# Like, Comment & Caption: A Decade of Social Media Video Caption Research (2015–2025)

Huong Nguyen
htn8@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Emma J. McDonnell
ejm249@uw.edu
University of Washington
Seattle, Washington, USA

Lloyd May
lloydmay@stanford.edu
Stanford University
Stanford, California, USA

Alexander Druzenko
adruzenk@epic.com
Epic Systems
Madison, Wisconsin, USA

Zoobia Saifullah Syeda
zs295@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Mark Cartwright
mc232@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Sooyeon Lee
sooyeon.lee@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

## Abstract

As video has become the dominant mode of content on platforms such as YouTube, TikTok, and Instagram, captioning has emerged as a critical factor for accessibility, engagement, and visibility. While prior studies have examined different types of social media video captions or communities' captioning usage, a systematic synthesis has not been undertaken, leading to the risk of proposing interventions that overlook core platform constraints or miss critical accessibility needs. This paper reviews 36 peer-reviewed papers published between 2015 and 2025 across fields such as Human-Computer Interaction (HCI), accessibility, media studies, education, and language learning. We note that captions operate as collective infrastructure co-produced by viewers, creators, and platforms. Deaf and Hard of Hearing (DHH), neurodivergent, and multilingual viewers depend on captions and increasingly expect mechanisms for feedback, while creators face inadequate tool support. Building on these insights, we propose the framework of Participatory Captioning and suggest design implications, highlighting future directions for social media video caption research.

## CCS Concepts

• **Human-centered computing → Accessibility systems and tools**; *HCI design and evaluation methods*; *Empirical studies in HCI*; Social media.

## Keywords

Accessibility; Caption; Deaf and Hard of Hearing; Social media platforms; Video accessibility

**ACM Reference Format:**

Huong Nguyen, Emma J. McDonnell, Lloyd May, Alexander Druzenko, Zoobia Saifullah Syeda, Mark Cartwright, and Sooyeon Lee. 2026. Like, Comment & Caption: A Decade of Social Media Video Caption Research (2015–2025). In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3772318.3791868



*CHI '26, Barcelona, Spain*



ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3791868

## 1 Introduction

During the past decade, video has emerged as the primary mode of communication on social media platforms such as YouTube, TikTok, and Instagram, playing a central role in how creators reach, engage, and retain viewers[32, 122, 150]. Video has accounted for the majority of global internet traffic in 2025, with TikTok exceeding 1.5 billion monthly active users and YouTube reaching 2.5 billion [138, 153]. Within this, social media video captions have become essential, ensuring accessibility while also shaping how videos are discovered, interpreted, and shared [82, 95, 119].

In addition, the rise of mobile technologies has fueled a global creator economy, with over 300 million creators monetizing their work in 2022 and a projected market value of $480 billion by 2027 [1, 39, 43]. In this context, captioning choices are not merely technical but used to strategically shape visibility and engagement [42, 45].

As platforms have grown, so too has the need to design captioning systems that can operate under fast-moving, interaction-heavy, and algorithmically mediated conditions. Because captions on platforms like TikTok, YouTube, and Instagram are generated, edited, displayed, and circulated through socio-technical pipelines, researchers across HCI, accessibility, communication, and media studies have examined different pieces of this design ecosystem. Prior work in HCI has reviewed caption accessibility research broadly [97] or surveyed video-sharing platforms at a high level [13]. Scholars have explored captions as pedagogical tools [46, 52, 109], cultural expressions [33], or legally protected accessibility tools [25, 64, 89, 131, 141].

However, there is little cross-literature understanding of how social media video captioning systems have been designed, evaluated, and theorized. We use *"systems"* in this case to refer to the entire captioning ecosystem: the end-to-end processes of creating, editing, displaying, and interpreting captions, along with the actors (creators, viewers, and platforms), tools, and interfaces that shape how these processes unfold. We refer to *social media video captioning* as the processes through which spoken language, non-speech information, and other communicative cues are transcribed, described,

or visually represented within videos on platforms. By *the design of social media video captioning systems*, we refer to research that develops, evaluates, or critiques tools, algorithms, interfaces, workflows, datasets, or platform features that support caption creation, editing, presentation, accuracy, accessibility, or interpretability. The absence of an integrated, design-oriented synthesis of SMVC is consequential: social media platforms differ fundamentally from traditional caption domains, embedding captions within creator tools, algorithmic feeds, moderation pipelines, and cross-cultural usage norms. Without an understanding of how different communities and disciplines have approached SMVC, researchers risk proposing interventions that overlook core platform constraints or miss critical accessibility needs.

In this paper, we reviewed 36 peer-reviewed articles published between 2015 and 2025 that specifically examine the design of social media video captioning systems. This scoped approach allows us to synthesize how social media video captioning systems are conceptualized within HCI and related fields, identify recurring design challenges, and surface opportunities for future research, infrastructure development, and accessibility-centered innovation.

This review is guided by three research questions (RQ):

> **RQ1:** How have social media video captioning systems evolved across platforms, communities, and caption types over the last decade?
> **RQ2:** How do viewers and creators use, interpret, and engage with social media video captions across platforms?
> **RQ3:** What design and infrastructural gaps remain, and what opportunities do they reveal for future SMVC research?

Through this process, we identify four interrelated themes: (1) Social media video caption types and their use in practice, (2) Viewers' perspectives on social media video captions, (3) Creators' perspectives on social media video captions, and (4) Social media video captioning systems, techniques, and datasets. We found that considerations for social media video captions are not only practical (e.g., automatic vs. manual captioning) but also about how the diverse needs of viewers make access contested, how creators perform everyday infrastructure to make content inclusive, and how current technical systems remain underdeveloped to meet various requirements. This review characterizes current trends in SMVC research, highlighting the viewers and creators who have received disproportionate attention while identifying communities and areas that remain underserved. It then proposes **Participatory Captioning** as a framework for understanding how caption practices are co-produced on social media platforms and for guiding future research and design recommendations.

## 2 Background and Related Work

### 2.1 History and Trends of Video Caption Research

We briefly review the development and study of captions across broadcast television, streaming, and videoconferencing. These eras established the norms, regulatory frameworks, and quality baselines that today's platforms inherit, adapt, or contest.

*2.1.1 Origins in Broadcast Media and The Presentation of Video Caption.* Research on television and live broadcast captions reveals a progression from accuracy to multifaceted examinations of usability, attention, and context. Also, from early readability studies to more recent work on occlusion and personalization, video caption research has consistently considered diverse contexts (e.g., news, sports, education) and viewers (DHH community, second language learners, students). This trajectory forms a critical foundation for understanding caption practices in newer platforms.

Captions first appeared on U.S. television in the early 1970s in an open captioning format, which was permanently visible text burned into the video, giving DHH viewers access to mainstream TV content for the first time [29, 106]. In 1976, the FCC authorized the use of line 21 for transmitting closed captions [101], and by 1980 the National Captioning Institute (NCI) had aired the first captioned broadcasts, introducing captions that could be toggled on and off, a critical shift from open captions [38, 101]. In 1982, NCI introduced the first implementation of real-time captioning at the Academy Awards using a Stenotype system operated by a trained court reporter [29], and in 1996, the Telecommunications Act mandated that all digital television receivers include built-in caption decoders [101]. In other countries like Australia, the government established the Australian Caption Centre in 1982 as a non-profit organization dedicated to promoting and producing captions for DHH Australians [27].

Scientific research began to examine both the effectiveness of captioned television and the impact of its presentation conventions. Studies have explored multiple aspects of video captions, such as two-line display limits [62, 63], consistent bottom-screen placement [86], and caption speed in relation to viewer comprehension, finding that comprehension declines at speeds >145 words per minute [60]. In addition, research also demonstrated that captioned television video supported not only DHH viewers, but also language learners and students in general [61, 70, 93, 103]. Over time, researchers have also examined the accuracy of the television caption. For example, Chavez et al. demonstrated that correlations between Word Error Rate (WER) and user experience are weak: viewer judgments were shaped as much by error salience, synchronization difficulty, and caption presentation as by raw accuracy [11]. Kafle and Huenerfauth proposed alternative metrics to replace WER, shifting the evaluation from transcription accuracy to user comprehension [66].

In addition, researchers have increasingly examined the visual and temporal dimensions of caption design. This body of work emphasizes that caption design is not neutral; the aesthetic and spatial presentation of captions directly shapes legibility, comprehension, and viewer trust. Early enhancement studies underscored the need to move beyond just caption provision toward the systematic design of caption usability. For example, Spina [136] explored video accessibility enhancements for DHH viewers, setting the stage for subsequent user-centered evaluations of caption format and presentation. This line of work framed television as the original testing ground for accessibility metrics. A large body of research has focused on caption appearance preferences, particularly font, size, and positioning. Berke et al. [17] investigated the preferences of DHH viewers for automatic captions, finding that participants valued clarity, accurate punctuation, and readability over raw error rates. In online news contexts, studies found that the position of

the caption significantly shaped the user experience. When captions obscured key visual content, viewers reported distraction and frustration [28].

Beyond stylistic preferences, research has also investigated visual attention and comprehension in live television. Amin et al. found that DHH viewers had to cognitively coordinate between reading captions and scanning visuals [6], and that tolerance for caption occlusion differed by genre, with news and drama requiring unobstructed visuals while entertainment allowed more flexibility [9]. In addition, personalization has been explored as a path forward. A growing body of work has explored dynamic and personalized caption design, with features such as size, contrast, and positioning [7, 9, 31, 55, 93].

*2.1.2 Streaming Era and Expansion to Broader Viewers.* During the era of streaming services, research began to focus on broader viewers, with online fan campaigns around caption quality emerging as early forms of participatory practices that would later shape social media video cultures.

The rise of video-on-demand platforms such as Netflix and Hulu marked a significant shift in captioning practices. Unlike broadcast or cable, which relied on embedded line-21 signals, streaming services adopted file-based caption formats such as Web Video Text Tracks [139, 152] and Timed Text Markup Language [102], allowing captions to scale across devices. Policy interventions also extended accessibility mandates to the digital domain: in the United States, the Twenty-First Century Communications and Video Accessibility Act required captions for online video previously aired with captions on television [38], while the European Accessibility Act (2025) requires broadcasters and streaming platforms to provide captions and audio descriptions to ensure accessibility for disabled viewers [2, 144].

Researchers in HCI and other fields have begun to examine how Netflix closed captions are designed and received. Dizon and Thanyawatpokin [34] found that captions that combine the native language of the learner and the target language improved comprehension and vocabulary acquisition. Fan-driven caption campaigns have long been part of media accessibility advocacy. More specifically, DHH viewers and allies mobilized around caption accuracy and availability in remastered releases of The Wizard of Oz (2009–2012) and later around caption quality issues in Netflix's Queer Eye (2018), urging platforms to correct errors and improve standards [24]. In addition, Netflix viewers also expressed genre-sensitive expectations for captions, evaluating quality not only in terms of accuracy but also in how well video captions aligned with the conventions of specific content types [74].

*2.1.3 Videoconferences and The Rise of User Agency since COVID-19 Pandemic.* The COVID-19 pandemic marked a further shift in captioning norms with the rise of videoconferencing platforms such as Zoom and Microsoft Teams [10, 40]. This differentiates videoconference captions governed by organizations or media and pushes toward a personalized infrastructure. Previous work has examined captioning as a key accessibility and engagement feature in these videoconference systems. It has shown that captions in videoconferences support language learners by improving comprehension [134], and that giving viewers control over caption presentation and prosody improves accessibility [31]. Other work demonstrates that incorporating richer features, such as inclusive design affordances and metadata (e.g., speaker identity, speech rate, background noise), can expand user agency [75, 149]. Captions also play a social role in shaping participation and dynamics in online meetings [98].

## 2.2 The Rise of Social Video Platforms and Trends on SMVC Research

In addition to these ecosystems above, social media platforms have developed rapidly, transforming from niche networking sites into a global ecosystem of platforms that structure daily communication and cultural production [67]. In scholarly terms, social media platforms are defined as *'the set of interactive Internet applications that facilitate (collaborative or individual) creation, curation, and sharing of user-generated content'* [30]. Beyond text and image-based platforms like Facebook, Instagram or X (Twitter), video-centric environments such as YouTube, Twitch, and TikTok are also classified as social media platforms, since they allow viewers to connect, learn and share through interactive user-generated content [22]. Unlike traditional television, social media fosters a participatory culture in which everyone can create and interact with content [41, 59].

With the widespread adoption of mobile technologies and global Internet access, anyone, often equipped with only a smartphone, can now produce and upload video content anytime and anywhere [39, 59]. Within this environment, the creator economy has become a significant site of labor and value generation: in 2022, it encompassed more than 300 million creators in nine countries, with more than half monetizing their work [1], and by 2027, its market value is projected to be approaching $480 billion [43]. Today, creators also cultivate personal branding and aesthetic identity through deliberate choices in design and presentation [99, 147]. In such contexts, every creative and technical decision, including whether and how to caption social media videos, becomes strategically significant [42, 45].

Video has emerged as the dominant medium on social platforms, valued by creators for its expressive capabilities and ability to drive engagement [32, 122, 150]. While captions were once conceived primarily as an accessibility mandate for DHH communities [29, 106], in contemporary social media, they now serve multiple functions, such as expanding access, enhancing search engine visibility, and sustaining viewer attention [14, 82, 95, 119]. Also, much of the online video today is consumed in silence. On Facebook, an estimated 85% of videos are viewed without sound, while captioned videos generally sustain higher engagement, with viewers being 80% more likely to watch an entire video when captions are present [48, 110].

*In this paper, we use the term* ***Social Media Video Captions*** *(SMVC) to refer to any textual or symbolic elements that are temporally aligned with video and serve to support speech comprehension, meaning-making, or interpretive access. This includes platform-generated captions, creator-edited or creator-designed captions, and captions produced by professional services. Our scope intentionally includes a wide range of expressive forms such as words, punctuation, onomatopoeia, emojis, and other visual markers in the text of captions when they are used to represent spoken language or non-speech information, or when they function to help viewers interpret tone, emotion, or narrative cues. We exclude decorative or purely aesthetic text that*

*does not contribute to comprehension, such as branding overlays, creator handles, stickers unrelated to the audio, watermarks, or other graphic elements that are not intended to convey speech, sound, or interpretive meaning. We acknowledge that the boundary between SMVC and other forms of on-screen text can be blurry in practice, such as "vibe text," keyword emphasis, or emoji-based captioning. In such cases, we treat these elements as video captioning practices when they play an interpretive, communicative, or accessibility-supporting role. Broadcast media uploaded to social platforms (e.g., CNN or BBC on YouTube) are also included within this scope.*

Building on the trajectories of television, streaming platforms, and videoconferences along with the rise of social media video, researchers have explored different aspects of SMVC. For example, in the education field, Parton et al. explored whether automatic captioning on YouTube met Deaf students' needs [109], while Hernandez et al. assessed the closed caption quality of YouTube videos dedicated to English language teaching and learning [52]. In business and marketing field, researchers found that the insertion of an English caption influenced the number of views on YouTube videos [50]. In the psychology field, Yang et al. found that enhanced captioning positively affected the motivation of intermediate to advanced German learners by supporting efficient processing and preserving linguistic integrity, allowing more effective incidental attention [154]. In HCI, researchers have examined captioning practices from multiple perspectives, such as captions on TikTok [35, 96, 131], captions on YouTube or other platforms [25, 26, 141].

