6 Limitations

A notable limitation encountered was difficulty in recruiting professional captioners due to industry standard non-disclosure agreements (NDA) generally required by large captioning firms. This resulted in the majority of professional captioners we contacted expressing interest in the study but later withdrawing from the study due to fear of breaching the NDA. Due to similar practices, the captioning software used by many firms appears to be proprietary, making it challenging to incorporate software analysis into our analysis. This study focused solely on research about closed-captioning that was published in English and recruited solely from the United States. Therefore, other languages and cultural contexts may have their own unique challenges and opportunities in NSI captioning that were out of scope of the current study. Our survey and interview study focused primarily on the DHH viewer perspective and did not actively seek to include people from other aurally diverse communities, or other communities of caption users. While we believe that DHH perspectives should always be actively included in closed captioning research, we recommend that future work incorporate perspectives from other caption user communities. Our participants were recruited from the USA and were fluent in written English and spoken English and/or ASL. Therefore, the findings of this paper should not be interpreted as emblematic of other cultural contexts. Finally, our survey and interview questions focused primarily on gathering nuanced insights into previously identified themes in NSI captioning. Therefore, our questions may have introduced a bias against novel themes.

7 Conclusion

We presented the results of (1) a systematic literature review of 36 NSI research papers, (2) a survey study of 168 DHH caption users, and (3) an interview study with 15 DHH caption viewers and 5 professional caption creators. These results provided insights into existing challenges in NSI captioning, such as a lack of agency for users and the diversity of NSI captioning preferences held by a diverse range of DHH caption-users. Previous research and the findings presented in this paper highlight that a one-size-fits-all approach to NSI closed-captioning leads to a sub-optimal viewing experience for most DHH caption users. We recommend that future work include the development of customizable captions that allow for personalization and customization, the utilization of longitudinal and comparative

methodologies in future NSI communication studies, and the inclusion of other communities of caption users in addition to DHH communities.

Data availability statement

The survey data generated by this study is publicly available here: https://zenodo.org/records/15411215.

Ethics statement

The studies involving humans were approved by Stanford University and New York University Institutional Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

LM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. MCl: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. KD: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. KO: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. SS: Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. PW: Data curation, Writing – original draft, Writing – review & editing. MF: Conceptualization, Funding acquisition, Resources, Supervision, Writing – original draft, Writing – review & editing. SL: Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing. MCa: Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. We thank Stanford University's Research, Action, and Impact through Strategic Engagement (RAISE); Music and Audio Research Laboratory (MARL) Seed Award Program for funding this project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomp.2025.1575176/full#supplementary-material

References

3Play Media (2017). "2017 state of captioning report," in *Technical Report*. Boston, MA: 3Play Media. Available online at: https://go.3playmedia.com/soc-2023 (accessed December 5, 2023).

3Play Media (2023). "2023 state of captioning report," in *Technical Report*. Boston, MA: 3Play Media.

Alonzo, O., Shin, H. V., and Li, D. (2022). "Beyond subtitles: captioning and visualizing non-speech sounds to improve accessibility of user-generated videos," in *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility* (New York, NY: Association for Computing Machinery), 1–12. doi: 10.1145/3517428.3544808

Arroyo Chavez, M., Thompson, B., Feanny, M., Alabi, K., Kim, M., Ming, L., et al. (2024). "Customization of closed captions via large language models," in *International Conference on Computers Helping People with Special* Needs (Cham: Springer), 50–58.

Austin, B. A., and Myers, J. W. (1984). *Hearing-Impaired Viewers of Prime-Time Television*.

Barbero, J. M., Bollaín, M., and Santos, E. (2010). "Production and distribution workflow for closed captioning," in *2010 International Conference on Distributed Frameworks for Multimedia Applications* (Jogjakarta: IEEE), 1–6.

Berke, L., Caulfield, C., and Huenerfauth, M. (2017). "Deaf and hard-of-hearing perspectives on imperfect automatic speech recognition for captioning one-on-one meetings," in *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore MA: ACM), 155–164.

Caldwell, B., Cooper, M., Reid, L. G., Vanderheiden, G., Chisholm, W., Slatin, J., et al. (2008). *Web Content Accessibility Guidelines (WCAG) 2.0*. Cambridge, MA: WWW Consortium (W3C), 1–34.