Although SMVC research is expanding, there has been little systematic review for the social media video captioning systems in the field. This gap limits visibility into platform-specific captioning dynamics, creates inconsistencies in how captioning systems are designed and evaluated, and contributes to design and policy interventions that do not fully address accessibility needs. To address this, we identify common methods and themes, highlight gaps, and offer future directions for researchers, as well as insights for creators, platform designers, and accessibility advocates.

## 3 Methodology

We used a two-stage Google Scholar search strategy covering SMVC work from 2015 to 2025. Phase 1 was exploratory: we mapped the breadth of captioning scholarship, refined our keyword taxonomy (platform, accessibility, and caption terms), and surfaced interdisciplinary work across HCI, media studies, education, and disability studies. This phase established the conceptual landscape and stabilized the search terms. Phase 2 applied the finalized taxonomy via SerpAPI but restricted results to the top 20 HCI venues (Table 1). Whereas Phase 1 cast a wide multidisciplinary net, Phase 2 served as a venue-specific sweep to confirm completeness within HCI and reduce the chance of missing papers embedded in specialized proceedings. Using the same taxonomy across both stages ensured comparability and minimized coverage gaps. We focused Phase 2 on HCI venues because HCI remains a primary intellectual home for research on system design, accessibility, multimodal interaction, and caption-related communication practices, which are the main focus of our literature review [16, 77, 151].

| # | Venue / Journal | Publisher |
|---|---|---|
| 1 | ACM CHI Conference on Human Factors in Computing Systems | ACM |
| 2 | Proceedings of the ACM on Human-Computer Interaction | ACM |
| 3 | International Journal of Human-Computer Interaction | Taylor & Francis |
| 4 | Behaviour & Information Technology | Taylor & Francis |
| 5 | IEEE Transactions on Affective Computing | IEEE |
| 6 | International Journal of Human-Computer Studies | Elsevier |
| 7 | Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies | ACM |
| 8 | Virtual Reality | Springer |
| 9 | International Journal of Interactive Mobile Technologies | Kassel University Press (open access) |
| 10 | ACM Transactions on Computer-Human Interaction | ACM |
| 11 | ACM Designing Interactive Systems Conference | ACM |
| 12 | ACM Symposium on User Interface Software and Technology | ACM |
| 13 | ACM/IEEE International Conference on Human Robot Interaction | ACM / IEEE |
| 14 | Frontiers in Virtual Reality | Frontiers |
| 15 | IEEE Virtual Reality Conference | IEEE |
| 16 | International Conference on Intelligent User Interfaces (IUI) | ACM |
| 17 | Universal Access in the Information Society | Springer |
| 18 | IEEE Transactions on Human-Machine Systems | IEEE |
| 19 | HCI International | Springer |
| 20 | Multimodal Technologies and Interaction | MDPI |

**Table 1: The list of proceedings and journals for paper retrieval was derived from the top 20 HCI-related venues ranked on Google Scholar as of May 2025**

.

### 3.1 Identification Process

To identify papers for our review, we first systematically selected data sources and search strings.

*3.1.1 Source Selection.* We conducted both manual (Phase 1) and programmatic (Phase 2) searches using Google Scholar. We chose Google Scholar as our meta-search engine for two reasons: (i) it has broad coverage across computing, communication, accessibility, and interdisciplinary social science literature, and (ii) it has high intersection with major digital libraries such as ACM Digital Library, IEEE Xplore, SpringerLink, and Scopus. Prior research has shown that Google Scholar retrieves the largest proportion of overlapping

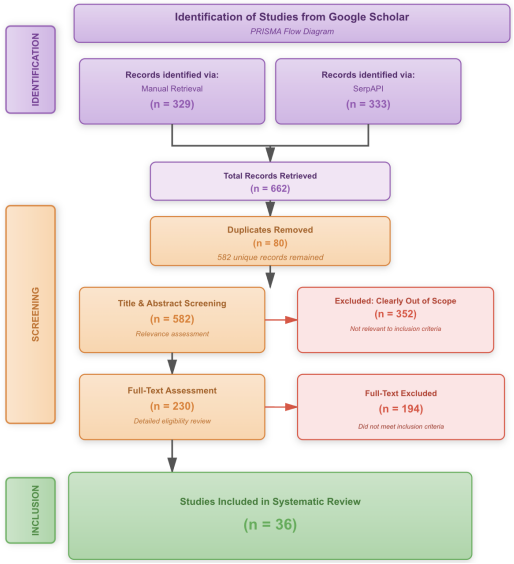


**Figure 1: PRISMA flow diagram showing the identification, screening, and inclusion process for the systematic review**

records compared to Web of Science and Scopus, making it suitable for multidisciplinary topics such as accessibility and SMVC [94].

*3.1.2 Search String Formulation.* To construct our search queries, we developed keyword groups aligned with the scope of this review: (i) platform-related terms, (ii) accessibility and captioning-related terms, (iii) caption-specific terminology, and (iv) user and stakeholder terms. We additionally identified a set of excluded terms to avoid retrieving studies unrelated to social-media video captioning (e.g., image captioning, classroom transcription, or translation studies).

We formulated a set of search terms aligned with our scope on video captioning in social media environments. Platform-related terms (e.g., *TikTok, YouTube, Instagram Reels, Facebook Watch*) were included to capture both global and regionally specific video-sharing ecosystems. Accessibility-related terms (e.g., *accessibility, disability, DHH, deaf, neurodivergent, language learners*) ensured coverage of literature focused on inclusion and access. Caption-specific terms (e.g., *caption, video caption, automatic caption, user-generated caption, non-speech information, sign-language captioning*) targeted work addressing captioning practices and techniques. We excluded terms such as *Zoom, online lecture, transcripts, image caption* to avoid retrieving videoconferencing- or classroom-based research, which is outside our focus.

Because our approach centers specifically on captioning and caption-related accessibility, it may overlook broader platform research where captions play a role but are not explicitly discussed. Together, Phase 1 returned 329 results.

In our second stage we used a Python script with SerpAPI to replicate the refined keyword structure from Phase 1, with systematic focus on HCI's top 20 venues ranked on Google Scholar (see Table 1). This phase yielded 333 results. This phase targeted the 20 HCI venues previously identified in Table 1. To minimize noise, the SerpAPI script applied the same exclusion filters as in Phase 1 and followed the same publication date restriction. For each entry, the script extracted metadata, including title, venue, publication year, and source URL.

## 3.2 Inclusion and Exclusion Criteria

We now explicate the inclusion and exclusion criteria we used to screen candidate results for analysis in our review. We designed the inclusion and exclusion criteria based on the main purpose of the paper, which is directly relevant to the design of the social media video captioning systems. Because captioning on social media emerges from the interplay of creators, viewers, and platform interfaces, we included studies that examined these environments and the caption-related activities of creators and viewers within them. To ensure our corpus focused on this, we restricted it to empirical studies, system-building work, and papers focusing on using captioned social media videos as central data.

*3.2.1 Inclusion Criteria.* A paper was included in our final analysis if:

**(1)** It talked about the design of social media video captioning systems.
**(2)** It explicitly examined social media platforms or employed simulated social media video environments (e.g., YouTube-like or TikTok-like interfaces) as part of the study context[1]
**(3)** It presented empirical or design contributions, such as user interviews, analyzes of captioned video content, computational captioning models, or artifacts demonstrating captioning tools and systems within social media contexts.
**(4)** It incorporated captioned social media videos as datasets for tasks such as error detection or accessibility system development, reflecting how platform-generated caption data shapes SMVC technologies and design practices.

*3.2.2 Exclusion Criteria.* Papers were excluded from our final analysis if they fell into one of the following categories:

**(1)** Publication type exclusions: We excluded short-form or non-archival contributions such as posters, demos, extended abstracts, and workshop-only publications, as well as editorials, commentary pieces, and theoretical essays without empirical data.
**(2)** Language exclusions: Papers not written in English were excluded unless a complete and verified translation was available.
**(3)** Topical exclusions (caption relevance): Studies that mentioned captions only in a general sense, such as noting their benefits for comprehension or visibility without examining social media video captioning systems, were excluded.
**(4)** Context exclusions: We excluded studies in the context focusing only on television, podcasts, surveillance footage (e.g., CCTV), streaming services, and videoconference platforms.

[1]Simulated social media environments are considered a valid form of social media setting in academic research. These systems reproduce core platform features—including interface design, recommendation mechanisms, interaction structures, and user behavioral patterns—enabling controlled investigation of online behavior, misinformation dynamics, and digital literacy. Amid increasingly restricted access to real platform data, simulated environments have become a widely accepted methodological alternative for generating behavioral evidence, offering both causal inference and ethical feasibility that real-world platforms cannot reliably support [57, 85].

## 3.3 Final Corpus

We documented our screening and selection process using a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)-style flow diagram, adapted for the scope of this review (see Figure 1). In two phases, our search retrieved 662 records. After removing 80 duplicates, 582 unique papers remained. We conducted title and abstract screening on these 582 papers, excluding 352 papers that were clearly out of scope. This left 230 papers for full-text review. During full-text screening, we assessed each paper's relevance to our inclusion criteria. Of the 230 full texts, 194 were excluded, spanning a wide range of disciplines. We excluded 43 papers from HCI, 26 from Marketing and Communication, 28 from Social Media Studies, 15 from Education, 23 from Linguistics and Language Studies, 23 from Artificial Intelligence (AI)/Machine Learning, 10 from Health and Medicine, 14 from Psychology, and 12 from Disability Studies. After these exclusions, 36 papers remained in the final corpus (see Appendix A).

## 3.4 Analysis Approach

*3.4.1 Coding Procedure.* We conducted iterative qualitative coding following established grounded-theory procedures [151]. This involved (1) **Open coding**, where we generated initial labels to capture emerging concepts across the corpus; (2) **Axial coding**, where we organized related codes into categories and linked study features to broader conceptual groupings; and (3) **Selective coding**, where we integrated these categories into a coherent thematic structure that described the landscape of SMVC research. Two researchers collaboratively coded all 36 papers.

Using affinity diagramming in Miro, we iteratively clustered related findings into small groups with Miro cards, discussed overlaps and distinctions across papers, and gradually linked these groups to emerging subcategories. We began with **open coding**, working individually to extract key excerpts from each paper's results, findings, and discussion sections. Each card was compared, regrouped, and refined over multiple sessions to ensure conceptual alignment. To structure this stage, we used the paper's subsections and paragraph boundaries as natural units of meaning. As a result, a single findings section could receive multiple codes when it contained several analytically distinct points. Each open-code card included either the section header or a concise excerpt from the paper to preserve contextual clarity. We then moved into **axial coding**, gathering all open-code cards and reflecting on one another's interpretations. Finally, we conducted **selective coding**, revisiting the full corpus and all card clusters to sharpen category definitions and verify consistent code assignments.

*3.4.2 Coder Process and Reliability.* In the initial round, two researchers independently coded a subset of 15 papers and then compared our interpretations to identify points of divergence. Any discrepancies were resolved through a negotiated-consensus process, which led to iterative refinements of the coding schema and sharper category boundaries, following guidelines for intercoder reliability in qualitative research [90]. After achieving stable agreement, we applied the refined coding schema to the full corpus. Throughout this stage, we met weekly with the whole author team to review analysis choices, discuss emerging themes, and ensure consistency and coherence across the developing thematic structure.

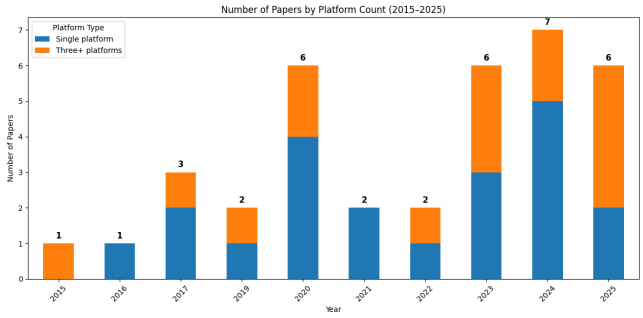


**Figure 2: Platform coverage across years, showing a shift from predominantly single-platform studies toward increasing multi-platform ones in social media video caption research**

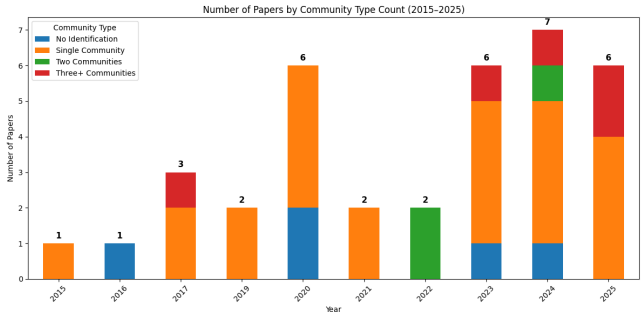


**Figure 3: Community diversity across years, highlighting growing inclusion of multiple participant communities in social media video caption research**

# 4 Summary of Paper Corpus

## 4.1 Publication Trends

Our corpus includes 36 papers published between 2015 and 2025. As shown in Figure 2, the field started quietly, with only one to three papers per year from 2015 to 2017. Activity did not pick up until after 2019, when interest in SMVC began to accelerate, reaching six papers in 2020. Since then, growth has been steady, with six publications in 2023, seven in 2024, and already six in 2025 as of May.

## 4.2 Publication Venues

In addition, we examined the distribution of the venues in our review pool, which can be seen in Table 2. Of the 36 papers, publication activity was highly uneven across venues. *Conference on Human Factors in Computing Systems* accounted for the largest share with 13 papers, while Computer-Supported Cooperative Work followed with 4. The remaining papers were dispersed across many venues, each appearing once, such as Web for All Conference or Universal Access in HCI. This dispersion across HCI & Accessibility venues, Education & Language Learning journals (e.g., Language Learning & Technology, Computer Assisted Language Learning), Media & Society outlets (e.g., Social Media + Society, Television

| Primary Field | Publication Venue (Journal/Conference) | # |
|---|---|---|
| **HCI & Accessibility** | Conference on Human Factors in Computing Systems [5, 25, 26, 33, 54, 64, 68, 69, 95, 96, 127, 141, 155] | 13 |
| | Computer-Supported Cooperative Work [81, 89, 131, 156] | 4 |
| | ACM SIGACCESS Conference on Computers and Accessibility [128] | 1 |
| | Web for All Conference [17] | 1 |
| | International Conference on Information Technology for Social Good [108] | 1 |
| | Universal Access in Human–Computer Interaction [8] | 1 |
| | Universal Access in the Information Society [113] | 1 |
| **Education & Language Learning** | Language Learning & Technology [4] | 1 |
| | Computer Assisted Language Learning [143] | 1 |
| | Online Learning [133] | 1 |
| | Journal of Open, Flexible and Distance Learning [109] | 1 |
| | The Journal of Asia TEFL [79] | 1 |
| | Journal of English and Applied Linguistics [52] | 1 |
| | Complutense Journal of English Studies [145] | 1 |
| **Media & Society** | Social Media + Society [118] | 1 |
| | Journal of Audiovisual Translation [35] | 1 |
| | Television & New Media [23] | 1 |
| | Journal of Global Literacies, Technologies, and Emerging Pedagogies [46] | 1 |
| | The European Proceedings of Social and Behavioural Sciences [71] | 1 |
| | Journal of Social Sciences [112] | 1 |
| **Other** | Journal of Advertising Research [116] | 1 |
| **Total** | | **36** |

**Table 2: Breakdown of 36 papers across three primary fields, with counts and interdisciplinary venue (N=36)**

& New Media), as well as other applied venues, highlights the interdisciplinary character of SMVC research, spanning computing, accessibility, education, media studies, language learning, and the social sciences.

# 5 Analysis and Findings

## 5.1 The Rise of Diversity Across Platforms and Communities

We analyzed our included papers through two dimensions: *platform coverage* and *community type*. As illustrated in Figure 2, most studies concentrated on a *single platform*, while a smaller subset examined *three or more platforms* within the same design; we did not observe a sustained pattern of two-platform studies. For community type, we categorized studies as focusing on a *single community*, *two communities*, or *three or more groups*, depending on how many distinct participant groups were explicitly included. We also had *No Identification*, for papers that did not specify any user group, commonly technical evaluations of auto-captioning or dataset papers with no defined participant group [79, 95].

Overall, **platform and community diversity** has increased steadily over the past decade. Early years (2015-2019) show predominantly single-platform studies with occasional multi-platform papers. Mid-period (2020-2022) shows moderate activity with mostly a single-platform focus. Recent years (2023-2025) demonstrate a notable increase in both total papers and the proportion of multi-platform studies. This shift reflects two broader trends: (1) users' growing engagement across multiple platforms simultaneously [36], and (2) the rapid expansion of short-form video ecosystems such as TikTok, Douyin, [2] Bilibili[3], Kuaishou[4], and RedNote. [5] (see Figure 2) [25, 26, 96, 140, 141]. A similar pattern emerges: early work predominantly centered on DHH communities or other single-community samples, whereas recent studies increasingly involve neurodivergent users, language learners, or mixed-community groups. This shift reflects a growing recognition that SMVC functions within a diverse ecosystem of audiences rather than serving a single community (see Figure 3 for trends in community type and Table 4 for detailed community type).

## 5.2 Overview of the Research Landscape

As SMVC research spans across different communities, platforms and studies, we closely examine the research landscape, including understanding how communities, platforms, and methodological traditions intersect.

*5.2.1 Analysis.* We coded each paper on six dimensions: (1) *Stakeholder (Creator/Viewer): Creator* refers to individuals who produce and upload videos on social media platforms, regardless of whether the videos are captioned, while *Viewer* refers to individuals who watch and interact with these videos. (2) *Community Type:* the accessibility-relevant population placed as beneficiaries of the research (e.g. DHH, neurodivergent groups, language learners), with "Other disabilities" covering additional disability groups, "Public"

[2]Douyin is the Chinese counterpart of TikTok.
[3]Bilibili is a Chinese video-sharing platform known for animation, fandom communities, and dense on-screen comment cultures.
[4]Kuaishou is a major Chinese short-video and livestreaming platform popular among diverse user groups.
[5]RedNote (Xiaohongshu) is a Chinese social media and e-commerce hybrid platform.

| Category | Codes |
|---|---|
| Stakeholder | Creators, viewers |
| Community Type | DHH, Neurodivergent groups, Language learners, Other disabilities, No identification, Public |
| Country of Residence | USA, Canada, Saudi Arabia, Poland, Australia, China, Cyprus, South Korea |
| Study Setting | In person, Online |
| Data Collection Method | Semi-structured interviews, Social media video data analysis, Surveys, Focus groups |
| Contribution Type | Empirical, Dataset, Artifact |

**Table 3: Summary of coding categories, including Stakeholder, Community Type, Country of Residence, Study Setting, Data Collection Method, and Contribution Type**

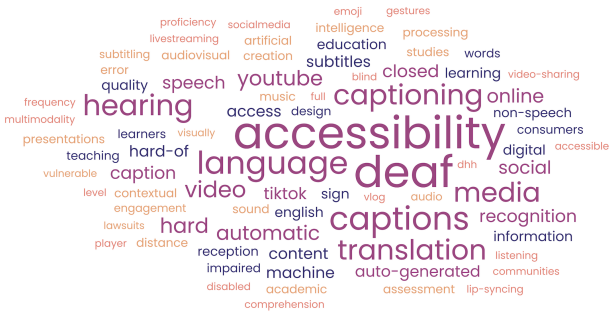


**Figure 4: Word cloud of key terms across 36 papers**

covering general public explicitly identified by the authors. (3) *Country of Residence:* the geographic location of the participants as reported in the study, used to situate the findings in cultural, linguistic and regional contexts. (4) *Study Setting:* the mode of execution of the user study (e.g., in person, online) (5) *Data Collection Method:* techniques used to collect data (e.g., interviews, surveys, diary study). For data collection methods, we report those used in at least two papers (6) *Contribution Type:* the knowledge contributions following Wobbrock et al. [151], with categories identified in our pool including Empirical, Dataset, and Artifact. More details can be seen in Table 3.

*5.2.2 Findings.* Across 36 papers, SMVC research has centered on the viewer experience, with nearly three-quarters prioritizing caption viewers over creators (Table 4). DHH participants were the most frequently studied community, alongside work involving language learners and general-population samples, while only a small number of studies focused on neurodivergent users or other disability groups. Most participant studies were conducted online, reflecting both the global distribution of platforms and the accessibility of remote methods. Semi-structured interviews dominated data collection, underscoring the field's emphasis on lived experience, while video-content analysis and surveys appeared less consistently. Empirical contributions drove the majority of papers,


**Figure 5: Four key research themes of the literature review**

with only a few introducing artifacts or datasets. Geographic representation remained narrow: U.S. and China-based participants accounted for most studies, with occasional work from other countries represented only once. These patterns indicate that although SMVC research spans diverse methodological approaches, it remains grounded in a small set of communities, countries, and empirical studies, revealing future research opportunities. (see Table 5).

## 5.3 Key Research Themes

Our synthesis revealed four key themes that map the contours of SMVC research, as can be seen in Figure 5. First, studies have examined **SMVC Types and Their Use in Practices**-automatic, user-generated, non-speech information, and captions for sign language, each raising tensions between efficiency, creativity, accessibility, and technical limitations. Second, research foregrounds **Viewers' Perspectives on SMVC**, highlighting not only interaction patterns and benefits but also recurring challenges and evolving expectations among viewers. Third, **Creators' Perspectives on SMVC** analyzes captioning as both a site of motivation and a labor-intensive practice constrained by platform tools and infrastructures. Finally, **SMVC: systems, techniques, and datasets** illustrates efforts to develop and evaluate systems, techniques, and datasets for captioning support.

*5.3.1 SMVC Types and Their Use in Practice.*
Throughout our coding process, we identified four types of SMVC investigated in papers: (1) automatic captions, (2) user-generated captions, (3) non-speech information, and (4) captions for sign language.

**Automatic Captions - Efficiency at the Cost of Accuracy**
Research on automatic and auto-translated captions consistently reveals systemic shortcomings in social media's automatic captions, from recurrent errors in recognition and translation to broader usability failures in educational and cross-cultural contexts. Parton et al. [109] analyzed the errors in automatic captions for educational videos on YouTube. Their study highlighted how seemingly

| Stakeholder | | Community Type | | Country of Residence | |
|---|---|---|---|---|---|
| **Category** | **Number** | **Category** | **Number** | **Country** | **Number** |
| Creators [5, 25, 81, 113, 118, 131, 141, 145, 155] | *9* | DHH [5, 8, 17, 23, 25, 26, 33, 54, 69, 81, 89, 96, 108, 116, 118, 127, 128, 141, 155] | *19* | United States [8, 23, 33, 89, 96, 108, 127, 128, 133] | *9* |
| Viewers [4, 5, 8, 17, 23, 33, 35, 46, 54, 64, 68, 69, 81, 89, 96, 108, 116, 127, 128, 133, 143, 155, 156] | *23* | No identification [71, 79, 95, 109, 112] | *5* | China [25, 141, 143] | *3* |
| | | Language learners [4, 33, 46, 52, 54, 143] | *6* | Other — one each (Canada [33], Saudi Arabia [4], Poland [35], Australia [46], Cyprus [108], South Korea [69]) | *6* |
| | | Public [35, 68, 113, 118, 133, 145] | *6* | | |
| | | Neurodivergent groups [33, 64, 96, 131] | *4* | | |
| | | Other disabilities [33, 64, 96, 116] | *4* | | |

**Table 4: Summary of Stakeholder, Community Type, and Country of Residence**

minor errors could lead to major meaning distortions. Similarly, Lee et al. [79] conducted an analysis of YouTube's automatic caption in videos of well-known speakers with varied accents and backgrounds. Their work proposed a ten-category taxonomy of caption errors and also classified caption errors based on function words versus content words, providing a lens for understanding where and how automatic caption systems fail. Martin et al. [113] investigated student-produced captions and compared them with those created by a human expert and YouTube's automatic caption system. These articles attributed the recurring problems to key system-level failings, including accuracy limitations, lack of speaker identification, and poor formatting quality [79, 113].

Several studies examined practices with automatically translated captions on social media, particularly YouTube. Prasetio and Wahyuningsih [112] analyzed YouTube's autotranslation of movie trailers into Indonesian, identifying lexical errors, disambiguation, word order, and compound words that reflect deeper structural mismatches between English and Indonesian. Problems with automatic caption translation were also examined in specific content domains. Hernandez et al. [52] studied auto-translate caption quality in English-language teaching videos on YouTube by Filipino creators. Veroz-González and Bernal [145] evaluated the usability of captions automatically translated from English to Spanish at academic events. These papers concluded that YouTube's auto-translate system is consistently limited by narrow training datasets, leading to cultural mismatches and reduced caption quality and accessibility [52, 112, 145].

In the educational domain, researchers have foregrounded automatic and auto-translated captions on social media, analyzing their quality and usability [52, 109, 145] through different frameworks, such as the Multidimensional Quality Metrics and Universal Design for Learning standards. Their findings consistently showed that YouTube auto-translated captions fell short of educational standards, required extensive post-editing, and required significant revision before being made available to students and language learners in general. One study also raised concerns about whether auto-translate captions on YouTube could ever meet the requirements for DHH students required by educational accessibility guidelines [109].

**User-generated Captions - Creativity Without Standards**

User-generated captions are typically implemented as open captions, embedded directly into the video and manually added by creators, who control their textual styling and spatial placement [96]. Studies of user-generated captions on social media show a landscape marked by inconsistency, stylistic experimentation, and recurring accessibility gaps. Even when caption presence appears high, quality can still be uneven. McDonnell et al. [96] reported that although all sampled videos included captions, only 72.7% applied them consistently across audio sources, yet this consistency did not necessarily translate into completeness or accessibility. Cao et al. [26] reported missing or incomplete captions in videos uploaded by deaf creators, while Tang et al. [141] observed that half of the videos by deaf creators on Kuaishou and WeChat lacked captions entirely.

Beyond completeness, researchers have documented frequent errors in user-generated captions. Li et al. [81] noted spelling, grammar, and punctuation issues in both automatic and human-created captions on YouTube. Duraj & Szarkowska [35] found that TikTok captions often deviated from linguistic norms (e.g., minimal grammar or sentence structure), shaping how viewers interpreted videos. Cao et al. [26] additionally identified overlapping text, missing captions during speech or signing, and intentional omissions that simulated inaccessibility. Research shows a clear tension between creative expression and accessibility in user-generated captions. Prior work [35, 96] found that TikTok lacks consistent captioning standards, raising questions about whether this issue appears across other platforms. Duraj and Szarkowska [35] argued that TikTok captions function less like traditional captions and more like a hybrid social-media writing style. Creators often use "textese" or "netspeak"—all lowercase, alternating caps, or playful typography—to signal mood, humor, or generational identity. McDonnell et al. [96] similarly noted that, unlike the formal captioning industry, TikTok caption practices remain highly unregulated and stylistically open-ended.

In addition, research also reported problems of text overlap between the caption and other elements in terms of the video user interface across social media channels [8, 26, 96]. For example, Cao

| Study Setting | | Data Collection Method | | Contribution Type | |
|---|---|---|---|---|---|
| **Category** | **Number** | **Category** | **Number** | **Category** | **Number** |
| In person [4, 8, 17, 23, 54, 113, 133, 143] | *8* | Semi-structured interviews [5, 8, 17, 25, 33, 46, 64, 69, 81, 89, 98, 116, 127, 128, 131, 141, 155] | *17* | Empirical (see Appendix A) | *36* |
| Online [5, 8, 25, 33, 64, 68, 69, 81, 89, 96, 116, 127, 128, 131, 141, 155] | *16* | Surveys [17, 35, 46, 81, 89, 108, 145] | *7* | Dataset [69, 95] | *2* |
| | | Social media video data analysis [26, 52, 79, 81, 95, 96, 109, 112, 118, 141] | *10* | Artifact [5, 54, 69, 108] | *4* |
| | | Focus groups [23, 46, 116] | *3* | | |

**Table 5: Summary of Study Setting, Data Collection Method, and Contribution Type**

et al. [26] observed that in some cases multiple captions appeared simultaneously, which reduced the clarity of the videos. McDonnell et al. [96] similarly reported that user-generated captions often conflicted with TikTok's dense user interface elements, while Amin et al. [8] found that caption occlusion from YouTube videos, such as covering faces or text-could lead to negative user responses.

**Non-Speech Information - The Missing Layer of Sound Access**

Researchers have increasingly examined non-speech information on social media platforms. May et al. [95] showed that non-speech information captions commonly include language markers, sound effects, paralanguage, speaker or manner identifiers, and musical or channel cues, often bracketed. Their analysis of popular and studio-produced YouTube videos found that although most captions contained some non-speech information, coverage was limited and rarely went beyond default YouTube tags such as music, applause, or laughter. As video duration increased, non-speech information density declined, with contextual text (supporting spoken or signed content) becoming the dominant category [95].

Other studies similarly reported sparse non-speech information use: Alonzo et al. [5] found that most creators rarely included non-speech information, and Zhou et al. and McDonnell et al. [96, 156] noted limited non-speech information in music videos on social media. The expressive range was also narrow—Zhou et al. [156] observed that non-speech information was often reduced to generic placeholders (e.g., "[music playing]"), even in music-focused videos. In education, Hernandez et al. [52] found higher non-speech information inclusion on YouTube (42 of 48 teaching videos) but still mostly simple descriptions such as background music or applause. Non-speech information presence and density also varied by topic: May et al. [95] reported richer non-speech information diversity in lifestyle and music videos compared to sports or military content.

Research also underscored major limitations in current non-speech information datasets and systems. May et al. [95] showed that automated non-speech information captions often lack richness and interpretability, with small errors—especially with homophones, technical terms, or idiomatic sounds—significantly reducing comprehension. Alonzo et al. [5] likewise found that ambiguous sounds remain challenging for both humans and automated systems, such as off-screen noises being misinterpreted (e.g., a child mistaken for a horse).

**The Absence of Captions for Sign Language**

Researchers emphasized the near-total absence of SMVC for sign language. For DHH creators, captions are essential for reducing the heavy labor required to make signed content understandable [25, 26, 89]. Yet current infrastructures remain deeply insufficient. Studies show that major platforms provide no standardized tools for embedding sign-language interpretation, and their interfaces often crop or obscure the signing space [89, 96]. This reflects a core design limitation: YouTube, TikTok, and similar platforms were not built with signed communication in mind. Extending this critique, Cao et al. [25] found that Chinese platforms lack any real-time sign-to-text translation support, making live-stream interaction especially difficult. They also noted that training datasets for sign-language captioning on social media remain scarce, and regionally specific sign languages further complicate model development [25].