Carroll, C., Booth, A., Leaviss, J., and Rick, J. (2013). "best fit" framework synthesis: refining the method. *BMC Med. Res. Methodol.* 13, 1–16. doi: 10.1186/1471-2288-13-37

Cavender, A., and Ladner, R. E. (2008). "Hearing impairments," in *Web Accessibility*, eds. S. Harper, and Y. Yesilada (London: Springer), 25–35.

Choi, Y., Jeon, J., Lee, C., Noh, Y.-G., and Hong, J.-H. (2024). "A way for deaf and hard of hearing people to enjoy music by exploring and customizing cross-modal music concepts," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu HI: ACM), 1–17.

Climent, M. M., Soler-Vilageliu, O., Vila, I. F., and Langa, S. F. (2021). Vr360 subtitling: Requirements, technology and user experience. *IEEE Access* 9, 2819–2838. doi: 10.1109/ACCESS.2020.3047377

de Lacerda Pataca, C., and Costa, P. D. P. (2023). Hidden bawls, whispers, and yelps: Can text convey the sound of speech, beyond words? *IEEE Trans. Affect. Comput*. 14, 6–16. doi: 10.1109/TAFFC.2022.3174721

de Lacerda Pataca, C., Hassan, S., Tinker, N., Peiris, R. L., and Huenerfauth, M. (2024). "Caption royale: exploring the design space of affective captions from the perspective of deaf and hard-of-hearing individuals," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–17.

de Lacerda Pataca, C., Watkins, M., Peiris, R., Lee, S., and Huenerfauth, M. (2023). "Visualization of speech prosody and emotion in captions: accessibility for deaf and hard-of-hearing users," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–15.

Downey, G. J. (2008). *Closed Captioning: Subtitling, Stenography, and the Digital Convergence of Text With Television*. Baltimore: JHU Press.

Drever, J. L., and Hugill, A. (2022). "Aural diversity: general introduction," in *Aural Diversity* (London: Routledge), 1–12.

Evans, M. K. (2019). *Here's How Automatic Captions Earned Their Nickname*. Mountainview, CA: YouTube. Available online at: https://www.youtube.com/watch?v=N7MfajxyWDY

Fels, D. I., Lee, D. G., Branje, C., and Hornburg, M. (2005). "Emotive captioning and access to television," in *AMCIS 2005 Proceedings* (Atlanta, GA: Association for Information Systems).

Fitzgerald, M., and Jensema, C. (1981). Closed-captioned television viewing preferences. *Am Ann Deaf*. 156, 536–539.

for Standardization, I. O. (2023). *Acoustics — Normal Equal-Loudness-Level Contours*. Geneva, CH: Standard.

Fourney, D., and Fels, D. (2008). ""thanks for pointing that out." making sarcasm accessible for all," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles, CA: SAGE Publications Sage CA), 571–575.

Gorman, B. M., Crabb, M., and Armstrong, M. (2021). "Adaptive subtitles: preferences and trade-offs in real-time media adaption," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–11.

Harrenstien, K. (2009). *Automatic Captions in Youtube*. Mountainview, CA: Google LLC. Available online at: https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

Harrington, R. P., and Vanderheiden, G. C. (2013). "Crowd caption correction (CCC)," in *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue WA: ACM), 1–2.

Hearing Loss Association of America (2023). *Hearing Loss Facts and Statistics*. Chicago, IL: Hearing Loss Association of America. Available online at: https://www.hearingloss.org/wp-content/uploads/2023/09/HLAA_Hearing_Loss_Facts_and_Statistics.pdf

Henry, S. (2022). *Captions/Subtitles*.

Hersh, M. (2013). Deaf people's experiences, attitudes and requirements of contextual subtitles: a two-country survey. *Telecommun. J. Austral*. 63:2. doi: 10.7790/tja.v63i2.406

Holcomb, T. K. (2013). *Introduction to American Deaf Culture*. Oxford: Oxford University Press.

Jain, D., Junuzovic, S., Ofek, E., Sinclair, M., R., Porter, J., et al. (2021). "Towards sound accessibility in virtual reality," in *Proceedings of the 2021 International Conference on Multimodal Interaction* (New York, NY: Association for Computing Machinery), 80–91.

Jensema, C. (1998). *Viewer Reaction to Different Television Captioning Speeds*, 318–324. doi: 10.1353/aad.2012.0073

Jeon, H.-S., Kyung, S.-Y., Lee, S.-J., and Yu, H.-Y. (2024). "Emotional subtitles through speech in films: a case study," in *2024 18th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (Kuala Lumpur: IEEE), 1–5.