#### *5.3.2 Viewers' Perspectives on SMVC: Interaction, Benefits, Challenges, and Expectations.*

**Viewer Interaction with SMVC**

Interaction between DHH viewers and social media platform's captioned videos often takes the form of workaround strategies rather than seamless access. For example, Li et al. [81] found that DHH viewers on YouTube used hashtags (e.g., #CaptionYourVideos, #ClosedCaptions, or #CaptionPlease) to identify trustworthy captioned YouTube videos. Mack et al. reported DHH viewers rated the effort to find accessible versions of uncaptioned content as highly burdensome, describing long searches for captioned versions or similar niche content on social media channels [89].

Despite missing or inaccurate captions, viewers still find ways to make sense of videos and actively contribute feedback, often through comments or private messages. Mack et al. [89] observed that during emergencies, some viewers asked others to clarify what was happening. Prior work [25, 81, 128] also shows that DHH viewers sometimes contacted creators for transcripts, asked interpreters or hearing friends for help, or even provided corrected translations in the comment section for others. Kraeva and Krasnopeyeva [71] found similar patterns in YouTube comments on a popular Russian voiceover of a YouTuber - PewDiePie: viewers frequently praised translators, debated translation choices, or criticized caption quality.

**SMVC's Benefits to Viewers**

In addition, researchers examined how SMVC brought benefits to viewers on social media. Researchers overwhelmingly emphasized that captions are a primary means for DHH viewers to participate in these digital spaces [5, 23, 35, 46, 54, 68, 71, 81, 96, 109, 116, 127, 155]. For example, McDonnell et al., Sharevski et al., and Li et al. found that DHH users intended to skip videos without captions or with poor quality captions, viewing them as inaccessible and a waste of time [81, 96, 127, 155]. Guo et al. found that captions were not just accessibility tools but were central to the viewing experience for DHH viewers, and reported that approximately half of their visual attention was spent reading captions [46]. Sharevski et al. [127] observed that many expert-led debunking videos, such as those on TikTok, lacked proper captions, limiting accessibility for DHH viewers.

Researchers extended the importance of SMVC to other types of viewers, such as language learners, neurodivergent viewers, or all [4, 5, 46, 64, 68, 128, 133, 143]. Jiang et al. [64] emphasized that ADHD ones used captions to understand speech, maintain focus, and recover attention after distractions. More clearly, the article noted that captions were especially valued by people with ADHD to manage sensory processing challenges and proved helpful when videos had poor audio, noisy backgrounds, or unfamiliar accents on TikTok, YouTube, and Instagram. Captions also supported people with ADHD in maintaining focus, catching up, and retaining information, with some participants taking screenshots with captions on screen and taking notes for later. Aldukhayel et al. [4] found that learners of both a first language and a second language experienced declines in listening comprehension without captions. They noted that learners' comprehension with captioned videos, such as in YouTube vlogs. Teng et al. similarly showed that for global story comprehension, students with full captions outperformed all others, regardless of English proficiency [143]. In the study by Teng et al. [143], the comprehension scores were higher when the videos included full captions, followed by the videos in which only keywords of each sentence were captioned, and the lowest when no captions were provided.

Researchers also found that SMVC could help boost community participation and cultural relatability among viewers. Desai et al. [33] reported that high-quality captions can enable and low-quality captions can hinder access to culture and community. They described cases where YouTube viewers in North America, far from their cultural networks, found that inadequate captioning limited their ability to follow spoken content in non-English languages and, in turn, made cultural connection more difficult.

**Viewers' Challenges with SMVC**

In addition to benefits, researchers identified both technical and social barriers surrounding SMVC. Technically, viewers lack tools to assess caption quality. Li et al. [81] noted that DHH viewers on YouTube had no way to preview caption accuracy and were forced to manually check each video. Timing and synchronization problems were also persistent: captions frequently lagged behind speech or action, hindering comprehension [23, 33, 46, 96]. McDonnell et al. [96] found TikTok captions sometimes disappeared before viewers could finish reading them. Guo et al. [46] showed that hearing level and reading speed shaped preferences: Hard of Hearing viewers benefited from tightly synced captions, while profoundly deaf ones struggled with fast-paced formats. Butler et al. [23] added that viewers need time to adjust to unfamiliar caption speeds and that some pacing choices can unintentionally create barriers or reflect ableist assumptions.

Socially, studies show that captions on social media are often treated as an afterthought, leaving viewers feeling overlooked or excluded [89, 96, 141]. Desai et al. [33] found that social stigma can shape caption use; in some cultures (e.g., South Korea, India), requesting captions may be seen as burdensome. Desai et al. [33] and Li et al. [81] also noted that captions frequently miss cultural nuance, humor, or sign-language tone, creating moments where DHH viewers feel left out when others understand content they cannot. Raymond et al. [116] observed that missing captions in brand videos can even discourage deaf viewers from purchasing products or services on social media.

**Viewers' Expectations for SMVC**

Researchers have also emphasized the centrality of viewer expectations in shaping SMVC. Across platforms, accuracy remains a baseline value [35, 46, 81, 96, 133]. Smith et al. [133] demonstrated that typical adult readers struggled to comprehend automatically captioned YouTube videos with high error rates, particularly when audio was unavailable. However, expectations vary by platform: studies report that TikTok users often tolerate casual, stylized, or error-prone captions, while YouTube viewers expect higher accuracy and professionalism [35, 68, 71, 96]. Duraj et al. [35] found that TikTok viewers accepted flexible captioning when it enhanced humor, and McDonnell et al. [96] noted that DHH viewers on the same platform welcomed creative variation.

Expectations are also shaped by content genre. Berke et al. [17] showed that DHH YouTube viewers prioritized captions for news and politics, education, technology and science, film and animation, and entertainment, while considering captions as less critical for games, pets, sports, or music. Alonzo et al. [5] and Kim et al. [69] reported that DHH participants preferred graphic captions for entertainment content but text-based captions for more serious genres.

Finally, a growing body of work points to the demand for adjustable caption features. Viewers increasingly demand control over caption size, position, background, and language, as well as mechanisms for error reporting and feedback [5, 64, 69, 79, 81, 96, 109, 127, 128]. Sharevski et al. [127] highlighted the desire of DHH viewers for tools to flag caption errors and suggest corrections. Jiang et al. [64] documented calls for a clearer speaker distinction and customizable font, size, and color. Kim et al. [69] further underscored the need for adjustable levels of non-speech information captioning to accommodate individual preferences.

*5.3.3 Creators' Perspectives on SMVC: Motivations, Captioning Methods, and Challenges.*

**Creators' Motivations for Adding SMVC**

Researchers have noted that creators approach captioning on social media with varied motivations. Simpson et al. [131], drawing on

interviews with neurodivergent TikTok creators, observed that captioning was often motivated by personal encounters with accessibility barriers and framed as a practice of inclusion and community care. Cao et al. [25] found that Deaf TikTok creators were motivated by the desire to attract the attention of diverse viewers and broaden the participation. Li et al. [81] further reported that captioning practices sometimes emerged in response to user feedback, and creators gradually developed a greater awareness of its importance.

**Creators' Captioning Methods on SMVC**
Creators have different methods on how to caption on social media platforms [81, 95, 131]. For example, Li et al. [81], interviewed YouTube creators outside disability communities, reported a clear preference for self-captioning, followed by the automatic caption option with manual edits, and finally one without edits or third-party services.

Researchers also documented how creators exercised creativity to compensate for the lack of a proper captioning support system on social media platforms. Cao et al. [25] observed that Chinese streamers used low-tech strategies, such as typing or writing short phrases on a pad or screen and holding them to the camera to communicate. Similarly, Rettberg et al. [118] argued that when platforms restrict expressive modes (e.g., Musical.ly's largely visual format), users create new conventions, such as hand signs, to add emotional and gestural layers in their videos.

**Creators' Challenges with SMVC**
Furthermore, captioning poses substantial challenges for creators, requiring significant time and effort [79, 131, 141, 155]. Li et al. [81] described this as a dilemma: Although creators recognized the value of captions, they struggled to balance caption labor with personal obligations. Yoo et al. [155] further emphasized the intricacy of song signing captions, which requires technical synchronization of glosses with rhythm, artistic conveying of mood and musicality, and cultural negotiation of Deaf–hearing dynamics. Beyond the time and effort involved, creators contend with inadequate tools, limited interface support, and insufficient platform guidance [5, 52, 89, 113, 131]. Martin et al. [113] identified critical flaws in YouTube's subtitle editor, including a misleading layout and the absence of alerts when captioning standards were violated. Simpson et al. [131] noted the cognitive burden of TikTok's interface, which required creators to edit captions while simultaneously attending to video stimuli, leading creators to use alternative editor applications such as Adobe Premiere or CapCut. Lastly, creators faced the systemic opacity of platform moderation: takedowns arrived without explanation, and clarity, if it came at all, was always after the fact [25, 26, 131]. Deaf signers were especially vulnerable in this regime. In the absence of Deaf-aware moderators, routine signing gestures were misclassified as violations and wrongfully purged [25, 26].

*5.3.4 SMVC Systems, Techniques and Datasets.*
First, work on integrated accessibility interfaces demonstrates the potential for captioning to be embedded within broader multimodal accessibility ecosystems. Panagi et al. [108] introduced the Signifier Accessible Media Player (SiAMP), a unified media player that embeds multiple accessibility modalities—English captions, American Sign Language fingerspelling, sign-language interpretation video, and SignWriting, into a single customizable interface. Evaluated with native American Sign Language and Cypriot Sign Language users, SiAMP highlights how multimodal caption infrastructures can support cross-lingual preferences and diverse communication practices.

Second, researchers explored techniques that move beyond conventional Automatic Speech Recognition (ASR) pipelines. Huang et al. [54] introduced BandCaption, which decomposes captioning into micro-tasks distributed across DHH users, low-proficiency speakers, high-proficiency speakers, and native speakers. Their evaluations show that hybrid human–AI pipelines can substantially reduce ASR errors and maintain cost efficiency. Complementing this, Kim et al. [69] introduced OnomaCap, a system for generating culturally grounded onomatopoeic captions for non-speech information in Korean. Drawing on nearly 8,000 listener-annotated sound-onomatopoeia pairs, OnomaCap demonstrates how culturally specific representations expand caption expressivity beyond what ASR or sound-event detection can capture.

Along with that, May et al. [95] released the YouTube NSI Captioning Dataset, a corpus of 715,000 videos with millions of non-speech information annotations across diverse genres and acoustic environments. Addressing longstanding gaps in the ecological coverage of non-speech information corpora, this dataset enables new computational approaches to multimodal captioning that reflect real social media soundscapes in English.

Taken together, these systems, techniques, and datasets illustrate (a) the need for multimodal and culturally grounded caption representations, (b) the limitations of current platform pipelines, and (c) the technical possibilities that could support more collaborative, viewer-creator caption practices.

## 6 Discussion and Future Research Suggestions

Our findings address the three research questions by revealing how SMVC systems have evolved into a socially embedded practice. First, our analysis shows that SMVC systems have undergone substantial changes over the past decade, diversifying across platforms, communities, and caption types rather than following a single linear trajectory. Second, we find that viewers and creators actively use, interpret, and negotiate captions in practice, treating them not only as accessibility features but also as tools for engagement, learning, and expression. Finally, our findings surface persistent design and infrastructural gaps, such as limitations in captioning tools, feedback channels, non-speech information, and platform-level support. Together, these insights motivate our discussion of Participatory Captioning as a lens for understanding SMVC as a collective, community-involved infrastructure, and they inform the future research directions we outline in this section.

### 6.1 Participatory Captioning

*6.1.1 Articulating Participatory Captioning.*
Through our paper analysis, we observed that social media viewers are not silent spectators. They pause, comment, take screenshots, send messages to creators, share with others, take notes, or even start a campaign related to SMVC [64, 82, 96, 111, 141]. This was also central to the #NoMoreCraptions campaign in 2016, led by deaf YouTuber Rikki Poynter, who underscored that sustainable

captioning practices required creators to engage their communities rather than ignore the issue or rely solely on self-captioning and automated systems [18, 111]. In doing so, viewers do more than react to text on a screen; they help shape what captions mean, how creators are perceived, and how videos travel. And in return, based on their feedback and the platforms' regulations, the creators tried to change or improve video caption quality [25, 81]. This resonates with Jenkins et al.'s vision of participatory culture on social media as one that supports the free expression of artistic talent and civic engagement, enabling one to create and circulate their work among others [59]. Participatory design offers, as Schuler and Namioka wrote, of "a new approach to computer system design in which people destined to use the system play a critical role in its design" [126]. Also, in HCI, McDonnell et al.'s Collective Communication Access (CCA) framework argues that communication access is never the work of a single actor; it is a shared, co-constructed practice distributed across everyone involved in an interaction [96, 97]. These frameworks help explain why the cycle of viewer feedback and creator adjustments is significant: SMVC evolves as a participatory process: Viewers, creators, and platforms continually improve and adapt captions to meet diverse needs.

Taken together, we name this dynamic **Participatory Captioning**. *This can be considered a collaborative approach in which viewers and creators actively contribute to the production, evaluation, and refinement of SMVC, ensuring that they reflect diverse needs and expectations.* **Participatory Captioning** disrupts the traditional top-down model, where SMVC is delivered solely by platforms or creators, by showing how viewers themselves actively shape accessibility and meaning. While CCA provides the theoretical foundation that communication access is co-constructed by all participants [97], **Participatory Captioning** shows how this co-construction happens through creator practices, viewer corrections, community-driven norms, and platform infrastructures. It also opens avenues for future research in the design of feedback systems, moderation tools, and creator support, while foregrounding captioning as a form of collective access that raises questions about collective labor, individual ownership, and platform responsibility [97].

### Epistemological Foundations

Firstly, **Participatory Captioning** names a dynamic with a long history, but a new medium: advocacy, is required to ensure that captioned meaning stays intact and that no one is pushed out because a caption is missing or wrong. Caption access has always relied on ongoing community-shaped practice, and this continues as media environments evolve. For decades, captions were championed and advanced largely by DHH advocates - people who persuaded broadcasters and insisted, often against resistance, that captions were a public good [106]. Today, that responsibility is shifting as public life increasingly happens online. Now, people maintain the expectation of quality captioning through commenting, sharing, reporting, or starting a campaign on the very platforms that shape contemporary communication [64, 82, 96, 111, 141]. In this way, social media becomes not merely a site of entertainment but also one of the arenas where the ongoing community-shaped practice continues.

Secondly, beyond correcting mistakes, captioning conventions emerge as users remix formats, experiment with style, and embed cultural references or in-group humor. Captioning today is also a medium of expression shaped collectively by viewers, creators, and cultural communities [14, 82, 119]. Creative expression is inherently variable—individuals draw on their own linguistic repertoires, cultural backgrounds, and lived experiences when interpreting and describing moments on screen [33, 99, 147]. **Participatory Captioning** recognizes this multiplicity as a central feature of captioning practices rather than a limitation and pushes this approach beyond the narrow frame of crowdsourced error correction. The effectiveness of **Participatory Captioning** might not be tied to mass engagement. However, even limited contributions can have a disproportionate impact because they shape how entire viewers sustain accessibility norms that platforms benefit from [44, 137]. The question is not "How many users will contribute?" but "What is the cost of ignoring the people who already do and whose work upholds interpretability for everyone? "

Lastly, recent work has reimagined the purpose of access from achieving measurable outcomes (e.g., comprehension) to increasing agency, as defined by individuals [117]. This perspective resonates strongly with **Participatory Captioning**. When SMVC function as a participatory layer, where viewers annotate sound, reinterpret meaning, contest misrepresentation, or express SMVC as culturally grounded perspectives, they become a site of connection, co-presence, and epistemic contribution. It enables people to speak in their own language, debate accessibility on their own ground, and surface forms of knowledge through their norms [146]. Notably, **Participatory Captioning** functions differently than other participatory frames within HCI. Unlike researcher-initiated participatory design processes or crowdsourcing efforts, the participation we observe in SMVC emerges outside designer control [25, 96, 131, 141]. Viewers intervene not because a system asks them to, but because caption errors affect how their communities interpret and experience a video. These acts of correction, annotation, and reinterpretation are therefore not contributions to a design process; they are expressions of care, responsibility, and community governance [78, 84**?** ].

### Critical Perspectives

We acknowledge that **Participatory Captioning** is not without critique, and we view these critiques as an important part of advancing the idea. We invite researchers, platforms, and communities to consider them together in further conversations. First, we consider long-standing critiques in participatory design that caution against treating participation as inherently democratizing. As Harrington et al. [47] argued, participatory design often privileges those who already hold social, linguistic, or institutional power, while marginalizing participants. **Participatory Captioning** could inadvertently reproduce structural inequities, placing disproportionate labor on viewers, particularly those from non-English contexts or underrepresented communities. Treating **Participatory Captioning** as a negotiated, power-aware practice means attending to who is allowed to shape captions, who bears the burden of correction, and how platform policies distribute or concentrate authority.

Moreover, community captioning initiatives, such as YouTube's now-retired community caption feature, also illustrate both the promise and fragility of distributed accessibility labor [87]. Viewers consistently call for mechanisms to provide feedback to creators.

At the same time, in practice, Li et al. [81] found that the captions contributed by the YouTube community often lacked consistency, professionalism, and fidelity to standards, introducing more errors and harassment. The research question that follows is not simply whether platforms should reintroduce community captioning, but how feedback infrastructures might be designed to balance openness with quality control and creator safety. Future work must consider the challenge of feedback as a socio-technical system: without carefully designed channels and moderation, 'participation' risks devolving into uncoordinated labor that reproduces the very inaccessibility it sought to resolve.

Lastly, we also recognize the risk that **Participatory Captioning** could reproduce forms of invisible labor, echoing what Simpson and Semaan [132] described as the community work that quietly sustains platforms. Naming this risk is essential, as it shifts the focus from what communities are already doing to what platforms should shoulder. If viewers and creators are providing linguistic expertise, correcting errors, and maintaining caption quality, then the central question becomes how platforms acknowledge, support, and sustain this labor. Future research could examine mechanisms for recognizing user contributions, motivating creators to improve captions, and designing moderation systems that complement community efforts while respecting cultural nuance.

Considering platforms also raises governance questions: how captioning standards adapt as language evolves, how feedback quality is assessed [76, 80], and how platform practices align with accessibility regulations and differing regional policies. Together, these issues position **Participatory Captioning** not only as a collaborative practice but also as a design and policy challenge that requires platforms to actively sustain and build on the voluntary contributions that currently uphold accessibility.

#### *6.1.2 Towards Design Opportunities and Challenges for Participatory Captioning.*

This section outlines key design opportunities and challenges emerging from **Participatory Captioning** (see Figure 6).

**Creator-side Design Opportunities**

*Understanding Creator Motivations and Recognition Systems* **Participatory Captioning** points toward several design opportunities for platforms to more effectively support the shared labor of SMVC production. Existing work suggests that creators caption for a wide range of reasons, from accessibility commitments to audience engagement or lived experience [25, 81, 131]. In addition to these motivations, creators also face practical incentives to monitor caption accuracy (e.g., mis-transcribing benign speech as inappropriate or offensive terms that trigger demonetization) [33, 81, 96, 131]. These risks create an additional layer of motivation for creators to review and correct captions, not only to support viewers but also to achieve their goals. Quantifying and categorizing these motivations could reveal opportunities for platforms to provide targeted support, encouraging creators to craft captions that align with their goals.

*Recognizing Captioning Labor and Growth.* Platforms could also offer cues that show how a creator's captioning practices evolve over time, highlighting small improvements, recurring issues, or shifts in clarity and accessibility. Such cues could provide indicators of progress, helping creators understand which captioning choices strengthen accessibility, audience comprehension, or their own goals. Platforms might also explore recognition systems that acknowledge captioning labor, similar to how existing awards highlight creative excellence [121].

**Viewer-Side Design Opportunities**

*Designing Actionable Viewer Feedback.* **Participatory Captioning** highlights opportunities for platforms to better support the interpretive and corrective work that viewers already perform. Rather than relying solely on comment sections, platforms could explore mechanisms that make caption-related feedback more actionable for creators. For instance, structured suggestion interfaces or mini-forum-like spaces could allow viewers to propose edits, upvote helpful corrections, or collectively identify recurring issues. These systems would not replace creator judgment but could help organize community contributions into clearer, more interpretable signals.

Another opportunity lies in giving viewers more control over the captioning experience itself. Future systems could allow viewers to optionally self-identify the communities or viewing preferences that shape their caption needs—such as selecting "DHH viewer," "language learner," or "neurodivergent viewer." These preferences could inform how captions are displayed or prioritized (e.g., richer non-speech information for DHH viewers, vocabulary scaffolds for learners, or simplified pacing for neurodivergent audiences). Additionally, videos that consistently receive strong community evaluations on caption clarity or accuracy could be marked with indicators that help viewers discover well-captioned content [81].

*Prioritizing Epistemic Expertise in Viewer-Driven Caption Corrections.* **Participatory Captioning** also raises questions about how to prioritize different forms of viewer expertise. Not all feedback serves the same purpose: for example, DHH viewers depend on captions more directly for accessing video content; their feedback often reflects needs that are especially critical for comprehension [106]. Future systems may therefore need mechanisms that surface accessibility-critical expert feedback without diminishing broader community participation. Future work should investigate how platforms might support voluntary, privacy-preserving forms of self-identification, surface heterogeneous types of expertise, and organize feedback in ways that respect differing needs without reproducing rigid tiers of authority.

**Design Challenges**

Multiple audience-specific caption versions can significantly improve accessibility, yet they introduce challenges for consistency, creative expression, industry standards, and long-term maintenance. Also, over-optimized AI systems might homogenize caption style and weaken captioning as a medium of creative and cultural expression. Future research should examine how captioning systems can support diverse audience needs without increasing cognitive load or eroding creative identity. This includes designing adaptive interfaces that surface features only when relevant, reducing unnecessary complexity for creators with different levels of expertise.

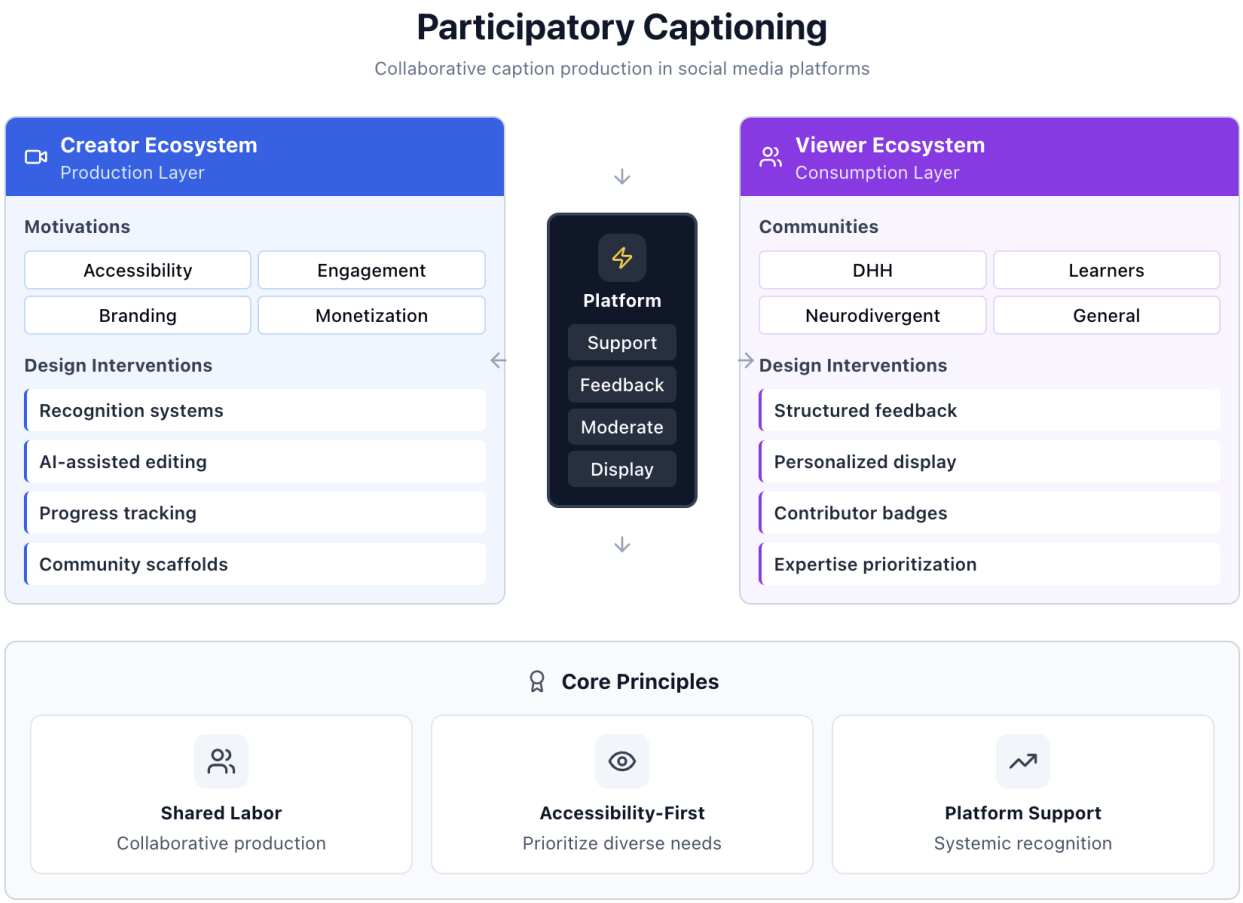


**Figure 6: Participatory Captioning system showing the design opportunities for the dynamics among creators, platforms, and viewers**

Another key consideration must be whether new systems create unsustainable long-term maintenance needs. Future research should further define the roles of creators, viewers, and platforms; provide transparency around how feedback is evaluated; and ensure that community contributions are enhanced.

## 6.2 Future Research Directions for SMVC

We now identify opportunities for future work around SMVC.

*6.2.1 Expansion of Who Is Considered an SMVC User.* Although captions have historically been legislated as an assistive tool for DHH viewers, there is also a long history of research demonstrating their utility for other groups. Prior work has shown that captioning benefits language learners, students, and neurodivergent people, and future captioning systems should be designed both to serve these groups and with the acknowledgment that DHH people also fall into these categories [4, 33, 34, 53, 60–63, 69, 109, 131, 133, 134]. What social media makes visible is not a sudden shift but a broadening of recognition: once captions circulate in an environment where everyone can produce, comment, and connect, their relevance to all becomes unavoidable, and future SMVC should be designed accordingly. Future work could also examine cross-group encounters such as DHH–hearing, DHH–ADHD, and interactions among DHH communities across different cultural or linguistic backgrounds. Such analysis would provide a more accurate account of practice by revealing where interpretations align and where they diverge.

*6.2.2 SMVC through Creator Motivations and Presentation Choices.* One recurring theme concerns creator motivations to caption. Some caption out of care work and a desire to make content accessible to others facing similar barriers; others develop awareness through long-term creation, shaped by viewer feedback and community norms [25, 81, 131], while others caption social media videos for visibility purposes [25]. Furthermore, researchers highlighted the difficulties that creators faced when adding captions, including TikTok's complex editing interface and a general lack of guidance [25, 81, 131, 141]. At the same time, researchers observed that in the absence of robust social media video captioning tools, creators improvised alternatives—for example, typing or writing short phrases on a pad or screen and holding them up to the camera [25], or using hand signs to layer gestures onto videos on Musical.ly [118].

However, few have captured creator workflows in depth. Diary studies or longitudinal methods could reveal the everyday realities of captioning, technical hurdles, time pressures, and creative trade-offs that are invisible in surveys or interviews [21].

Additionally, the content choices of the creators were often shaped by personal branding strategies and aesthetic orientations [99, 147]. Extending this insight to SMVC, an open question is whether caption presentation itself constitutes another site for negotiating branding and aesthetic identity. Addressing this question might deepen the understanding of creator captioning practices and inform the design of captioning tools and workflows that better align with these practices.

*6.2.3 Expansion Beyond English-speaking and Western contexts.* We observed work that extended beyond the English-speaking and Western contexts, such as papers with contexts from China, Saudi Arabia, or South Korea [46, 69, 141, 143]. These highlight how language, culture, and platform affordances intertwine, but the coverage is still thin, with numerous language and dialect problems reported in our analysis [33, 52, 71, 79, 145]. In addition, researchers have observed that languages continue to disappear each year [33]. Viewers also often rely on captions as tools for language learning and cultural connection. Taken together, this suggests that captions may serve not only as aids for individual comprehension but also as potential supports for language preservation. Researchers have pointed out the need to build datasets that encompass multiple scripts, dialects, non-speech information, sign languages, or culturally specific terms [5, 17, 25, 33, 69, 79, 89, 95, 108, 112, 145]. However, cultural norms differ across national and regional contexts, so that conversational practices that are considered appropriate and intelligible in one culture may not be easily transferable or comprehensible in another [19].

As such, we recommend future research examine how current social media automatic captioning systems account for cultural diversity and interrogate the risks that arise when captions reproduce multiple languages, potentially reinforcing inequities in access and interpretation [3, 142]. In addition, although the incorporation of multilingual languages and culturally specific terminology can enrich the expressive capacity of captions, it also introduces complex challenges related to moderation, privacy, and content ownership. While some regionally specific linguistic forms carry cultural authenticity, others may encode subtle or overt harms [15]. As creators address increasingly diverse viewers, future work should examine how they negotiate the tension between authenticity and accessibility, for example, whether to caption in their own language or dialect or prioritize intelligibility for heterogeneous viewer communities or certain standards [33, 49].

*6.2.4 Toward Platform Expansion.* We observed that existing research is disproportionately concentrated on a small set of dominant platforms, most notably YouTube, TikTok, and Facebook, along with several Chinese-based platforms such as RedNote, Douyin, and Bilibili [5, 25, 71, 95, 96, 128, 131, 141]. In contrast, emerging work in related fields has begun to examine how technical features shape addictive engagement in Instagram Reels, YouTube Shorts, and Facebook Watch, suggesting the need for a closer analysis of these formats in SMVC research in HCI [100, 115, 120]. Each platform fosters its own culture: YouTube's search and archiving functions support long-form educational content [130]; TikTok thrives on short-form, algorithmically curated entertainment [91]; and Instagram privileges lifestyle aesthetics [125].