Kim, H., Tao, Y., Liu, C., Zhang, Y., and Li, Y. (2023a). "Comparing the impact of professional and automatic closed captions on video-watching experience," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg: ACM), 1–6.

Kim, J., Ahn, S., and Hong, J.-H. (2023b). "Visible nuances: a caption system to visualize paralinguistic speech cues for deaf and hard-of-hearing individuals," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY: Association for Computing Machinery).

Kushalnagar, R. S., Behm, G. W., Stanislow, J. S., and Gupta, V. (2014). "Enhancing caption accessibility through simultaneous multimodal information: visual-tactile captions," in *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, 185–192.

Lasecki, W. S., Miller, C. D., Kushalnagar, R., and Bigham, J. P. (2013). "Legion scribe: Real-time captioning by the non-experts," in *Proceedings of the 10th International Cross-Disciplinary Conference on Web* Accessibility (Rio de Janeiro: ACM), 1–2.

Lee, D. G., Fels, D. I., and Udo, J. P. (2007). Emotive captioning. *Comp. Entertain. (CIE)* 5:11. doi: 10.1145/1281329.1281344

Li, F. M., Lu, C., Lu, Z., Carrington, P., and Truong, K. N. (2022). An exploration of captioning practices and challenges of individual content creators on youtube for people with hearing impairments. *Proc. ACM Human-Comp. Inter*. 6, 1–26. doi: 10.1145/3512922

Liu, X. B., Wang, R., Li, D., Chen, X. A., and Pavel, A. (2022). "CrossA11y: identifying video accessibility issues via cross-modal grounding," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, UIST '22* (New York, NY: Association for Computing Machinery), 1–14.

Martin, D. (2024). Experimental modalities: Crip representation and access with electronic arts intermix. *Leonardo* 57, 209–214. doi: 10.1162/leon_a_02490

May, L., Ohshiro, K., Dang, K., Sridhar, S., Pai, J., Fuentes, M., et al. (2024). "Unspoken sound: identifying trends in non-speech audio captioning on youtube," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–19.

May, L., Park, S. Y., and Berger, J. (2023). "Enhancing non-speech information communicated in closed captioning through critical design," in *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '23* (New York, NY: Association for Computing Machinery), 1–14.

McDonnell, E. J., Eagle, T., Sinlapanuntakul, P., Moon, S. H., Ringland, K. E., Froehlich, J. E., et al. (2024). ""Caption it in an accessible way that is also enjoyable": characterizing user-driven captioning practices on tiktok," in *Proceedings of the CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–16.

McGowan, J., Leplâtre, G., and McGregor, I. (2017). "Cymasense: a real-time 3D cymatics-based sound visualisation tool," in *Proceedings of the 2017 ACM Conference Companion Publication on Designing Interactive Systems* (New York, NY: Association for Computing Machinery), 270–274.

Mei, X., Liu, X., Plumbley, M. D., and Wang, W. (2022). "Automated audio captioning: an overview of recent progress and new challenges," in *EURASIP Journal on Audio, Speech, and Music Processing*. Berlin: Springer Nature.

Mendis, J., Oncy-Avila, R., Vogler, C., and Kushalnagar, R. (2022). "Caption UI/UX - display emotive and paralinguistic information in captions," *The Journal on Technology and Persons with Disabilities*, ed. J. Santiago (Northridge, CA: CSUN Assistive Technology Conference), 125.

Moore, B. C. (2012). *An Introduction to the Psychology of Hearing*. Leiden: Brill.

Mori, J., and Fels, D. I. (2009). "Seeing the music can animated lyrics provide access to the emotional content in music for people who are deaf or hard of hearing?," in *2009 IEEE Toronto International Conference Science and Technology for Humanity (TIC-STH)* (Toronto, ON: IEEE), 951–956.

Naim, I., Gildea, D., Lasecki, W., and Bigham, J. P. (2013). "Text alignment for real-time crowd captioning," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Kerrville, TX: Association for Computational Linguistics), 201–210.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71. doi: 10.31222/osf.io/v7gm2

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning* (New York, NY: PMLR), 28492–28518.

Rashid, R., Aitken, J., and Fels, D. I. (2006). *Expressing Emotions Using Animated Text Captions, volume 4061 of Lecture Notes in Computer Science* (Berlin, Heidelberg: Springer Berlin Heidelberg), 24–31.