We also encourage researchers to expand the scope of the study to include emerging and non-Western platforms, such as Moj in India [104] or Likee and Bigo Live in Singapore [123, 135]. Insights from cross-cultural HCI and other fields underscore that interaction practices are culturally situated [65, 92, 129], implying that platform-specific studies risk overlooking important socio-technical differences. Demographic profiles, creator practices, and technical amenities vary among platforms, uniquely shaping captioning behaviors and quality [51, 148]. Therefore, comparative research across platforms could illuminate the diverse functions of captions, including but not limited to how creators balance accessibility and engagement and how viewers evaluate caption quality, presentation, and utility. Such analysis would extend current understandings of SMVC, positioning captions not simply as a technical affordance but as a socially situated communicative practice embedded in platform cultures.

*6.2.5 The Need for SMVC Guidelines.* Although participants expressed openness to flexibility on platforms and video genres, they expected certain rules and guidelines for SMVC [25, 46, 52, 96, 109, 127, 128]. At the same time, the competitive and creativity-driven environment of social media continually pushes creators to innovate [45], raising design challenges around the balance of standardization and expressive freedom. A useful direction for future work is to study how communities co-create informal captioning norms and consider how these might be translated into platform-level standards. Another is to explore participatory feedback systems in which DHH, neurodivergent, and other viewers help establish quality benchmarks [25, 46, 52, 96, 109, 127, 128]. Beyond setting guidelines, future work should confront implementation and governance: how platforms embed accessibility standards into large-scale moderation while addressing uneven enforcement and the opacity of current practices.

*6.2.6 SMVC Interface Design for Visibility.* Current social media video interfaces often overload screens with captions, stickers, and visual effects, leading to overlap or occlusion of critical visual information. Such issues undermine both usability and accessibility, particularly for DHH viewers [7, 25, 96]. Although prior television research has addressed caption placement, occlusion, and style [7, 63], SMVC presents new challenges due to the density and dynamism of on-screen elements, especially with those burned into videos. Future work should investigate interface designs that balance information richness with clarity, with particular attention to open captions, which are permanently embedded and cannot be repositioned or customized by viewers [96].

*6.2.7 The Rapidly Changing Platform Landscape.* Our analysis suggests that SMVC research has consistently lagged behind the rapid development of platforms and creator practices. This gap has often been characterized as a form of cultural lag, in which technological innovations outpace scholarly interpretation and institutional adaptation [107]. Major platform disruptions and shifts in captioning practices remain under-examined. For example, while YouTube's discontinuation of its community captioning feature has been noted

[87], we still do not know how much invisible volunteer labor was lost, how heavily DHH viewers, language learners, and multilingual communities depended on that labor, or how creators adapted after the feature disappeared. Studying this is necessary to understand how governance decisions redistribute accessibility, how community labor is erased, and how platform power is exercised in ways that are neither transparent nor accountable. In addition, work on the implementation of automatic captions on YouTube and TikTok has highlighted implications for accessibility [66, 81, 96, 131, 145]. However, little is known about the broader consequences of this transition, compared with earlier contexts that lacked automatic captions. Similarly, TikTok research yields conflicting accounts of caption prevalence: some studies report widespread use, while others suggest the opposite [26, 96, 141].

*6.2.8 Integrating Creator, Viewer, and Platform Considerations.* As platforms now bring together a highly heterogeneous mix of creators and viewers [5, 25, 33, 35, 52, 64, 69, 81, 89, 95, 96, 131], the interactions among these groups have become central to how captioning is produced, interpreted, and negotiated. Recent advances in multimodal AI systems are beginning to reshape how creators produce and edit video content, impacting scripting, editing, and creative production support [20, 88]. These practices are little understood, and future work should explore use patterns (e.g., if generative AI is used for drafting video captions or detecting errors) to document this evolving workflow. We must also assess developing workflows with regard to limitations in the current captioning infrastructure. Automated speech recognition tools struggle with tone, humor, emotional nuance, prosody, and culturally situated expressions [83, 114]—elements that are central to how viewers make sense of social media videos [12, 58]. There is evidence that emerging large audio language models suffer from hallucination [72], limited temporal event reasoning [73] and generally remain far from capturing the contextual, affective, and culturally variable interpretations that viewers rely on [95]. Taken together, these dynamics point toward a future in which captions become increasingly hybrid, co-produced by creators, viewers, and AI systems. This shift raises new questions for future research around how captioning labor is distributed between humans and algorithms, what forms of interpretation become privileged or obscured when AI mediates meaning, how creators could be supported in prompting AI systems to capture culturally situated meaning, and whether current benchmarks [66] are appropriate to evaluate captions generated by multimodal AI models.

## 6.3 Limitations

Although work from diverse communities exists [46, 69, 141, 143], our review is limited to studies published in English. We see this limitation as a recommendation and invitation for future collaborative work. SMVC practices might evolve differently across regions and languages, and many culturally rich and visually inventive approaches emerge along with Anglo contexts [56, 92]. Incorporating cross-regional approaches helps counter the biases that arise when speech and language datasets are trained primarily on English-language [15, 52, 105, 112, 124, 145]. This also matters for DHH viewers everywhere, including those in Western settings. Captioning practices shaped by diverse regions might help them fully engage with content as they travel, migrate, and navigate unfamiliar environments, even on social media - dynamics documented in research on Deaf mobility [37].

Also, although our search strategy captured the core set of papers on SMVC, especially in terms of social media video captioning systems, we acknowledge that it is not fully comprehensive. Our review relied primarily on Google Scholar, and we did not conduct systematic searches across larger databases such as Scopus or IEEE Xplore. While Google Scholar surfaced papers from adjacent fields (e.g., social media studies, marketing, and disability studies), we did not manually screen the libraries in these broader domains, where additional relevant work may exist. A future review that incorporates multi-database searches and deeper cross-disciplinary coverage might strengthen the completeness and reproducibility of the evidence base.

# 7 Conclusion

Our review of 36 papers follows the trends in SMVC research, highlighting its role in accessibility, engagement, and creative practices. We propose Participatory Captioning as a way to involve viewers, creators, and platforms in shaping equitable captioning practices on social media platforms and suggest future design implications. Although SMVC research has expanded in scope, important gaps remain in areas, opening further research opportunities. We hope this serves as a useful resource for platforms, communities, and researchers working toward more accessible captioning practices.

# Acknowledgments

This work was supported by the National Science Foundation under Grant No. 2504642 and the National Library of Medicine under Grant No. T15LM007442.

We used ChatGPT for editing support, including grammar clarification and wording refinement during the revision process. All ideas, synthesis, and analysis come from the authors.

# References

[1] Adobe. 2022. *Future of Creativity: Creators in the Creator Economy.* Technical Report. Adobe (with Edelman Data & Intelligence). https://www.arpp.org/wp-content/uploads/2022/09/Adobe-Future-of-Creativity-Study_Creators-in-the-Creator-Economy.pdf Accessed: 2025-08-15.

[2] AI-Media. 2025. European Accessibility Act 2025: What Broadcasters Need to Know. https://www.ai-media.tv/knowledge-hub/insights/european-accessibility-act-2025/ Accessed: 2025-09-07.

[3] Abdullah Al-Momani, Ahmad S. Haider, Mohammed Dagamseh, and Wala' Mohammad Akasheh. 2025. Audience Responses to Cultural and Linguistic Gaps in English–Arabic Auto-Subtitles on YouTube. *Research Journal in Advanced Humanities* 6, 3 (aug 2025). doi:10.58256/x3sten23

[4] Dukhayel Aldukhayel. 2021. The effects of captions on L2 learners' comprehension of vlogs. *Language Learning & Technology* (2021).

[5] Oliver Alonzo, Hijung Valentina Shin, and Dingzeyu Li. 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 (ASSETS '22).* Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3517428.3544808

[6] Akhter Al Amin, Abraham Glasser, Raja Kushalnagar, Christian Vogler, and Matt Huenerfauth. 2021. Preferences of deaf or hard of hearing users for live-TV caption appearance. In *International Conference on Human-Computer Interaction.* Springer, 189–201.

[7] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Caption-occlusion severity judgments across live-television genres from deaf and hard-of-hearing viewers. In *Proceedings of the 18th International Web for All Conference*. 1–12.

[8] Akhter Al Amin, Saad Hassan, and Matt Huenerfauth. 2021. Effect of occlusion on deaf and hard of hearing users' perception of captioned video quality. In *International Conference on Human-Computer Interaction*. Springer, 202–220.

[9] Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2023. Understanding how deaf and hard of hearing viewers visually explore captioned live TV news. In *Proceedings of the 20th International Web for All Conference*. 54–65.

[10] Priya Anand. 2020. *Zoom Daily Users Surge to 300 Million Despite Privacy Woes*. https://www.bloomberg.com/news/articles/2020-04-22/zoom-daily-users-surge-to-300-million-despite-privacy-woes Accessed: 2025-08-31.

[11] Mariana Arroyo Chavez, Molly Feanny, Matthew Seita, Bernard Thompson, Keith Delk, Skyler Officer, Abraham Glasser, Raja Kushalnagar, and Christian Vogler. 2024. How users experience closed captions on live television: quality metrics remain a challenge. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[12] James M Barry and Sandra S Graça. 2018. Humor effectiveness in social video engagement. *Journal of Marketing Theory and Practice* 26, 1-2 (2018), 158–180.

[13] Ava Bartolome and Shuo Niu. 2023. A Literature Review of Video-Sharing Platform Research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 790, 20 pages. doi:10.1145/3544548.3581107

[14] Michael Baumgartner. 2025. 150+ Video Marketing Statistics You Can't Afford to Ignore in 2025. https://www.zebracat.ai/post/video-marketing-statistics Accessed: 2025-08-16.

[15] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[16] Cynthia L Bennett and Daniela K Rosner. 2019. The promise of empathy: Design, disability, and knowing the" other". In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[17] Larwan Berke, Matthew Seita, and Matt Huenerfauth. 2020. Deaf and hard-of-hearing users' prioritization of genres of online video content requiring accurate captions. In *Proceedings of the 17th International Web for All Conference*. 1–12.

[18] Linda Besner. 2019. When Is a Caption Close Enough? YouTube's notoriously nonsensical auto-captions are improving. But there's a deeper problem. *The Atlantic* (9 aug 2019). https://www.theatlantic.com/health/archive/2019/08/youtube-captions/595831/

[19] Heather Bowe, Kylie Martin, and Howard Manns. 2014. *Communication across cultures: Mutual understanding in a global world*. Cambridge University Press.

[20] Jasper David Brüns and Martin Meißner. 2024. Do you create your content yourself? Using generative artificial intelligence for social media content creation diminishes perceived brand authenticity. *Journal of Retailing and Consumer Services* 79 (2024), 103790.

[21] Jean Burgess, Kath Albury, Anthony McCosker, and Rowan Wilken. 2022. *Everyday data cultures*. John Wiley & Sons.

[22] Jean Burgess and Joshua Green. 2018. *YouTube: Online video and participatory culture*. John Wiley & Sons.

[23] Janine Butler. 2019. The Visual Experience of Accessing Captioned Television and Digital Videos. *Television & New Media* 21, 5 (2019), 1–18. doi:10.1177/1527476418824805

[24] Kelsey Cameron. 2025. Contesting Captions: Netflix, Fan Campaigns, and the Labor of Access. *Television & New Media* 26, 5 (2025), 570–585.

[25] Beiyan Cao, Changyang He, Muzhi Zhou, and Mingming Fan. 2023. SparklingSilence: Practices and Challenges of Livestreaming Among Deaf or Hard of Hearing Streamers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)* (Hamburg, Germany). Association for Computing Machinery, New York, NY, USA, 15. doi:10.1145/3544548.3581053

[26] Jiaxun Cao, Xuening Peng, Fan Liang, and Xin Tong. 2024. "Voices Help Correlate Signs and Words": Analyzing Deaf and Hard-of-Hearing (DHH) TikTokers' Content, Practices, and Pitfalls. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3613904.3642413

[27] Centre for Inclusive Design. 2025. History – Centre for Inclusive Design. https://centreforinclusivedesign.org.au/about/history/ Accessed: 2025-09-07.

[28] Michael Crabb, Rhianne Jones, Mike Armstrong, and Chris J Hughes. 2015. Online news videos: the UX of subtitle position. In *Proceedings of the 17th international ACM SIGACCESS conference on Computers & accessibility*. 215–222.

[29] Barry Jay Cronin. 1980. Closed-caption television: Today and tomorrow. *American Annals of the Deaf* 125, 6 (1980), 726–728.

[30] Jenny L Davis. 2012. Social media and experiential ambivalence. *Future Internet* 4, 4 (2012), 955–970.

[31] Caluã de Lacerda Pataca, Matthew Watkins, Roshan Peiris, Sooyeon Lee, and Matt Huenerfauth. 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*. 1–15.

[32] Deloitte Insights. 2025. Digital Media Trends: 2025 Survey. https://www.deloitte.com/us/en/insights/industry/technology/digital-media-trends-consumption-habits-survey/2025.html (accessed August 14, 2025).

[33] Aashaka Desai, Rahaf Alharbi, Stacy Hsueh, Richard E Ladner, and Jennifer Mankoff. 2025. Toward Language Justice: Exploring Multilingual Captioning for Accessibility. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[34] Gilbert Dizon and Benjamin Thanyawatpokin. 2021. Language learning with Netflix: Exploring the effects of dual subtitles on vocabulary learning and listening comprehension. *Computer-Assisted Language Learning Electronic Journal* 22, 3 (2021), 52–65.

[35] K. Duraj and A. Szarkowska. 2025. Beyond Traditional Subtitles: How Emojis and Non-Standard Typography in Subtitles Boost Engagement on TikTok. *Journal of Audiovisual Translation* 8, 1 (2025), 1–28. doi:10.47476/jat.v8i1.2025.339 Editor(s): J. Pedersen. Received: August 8, 2024. Published: April 11, 2025. ©2025 Author(s). This is an open access article under the Creative Commons Attribution License..

[36] Shira Dvir-Gvirsman, Daniel Sude, and Guy Raisman. 2024. Unpacking news engagement through the perceived affordances of social media: A cross-platform, cross-country approach. *New Media & Society* 26, 11 (2024), 6487–6509.

[37] Steven Emery, Sanchayeeta Iyer, Amandine Le Maire, Erin Moriarty, and Annelies Kusters. 2024. *Deaf mobility studies: Exploring international networks, tourism, and migration*. Gallaudet University Press.

[38] Federal Communications Commission. 2011. *Implementation of the Twenty-First Century Communications and Video Accessibility Act of 2010; Closed Captioning of Internet Protocol-Delivered Video Programming*. Technical Report FCC 11-138. Federal Communications Commission. https://docs.fcc.gov/public/attachments/FCC-11-138A1.pdf

[39] Richard Florida and Creative Class Group. 2024. *The Creator Revolution: Results From a Global Survey*. Technical Report. Creative Class Group. https://creativeclass.com/reports/The-Creator-Revolution.pdf Commissioned by Meta; based on a global survey across 20 countries.

[40] Fortune Business Insights. 2025. Web Conferencing Market Size, Share & COVID-19 Impact Analysis, By Component … 2020–2027. https://www.fortunebusinessinsights.com/web-conference-software-market-102993 Valued at USD 3.62 billion in 2019; projected USD 10.46 billion by 2027; CAGR 14.3%. Accessed: 2025-08-31.

[41] Sherice Gearhart and Seok Kang. 2014. Social media in television news: The effects of Twitter and Facebook comments on journalism. *Electronic News* 8, 4 (2014), 243–259.

[42] Morton Ann Gernsbacher. 2015. Video Captions Benefit Everyone. *Policy Insights from the Behavioral and Brain Sciences* 2, 1 (oct 2015), 195–202. doi:10.1177/2372732215602130

[43] Goldman Sachs Research. 2023. *The Creator Economy Could Approach Half-a-Trillion Dollars by 2027*. https://www.goldmansachs.com/insights/articles/the-creator-economy-could-approach-half-a-trillion-dollars-by-2027.html

[44] Mary L Gray and Siddharth Suri. 2019. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Harper Business.

[45] Daniel P. Gross. 2020. Creativity Under Fire: The Effects of Competition on Creative Production. *The Review of Economics and Statistics* 102, 3 (2020), 583–599. doi:10.1162/rest_a_00831

[46] Sijia Guo, Helena Sit, and Shen Chen. 2020. Effects of captioned videos on learners' comprehension. *Journal of Global Literacies, Technologies, and Emerging Pedagogies* 6, 1 (2020), 1062–1082.

[47] Christina Harrington, Sheena Erete, and Anne Marie Piper. 2019. Deconstructing community-based collaborative design: Towards more equitable participatory design engagements. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–25.

[48] Maria Harutyunyan. 2025. Video Marketing Statistics To Consider For 2025 Campaigns. https://loopexdigital.com/blog/video-marketing-statistics Published: October 31, 2024; Updated: September 2, 2025.

[49] Uday Sadiq Hasan and N Solomon Benny. 2025. Code-Switching in Digital Communication: A Pragmatic Approach to Multilingual Interactions on Social Media. *South Asian Journal of Social Sciences & Humanities* 6, 3 (2025).

[50] Husna Shafirah Binti Helmirzal, Zuraidah Binti Mohd Sulaiman, and Afifah Binti Fadhlullah. 2023. NoCap Captions: Providing Subtitle for Content Creators. *International Journal of Academic Research in Business and Social Sciences* 13, 1 (2023), 533–542. doi:10.6007/IJARBSS/v13-i1/16210

[51] Jonathan Hendrickx and Jorge Vázquez-Herrero. 2024. Dissecting social media journalism: A comparative study across platforms, outlets and countries. *Journalism Studies* 25, 9 (2024), 1053–1075.

[52] Q. M. G. Hernandez. 2024. A Qualitative Study of Closed Captions in English Language Teaching Videos on YouTube. *Journal of English and Applied Linguistics* (2024). Available at Animo Repository, De La Salle University.

[53] Hsin-Chuan Huang and David E Eskey. 1999. The effects of closed-captioned television on the listening comprehension of intermediate English as a second language (ESL) students. *Journal of Educational Technology Systems* 28, 1 (1999), 75–96.

[54] Yun Huang, Yifeng Huang, Na Xue, and Jeffrey P. Bigham. 2017. Leveraging Complementary Contributions of Different Workers for Efficient Crowdsourcing of Video Captions. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, Denver, CO, USA, 1830–1837. doi:10.1145/3027063.3053164

[55] Chris J Hughes, Mike Armstrong, Rhianne Jones, and Michael Crabb. 2015. Responsive design for personalised subtitles. In *Proceedings of the 12th International Web for All Conference*. 1–4.

[56] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1311–1320.

[57] Arvin Jagayat and Becky L Choma. 2024. A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond. *Current Opinion in Psychology* 55 (2024), 101726.

[58] Allan James. 2017. Prosody and paralanguage in speech and the social media: The vocal and graphic realisation of affective meaning. *Linguistica* 57, 1 (2017), 137–149.

[59] Henry Jenkins, Mizuko Ito, and danah boyd. 2015. *Participatory Culture in a Networked Era: A Conversation on Youth, Learning, Commerce, and Politics*. Polity Press, Cambridge, UK.

[60] Carl Jensema. 1998. Viewer reaction to different television captioning speeds. *American annals of the deaf* 143, 4 (1998), 318–324.

[61] Carl J. Jensema and Robb Burch. 1999. *Caption Speed and Viewer Comprehension on Programs. Final Report.* Technical Report ERIC ED434446. U.S. Department of Education.

[62] Carl J Jensema, Ramalinga Sarma Danturthi, and Robert Burch. 2000. Time spent viewing captions on television programs. *American annals of the deaf* 145, 5 (2000), 464–468.

[63] Carl J Jensema, Sameh El Sharkawy, Ramalinga Sarma Danturthi, Robert Burch, and David Hsu. 2000. Eye movement patterns of captioned television viewers. *American annals of the deaf* 145, 3 (2000), 275–285.

[64] Lucy Jiang, Woojin Ko, Shirley Yuan, Tanisha Shende, and Shiri Azenkot. 2025. Shifting the Focus: Exploring Video Accessibility Strategies and Challenges for People with ADHD. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[65] Dal Yong Jin. 2025. Platform Imperialism Theory From the Asian Perspectives. *Social Media+ Society* 11, 1 (2025), 20563051251329692.

[66] Sushant Kafle and Matt Huenerfauth. 2017. Evaluating the usability of automatically generated captions for people who are deaf or hard of hearing. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. 165–174.

[67] Martin Kenney and John Zysman. 2016. The rise of the platform economy. *Issues in science and technology* 32, 3 (2016), 61.

[68] Hyunju Kim, Yan Tao, Chuanrui Liu, Yuzhuo Zhang, and Yuxin Li. 2023. 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*. 1–6.

[69] JooYeong Kim and Jin-Hyuk Hong. 2025. OnomaCap: Making Non-speech Sound Captions Accessible and Enjoyable through Onomatopoeic Sound Representation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, 1–22. doi:10.1145/3706598.3713911

[70] Patricia S Koskinen, Robert M Wilson, and Carl J Jensema. 1985. Closed-captioned television: A new tool for reading instruction. *Literacy Research and Instruction* 24, 4 (1985), 1–7.

[71] Svetlana Kraeva and Ekaterina Krasnopeyeva. 2020. Judging Translation on Social Media: A Pragmatic Look at YouTube Comment Section. In *10th International Conference "Word, Utterance, Text: Cognitive, Pragmatic and Cultural Aspects" (WUT 2020) (The European Proceedings of Social and Behavioural Sciences, Vol. 95)*. European Publisher, 791–799. doi:10.15405/epsbs.2020.08.91

[72] Chun-Yi Kuan and Hung-yi Lee. 2025. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[73] Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992* (2025).

[74] Seda Kuscu-Ozbudak. 2022. The role of subtitling on Netflix: an audience study. *Perspectives* 30, 3 (2022), 537–551.

[75] Raja S Kushalnagar and Christian Vogler. 2020. Teleconference accessibility and guidelines for deaf and hard of hearing users. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–6.

[76] Cliff Lampe and Paul Resnick. 2004. Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 543–550.

[77] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction*. Morgan Kaufmann.

[78] Christopher A Le Dantec and Sarah Fox. 2015. Strangers at the gate: Gaining access, building rapport, and co-constructing community-based research. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 1348–1358.

[79] Jeong-Hwa Lee and Kyung-Whan Cha. 2020. An Analysis of the Errors in the Auto-Generated Captions of University Commencement Speeches on YouTube. *Journal of Asia TEFL* 17, 1 (2020), 143–159.

[80] Chao Li and Balaji Palanisamy. 2019. Incentivized blockchain-based social media platforms: A case study of steemit. In *Proceedings of the 10th ACM conference on web science*. 145–154.

[81] Franklin Mingzhe Li, Cheng Lu, Zhicong Lu, Patrick Carrington, and Khai N. Truong. 2022. An Exploration of Captioning Practices and Challenges of Individual Content Creators on YouTube for People with Hearing Impairments. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 75 (apr 2022), 26 pages. doi:10.1145/3512922

[82] Jiahui Li. 2023. Social Media Engagement: Can Video Captions Increase User Engagement?. In *Proceedings of the 3rd International Conference on Economic Development and Business Culture (ICEDBC 2023)*, Vol. 258. Springer Nature, 103.

[83] Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. 2023. ASR and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition. *arXiv preprint arXiv:2305.16065* (2023).

[84] Ann Light and Yoko Akama. 2014. Structuring future social relations: the politics of care in participatory practice. In *Proceedings of the 13th Participatory Design Conference: Research Papers-Volume 1*. 151–160.

[85] Ziyue Lin, Yi Shan, Lin Gao, Xinghua Jia, and Siming Chen. 2025. SimSpark: Interactive Simulation of Social Media Behaviors. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–32.

[86] Deborah L Linebarger. 2001. Learning to read from television: The effects of using captions and narration. *Journal of educational psychology* 93, 2 (2001), 288.

[87] Kim Lyons. 2020. YouTube is ending its community captions feature and deaf creators aren't happy about it. *The Verge* (31 jul 2020). https://www.theverge.com/2020/7/31/21349401/youtube-community-captions-deaf-creators-accessibility-google Accessed: 2025-09-07.

[88] Yao Lyu, He Zhang, Shuo Niu, and Jie Cai. 2024. A Preliminary Exploration of YouTubers' Use of Generative-AI in Content Creation. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[89] Kelly Mack, Danielle Bragg, Meredith Ringel Morris, Maarten W. Bos, Isabelle Albi, and Andrés Monroy-Hernández. 2020. Social App Accessibility for Deaf Signers. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2, Article 125 (October 2020), 31 pages. doi:10.1145/3415196

[90] Catherine MacPhail, Nomhle Khoza, Laurie Abler, and Meghna Ranganathan. 2016. Process guidelines for establishing intercoder reliability in qualitative studies. *Qualitative research* 16, 2 (2016), 198–212.

[91] Jessica Maddox and Fiona Gill. 2023. Assembling "sides" of TikTok: Examining community, culture, and interface through a BookTok case study. *Social Media+ Society* 9, 4 (2023), 20563051231213565.

[92] Aaron Marcus and Emilie West Gould. 2000. Crosscurrents: cultural dimensions and global Web user-interface design. *interactions* 7, 4 (2000), 32–46.

[93] Peter L Markham. 1993. Captioned television videotapes: Effects of visual support on second language comprehension. *Journal of Educational Technology Systems* 21, 3 (1993), 183–191.

[94] Alberto Martín-Martín, Enrique Orduna-Malea, Mike Thelwall, and Emilio Delgado López-Cózar. 2018. Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of informetrics* 12, 4 (2018), 1160–1177.

[95] Lloyd May, Keita Ohshiro, Khang Dang, Sripathi Sridhar, Jhanvi Pai, Magdalena Fuentes, Sooyeon Lee, and Mark Cartwright. 2024. Unspoken Sound: Identifying Trends in Non-Speech Audio Captioning on YouTube. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642162

[96] Emma J McDonnell, Tessa Eagle, Pitch Sinlapanuntakul, Soo Hyun Moon, Kathryn E Ringland, Jon E Froehlich, and Leah Findlater. 2024. "Caption It in an Accessible Way That Is Also Enjoyable": Characterizing User-Driven Captioning Practices on TikTok. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[97] Emma J. McDonnell and Leah Findlater. 2024. Envisioning Collective Communication Access: A Theoretically-Grounded Review of Captioning Literature from 2013–2023. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS '24)*. Association for Computing Machinery, St. John's, NL, Canada. doi:10.1145/3663548.3675649

[98] Emma J McDonnell, Ping Liu, Steven M Goodman, Raja Kushalnagar, Jon E Froehlich, and Leah Findlater. 2021. Social, environmental, and technical: Factors at play in the current use and future design of small-group captioning.

*Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.

[99] Cristina Miguel, Carl Clare, Catherine J Ashworth, and Dong Hoang. 2024. Self-branding and content creation strategies on Instagram: A case study of foodie influencers. *Information, Communication & Society* 27, 8 (2024), 1530–1550.

[100] Angela Molem, Stephann Makri, and Dana Mckay. 2024. Keepin'it Reel: Investigating how short videos on TikTok and Instagram reels influence view change. In *Proceedings of the 2024 Conference on Human Information Interaction and Retrieval*. 317–327.

[101] National Captioning Institute. 2020. History of Closed Captioning. https://www.ncicap.org/history-of-cc. Accessed: 2025-11-07.

[102] Netflix. 2018. Timed Text Style Guide: General Requirements. https://partnerhelp.netflixstudios.com/hc/en-us/articles/215758617-Timed-Text-Style-Guide-General-Requirements Accessed: 2025-09-03.

[103] Susan B Neuman and Patricia Koskinen. 1992. Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading research quarterly* (1992), 95–106.

[104] PR Newswire. 2025. ShareChat & Moj Announce Second Edition of Short Form Big Impact Leadership Summit 2025. https://www.prnewswire.com/in/news-releases/sharechat--moj-announce-second-edition-of-short-form-big-impact-leadership-summit-2025-302515985.html Accessed: 2025-08-29.

[105] Mikel K Ngueajio and Gloria Washington. 2022. Hey ASR system! Why aren't you more inclusive? Automatic speech recognition systems' bias and proposed bias mitigation techniques. A literature review. In *International conference on human-computer interaction*. Springer, 421–440.

[106] Malcolm J. Norwood. 1988. Captioning for Deaf People: An Historical Overview. In *Speech to Text: Today and Tomorrow*, J. E. Harkins and B. M. Virvan (Eds.). Gallaudet Research Institute, Washington, DC. https://dcmp.org/learn/80-captioning-for-deaf-people-an-historical-overview Proceedings of a Conference at Gallaudet University.

[107] William Fielding Ogburn. 1922. *Social Change with Respect to Culture and Original Nature*. B. W. Huebsch, Inc., New York, NY.

[108] Paraskevi Panagi, Alexandros Yeratziotis, Thomas Fotiadis, Christos Mettouris, and George Angelos Papadopoulos. 2024. "Signifier" Video Sharing Platform and Accessible Media Player for Deaf Users. In *Proceedings of the 2024 International Conference on Information Technology for Social Good*. 151–157.

[109] Becky Parton. 2016. Video captions for online courses: Do youtube's auto-generated captions meet deaf students' needs? *Journal of Open, Flexible, and Distance Learning* 20, 1 (2016), 8–18.

[110] Sahil Patel. 2016. *85 Percent of Facebook Video Is Watched Without Sound*. https://digiday.com/media/silent-world-facebook-video/ Accessed: 2025-09-02.

[111] Rikki Poynter. 2016. #NoMoreCraptions: How To Properly Captions Your Videos. https://www.youtube.com/watch?v=-O4YcVQt5NM Accessed: 2025-09-03.

[112] Noor Prasetio and Neneng Sri Wahyuningsih. 2023. An Analysis of the Error Translation in Movie Trailers by YouTube Auto-Translate. *Eligible: Journal of Social Sciences* 2, 2 (2023), 264–278. doi:10.53276/eligible.v2i2.81

[113] Jorge Pérez-Martín, Alejandro Rodriguez-Ascaso, and Elisa M. Molanes-López. 2021. Quality of the captions produced by students of an accessibility MOOC using a semi-automatic tool. *Universal Access in the Information Society* 20, 4 (2021), 677–690. doi:10.1007/s10209-020-00740-9

[114] Leyuan Qu, Taihao Li, Cornelius Weber, Theresa Pekarek-Rosin, Fuji Ren, and Stefan Wermter. 2023. Disentangling prosody representations with unsupervised speech reconstruction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023), 39–54.