Rashid, R., Vy, Q., Hunt, R., and Fels, D. I. (2008). Dancing with words: Using animated text for captioning. *Intl. J. Human-Comp. Interact*. 24, 505–519. doi: 10.1080/10447310802142342

Rogers, K., Hirzle, T., Karaosmanoglu, S., Palomino, P. T., Durmanova, E., Isotani, S., et al. (2024). "An umbrella review of reporting quality in chi systematic reviews: guiding questions and best practices for HCI," in *ACM Transactions on Computer-Human Interaction* (New York, NY: Association for Computing Machinery).

Square Enix (2010). *Final Fantasy XIV: A Realm Reborn*. [*PC Digtial*]. Shinjuku: Square Enix.

Tripp, E. (2023). *Coda Identity-Why Our Stories Are Important: A Qualitative Look at the Personal Narratives of Adult Hearing Children of Deaf Adults*.

Tsaousi, A. (2015). Making sound accessible: The labelling of soundeffects in subtitling for the deaf and hard-ofhearing. *Hermeneus* 17, 233–252.

Udo, J. P., and Fels, D. I. (2010). The rogue poster-children of universal design: closed captioning and audio description. *J. Eng. Design* 21, 207–221. doi: 10.1080/09544820903310691

United States Census Bureau (2021). *American Community Survey*. Suitland-Silver Hill, MD: United States Census Bureau. Available online at: https://www.researchondisability.org/annual-disability-statistics-collection/build-your-own-statistics-state-national-level-statistics

United States Congress (1990). *Television Decoder Circuitry Act of 1990*. Washington, DC: United States Congress. Available online at: https://www.congress.gov/bill/101st-congress/senate-bill/1974

United States Congress (1996). *Telecommunications Act of 1996*.

United States Congress (2012). *47 cfr 79.4 - Closed Captioning of Video Programming Delivered Using Internet Protocol*. Washington, DC: United States Congress. Available online at: https://www.ecfr.gov/current/title-47/chapter-I/subchapter-C/part-79/subpart-A/section-79.4l

Vy, Q. V., and Fels, D. I. (2009). *Using Avatars for Improving Speaker Identification in Captioning, volume 5727 of Lecture Notes in Computer Science* (Berlin, Heidelberg: Springer Berlin Heidelberg), 916–919.

Vy, Q. V., and Fels, D. I. (2010). *Using Placement and Name for Speaker Identification in Captioning, volume 6179 of Lecture Notes in Computer Science* (Springer Berlin Heidelberg, Berlin, Heidelberg), 247–254.

Vy, Q. V., and Fels, D. I. (2011). Enhanced captioning - speaker identification: text vs. images. *Telecommunic. J. Austral*. 61:2. doi: 10.7790/tja.v61i2.209

Vy, Q. V., Mori, J. A., Fourney, D. W., and Fels, D. I. (2008). *EnACT: A Software Tool for Creating Animated Text Captions, volume 5105 of Lecture Notes in Computer Science* (Berlin, Heidelberg: Springer Berlin Heidelberg), 609–616.

Wald, M. (2006). "Captioning for deaf and hard of hearing people by editing automatic speech recognition in real time," in *Computers Helping People with Special Needs*, eds. D. Hutchison, T. Kanade, J. Kittler, J. M. Kleinberg, F. Mattern, J. C. Mitchell, et al. (Berlin, Heidelberg: Springer Berlin Heidelberg), 683–690.

Wang, F., Nagano, H., Kashino, K., and Igarashi, T. (2016). Visualizing video sounds with sound word animation to enrich user experience. *IEEE Trans. Multimedia* 19, 418–429. doi: 10.1109/TMM.2016.2613641

Xu, X., Xie, Z., Wu, M., and Yu, K. (2024). Beyond the Status Quo: a contemporary survey of advances and challenges in audio captioning. *IEEE Trans. Multimedia* 32, 95–112. doi: 10.1109/TASLP.2023.3321968

Zdenek, S. (2011). Which sounds are significant? towards a rhetoric of closed captioning. *Disab. Stud. Quart*. 31:33. doi: 10.18061/dsq.v31i3.1667

Zdenek, S. (2015). "Reading sounds," in *Reading Sounds* (Chicago, IL: University of Chicago Press).

Zdenek, S. (2018). *Logocentrism: The Tendency to Privilege Speech Over Non-Speech in Closed Captioning*.