[115] Prajit T Rajendran, Kevin Creusy, and Vivien Garnes. 2024. Shorts on the rise: assessing the effects of YouTube shorts on long-form video content. *arXiv preprint arXiv:2402.18208* (2024).

[116] Mary Anne Raymond, Hillary Smith, Les Carlson, and Aditya Gupta. 2024. An Examination of Digital Accessibility Within Social Media Platforms: Problems for Vulnerable Consumers and Policy Implications. *Journal of Advertising Research* 64, 4 (2024), 430–450. doi:10.2501/JAR-2024-026

[117] Microsoft Research. 2020. What Does Access Mean? Re-thinking the Accessibility Research Problems That We Tackle. https://www.youtube.com/watch?v=3f_lVpAAFco Video from the Accessible Computer Science Education Fall Workshop; Accessed: 2025-09-03.

[118] Jill Walker Rettberg. 2017. Hand signs for lip-syncing: The emergence of a gestural language on Musical. ly as a video-based equivalent to emoji. *Social Media+ Society* 3, 4 (2017), 2056305117735751.

[119] Rev Press. 2021. The Ultimate Roundup of Compelling Closed Captions Statistics. https://www.rev.com/blog/ultimate-roundup-closed-captions-statistics Published: May 18, 2021. Accessed: 2025-08-16.

[120] James A Roberts and Meredith E David. 2025. Technology Affordances, Social Media Engagement, and Social Media Addiction: An Investigation of TikTok, Instagram Reels, and YouTube Shorts. *Cyberpsychology, Behavior, and Social Networking* 28, 5 (2025), 318–325.

[121] Yasmin Rufo. 2025. *TikTok award winner Max: 'It takes me hours to make the content you scroll on the toilet'*. https://www.bbc.com/news/articles/c17p1z5ppq8o BBC News, Published 13 November 2025.

[122] Abdul Saleem, Rashid Mehmood, Ayman Taj, Asadullah Lakho, et al. 2024. Impact of Video Content Marketing on Consumer Engagement. *Journal of Policy Research* 10, 3 (sep 2024), 83–95. doi:10.61506/02.00322 License: CC BY 4.0.

[123] Peggy Anne Salz. 2022. Striking Gold in the Global Livestreaming Creator Economy: A Q&A With BIGO Technology's Mike Ong. https://www.forbes.com/sites/peggyannesalz/2022/03/08/striking-gold-in-the-global-livestreaming-creator-economy-a-qa-with-bigo-technologys-mike-ong/ Accessed: 2025-08-29.

[124] Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. 2021. Re-imagining algorithmic fairness in india and beyond. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 315–328.

[125] Burçin Sarı. 2025. The Rise of Influencer Practices Among Psychologists: From Therapy Rooms to Instagram Reels. *Social Media+ Society* 11, 2 (2025), 20563051251353741.

[126] Douglas Schuler and Aki Namioka (Eds.). 1993. *Participatory Design: Principles and Practices*. CRC Press, Hillsdale, NJ.

[127] Filipo Sharevski, Oliver Alonzo, and Sarah Hau. 2025. "I Have Never Seen That for Deaf People's Content:" Deaf and Hard-of-Hearing User Experiences with Misinformation, Moderation, and Debunking on Social Media in the US. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 1–17. doi:10.1145/3706598.3713114

[128] Brent N. Shiver and Rosalee J. Wolfe. 2015. Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS '15)* (Lisbon, Portugal). Association for Computing Machinery, 231–238. doi:10.1145/2700648.2809857

[129] Ben Shneiderman. 2000. Universal usability. *Commun. ACM* 43, 5 (2000), 84–91.

[130] Abdulhadi Shoufan and Fatma Mohamed. 2022. YouTube and education: A scoping review. *IEEE Access* 10 (2022), 125576–125599.

[131] Ellen Simpson, Samantha Dalal, and Bryan Semaan. 2023. " Hey, Can You Add Captions?": The Critical Infrastructuring Practices of Neurodiverse People on TikTok. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–27.

[132] Ellen Simpson and Bryan Semaan. 2025. The Enshittification of the Creative Internet. *Proceedings of the ACM on Human-Computer Interaction* 9, 7 (2025), 1–24.

[133] Chad Smith, Tamby Allman, and Samantha Crocker. 2017. Reading between the Lines: Accessing Information via" Youtube's" Automatic Captioning. *Online Learning* 21, 1 (2017), 115–131.

[134] N Sooryah and KR Soundarya. 2020. Live Captioning for Live Lectures-An Initiative to Enhance Language Acquisition in Second Language Learners, through Mobile Learning. *Webology* 17, 2 (2020).

[135] South China Morning Post. 2020. Singapore-based Likee, led by a former factory worker, is gaining ground on TikTok. https://kr-asia.com/singapore-based-likee-led-by-a-former-factory-worker-is-gaining-ground-on-tiktok Accessed: 2025-08-29.

[136] Carli Spina. 2021. Video accessibility. *Library Technology Reports* 57, 3 (2021), 0024–2586.

[137] Susan Leigh Star and Anselm Strauss. 1999. Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer supported cooperative work (CSCW)* 8, 1 (1999), 9–30.

[138] Statista Research Department. 2025. Most Popular Social Networks Worldwide as of July 2025, Ranked by Number of Monthly Active Users. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/ Accessed: 2025-08-15.

[139] Thomas Steiner, Hannes Mühleisen, Ruben Verborgh, Pierre-Antoine Champin, Benoît Encelle, and Yannick Prié. 2014. Weaving the Web (VTT) of Data.. In *LDOW*.

[140] Chunmeizi Su. 2023. *Douyin, TikTok and China's online screen industry: The rise of short-video platforms*. Routledge.

[141] Xinru Tang, Xiang Chang, Nuoran Chen, Yingjie Ni, RAY LC, and Xin Tong. 2023. Community-driven information accessibility: Online sign language content creation within d/deaf communities. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–24.

[142] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, Dirk Hovy, Shannon Spruit, Margaret Mitchell, Emily M. Bender, Michael Strube, and Hanna Wallach (Eds.). Association for Computational Linguistics, Valencia, Spain, 53–59. doi:10.18653/v1/W17-1606

[143] Feng Teng. 2019. Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos. *Computer Assisted Language Learning* 32, 7 (2019), 665–691.

[144] U.S. Congress. 2010. Twenty-First Century Communications and Video Accessibility Act of 2010 (CVAA). https://www.govinfo.gov/content/pkg/PLAW-111publ260/pdf/PLAW-111publ260.pdf Accessed: 2025-09-03.

[145] María Azahara Veroz-González and Pilar Castillo Bernal. 2024. Automatic Closed Captions and Subtitles in Academic Video Presentations: Possibilities and Shortcomings. *Complutense Journal of English Studies* 32, e94649 (2024). doi:10.5209/cjes.94649

[146] Melanie Walker and Alejandra Boni. 2020. *Participatory research, capabilities and epistemic justice: A transformative agenda for higher education.* Springer Nature.

[147] Shaofu Wang. 2020. *Personal branding strategies of female entertainment influencers on TikTok.* Rochester Institute of Technology.

[148] Xueying Wang, Meng Chen, and Wei Jiang. 2024. Why is one social media platform not enough? A typology of platform-swinging behavior and associated affordance preferences. *Social Media+ Society* 10, 2 (2024), 20563051241254373.

[149] Yiwen Wang, Ziming Li, Pratheep Kumar Chelladurai, Wendy Dannels, Tae Oh, and Roshan L Peiris. 2023. Haptic-captioning: using audio-haptic interfaces to enhance speaker indication in real-time captions for deaf and hard-of-hearing viewers. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[150] We Are Social and Meltwater. 2025. Digital 2025: Global Digital Report. https://wearesocial.com/wp-content/uploads/2025/02/GDR-2025-v2.pdf Accessed: 2025-08-15.

[151] Jacob O Wobbrock and Julie A Kientz. 2016. Research contributions in human-computer interaction. *interactions* 23, 3 (2016), 38–44.

[152] World Wide Web Consortium (W3C). 2019. WebVTT: The Web Video Text Tracks Format. https://www.w3.org/TR/webvtt1/ W3C Candidate Recommendation. Editor: Silvia Pfeiffer. Accessed: 2025-09-03.

[153] Wyzowl. 2025. Video Marketing Statistics 2025 (11 Years of Data). https://wyzowl.com/video-marketing-statistics/ Accessed: 2025-08-15.

[154] Peter Yang. 2020. The Cognitive and Psychological Effects of YouTube Video Captions and Subtitles on Higher-Level German Language Learners. In *Technology and the Psychology of Second Language Learners and Users*. Springer International Publishing, 83–112. doi:10.1007/978-3-030-34212-8_4

[155] Suhyeon Yoo, Georgianna Lin, Hyeon Jeong Byeon, Amy S. Hwang, and Khai Nhut Truong. 2023. Understanding Tensions in Music Accessibility through Song Signing for and with d/Deaf and Non-d/Deaf Persons. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, 1–18. doi:10.1145/3544548.3581287

[156] Kyrie Zhixuan Zhou, Weirui Peng, Yuhan Liu, and Rachel F. Adler. 2025. Exploring the Diversity of Music Experiences for Deaf and Hard of Hearing Individuals. *Proceedings of the ACM on Human-Computer Interaction* 9, 2 (2025), 1–28. doi:10.1145/3710976

# A Full List of 36 Papers

| No. | Title | Authors & Year | Venue |
|---|---|---|---|
| 1 | "Caption It in an Accessible Way That Is Also Enjoyable": Characterizing User-Driven Captioning Practices on TikTok | McDonnell et al. (2024) | CHI |
| 2 | "Hey Can You Add Captions?": The Critical Infrastructuring Practices of Neurodiverse People on TikTok | Simpson et al. (2023) | CSCW |
| 3 | An Exploration of Captioning Practices and Challenges of Individual Content Creators on YouTube for People with Hearing Impairments | Li et al. (2022) | CSCW |
| 4 | An Examination of Digital Accessibility Within Social Media Platforms | Raymond (2024) | Journal of Advertising Research |
| 5 | The Effects of Captions on L2 Learners | Aldukhayel (2021) | Language Learning & Technology |
| 6 | Maximizing the Potential of Captions for Primary School ESL Students' Comprehension of English-Language Videos | Teng (2019) | Computer Assisted Language Learning |
| 7 | Beyond Traditional Subtitles: How Emojis and Non-Standard Typography in Subtitles Boost Engagement on TikTok | Duraj and Szarkowska (2025) | Journal of Audiovisual Translation |
| 8 | A Qualitative Study of Closed Captions in English Language Teaching (ELT) YouTube Videos | Hernandez et al. (2024) | Journal of English & Applied Linguistics |
| 9 | An Analysis of the Errors in the Auto-Generated Captions of University Commencement Speeches on YouTube | Lee et al. (2020) | The Journal of Asia TEFL |
| 10 | Toward Language Justice: Exploring Multilingual Captioning for Accessibility | Desai et al. (2025) | CHI |
| 11 | Reading Between the Lines: Accessing Information via YouTube's Automatic Captioning | Smith et al. (2017). | Online Learning |
| 12 | Video Captions for Online Courses: Do YouTube's Auto-Generated Captions Meet Deaf Students' Needs? | Parton (2016) | Journal of Open, Flexible & Distance Learning |
| 13 | "Voices Help Correlate Signs and Words": Analyzing Deaf and Hard-of-Hearing TikTokers' Content, Practices, and Pitfalls | Cao et al. (2024) | CHI |
| 14 | DHH Users' Prioritization of Genres of Online Video Content Requiring Accurate Captions | Berke et al. (2020) | W4A |
| 15 | Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia | Shiver & Wolfe (2015). | ASSETS |
| 16 | "I Have Never Seen That for Deaf People's Content": DHH User Experiences with Misinformation, Moderation, and Debunking on Social Media | Sharevski et al. (2025) | CSCW |
| 17 | Leveraging Complementary Contributions of Different Workers for Efficient Crowdsourcing of Video Captions | Huang et al. (2017) | CHI |
| 18 | Shifting the Focus: Exploring Video Accessibility Strategies and Challenges for People with ADHD | Jiang et al. (2025) | CHI |
| 19 | Effects of Captioned Videos on Learners' Comprehension | Guo et al. (2020) | Journal of Global Literacies, Technologies, and Emerging Pedagogies |
| 20 | The Visual Experience of Accessing Captioned Television and Digital Videos | Butler (2019) | Television and New Media |
| 21 | Quality of the captions produced by students of an accessibility MOOC using a semi-automatic tool | Pérez-Martín (2020) | Universal Access in the Information Society |
| 22 | Comparing the Impact of Professional and Automatic Closed Captions on Video-Watching Experience | Kim et al. (2023) | CHI |
| 23 | An Analysis of the Error Translation in Movie Trailers by YouTube Auto-Translate | Prasetio & Wahyuningsih (2023) | Journal of Social Sciences |
| 24 | Judging Translation On Social Media: A Pragmatic Look At Youtube Comment Section | Kraeva & Krasnopeyeva (2020) | The European Proceedings of Social and Behavioural Sciences |
| 25 | Automatic Closed Captions and Subtitles in Academic Video Presentations: Possibilities and Shortcomings | Veroz-González & Bernal (2024) | Complutense Journal of English Studies |
| 26 | Effect of Occlusion on DHH Users' Perception of Captioned Video Quality | Amin et al. (2021) | UAHCI |
| 27 | Exploring the Diversity of Music Experiences for Deaf and Hard-of-Hearing Individuals | Zhou et al. (2025) | CSCW |
| 28 | Beyond Subtitles: Captioning and Visualizing Nonspeech Sounds to Improve Accessibility of User-Generated Videos | Alonzo et al. (2022) | ASSETS |
| 29 | Unspoken Sound: Identifying Trends in Non-Speech Audio Captioning on YouTube | May et al. (2024) | CHI |
| 30 | OnomaCap: Making Non-speech Sound Captions Accessible and Enjoyable through Onomatopoeic Sound Representation | Kim et al. (2025) | CHI |

*Continued on next page*

| No. | Title | Authors & Year | Venue |
|---|---|---|---|
| 31 | Hand Signs for Lipsyncing The Emergence of a Gestural Language on Musically as a Video-Based Equivalent to Emoji | Rettberg (2017) | Social Media + Society |
| 32 | Signifier Video Sharing Platform and Accessible Media Player | Panagi et al. (2024) | GoodIT |
| 33 | Social App Accessibility for Deaf Signers | Mack et al. (2020) | CSCW |
| 34 | Community-Driven Information Accessibility: Online Sign Language Content Creation within d/Deaf Communities | Tang et al. (2023) | CSCW |
| 35 | Sparkling Silence: Practices and Challenges of Livestreaming Among Deaf or Hard-of-Hearing Streamers | Cao et al. (2023) | CHI |
| 36 | Understanding Tensions in Music Accessibility Through Song Signing for and With d/Deaf and Non-d/Deaf Persons | Yoo et al. (2023) | ASSETS |