Unspoken Sound: Identifying Trends in Non-Speech Audio Captioning on YouTube

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

Keita Ohshiro
ko89@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Khang Dang
khd3@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Sripathi Sridhar
ss645@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Jhanvi Pai
jp295@njit.edu
New Jersey Institute of Technology
Newark, New Jersey, USA

Magdalena Fuentes
mfuentes@nyu.edu
New York University
New York, New York, USA

Sooyeon Lee
sooyeon.lee@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

ABSTRACT

High-quality closed captioning of both speech and non-speech elements (e.g., music, sound effects, manner of speaking, and speaker identification) is essential for the accessibility of video content, especially for d/Deaf and hard-of-hearing individuals. While many regions have regulations mandating captioning for television and movies, a regulatory gap remains for the vast amount of web-based video content, including the staggering 500+ hours uploaded to YouTube every minute. Advances in automatic speech recognition have bolstered the presence of captions on YouTube. However, the technology has notable limitations, including the omission of many non-speech elements, which are often crucial for understanding content narratives. This paper examines the contemporary and historical state of non-speech information (NSI) captioning on YouTube through the creation and exploratory analysis of a dataset of over 715k videos. We identify factors that influence NSI caption practices and suggest avenues for future research to enhance the accessibility of online video content.

CCS CONCEPTS

• **Human-centered computing → Accessibility systems and tools**; *Accessibility technologies*; Empirical studies in accessibility.

KEYWORDS

datasets, closed captioning, non-speech information, extra-speech information, subtitles

ACM Reference Format:
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 (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3613904.3642162

1 INTRODUCTION

With over 500+ hours of video uploaded to YouTube every minute [70], online video has become a major medium of communication for entertainment, education, news, and more. However, with information presented in both auditory and visual modalities, online video on its own is not accessible to all people, e.g., the large worldwide d/Deaf and hard of hearing (DHH) population.

Captions can improve the accessibility of video for DHH audiences [71]. Captions are a textual version of the speech and non-speech audio information in the video and provide critical content to DHH viewers [32]. Closed captioning (i.e., captions that can be voluntarily turned on/off) debuted on United States network television in March 1980 [23], and over the next twenty years, US legislation was passed to require televisions to include caption decoders [20] and broadcasters to caption most of their content [21]. In 2010, additional legislation was passed that requires most previously broadcast content to be captioned when redistributed over the internet on official channels [19]. Yet, the vast majority of video content on the internet (e.g., YouTube) does not meet this criteria and thus is not mandated to be captioned by law. Without such regulation, many videos may remain uncaptioned due to the considerable time required to manually caption content.

In recent years, advances in automatic speech recognition (ASR) technology, which enables the transcription of spoken words, have facilitated the captioning process and thus increased the presence of captions on video-sharing and streaming platforms. As captions have become more prevalent, practitioners and researchers have strived to improve captions, thereby enhancing the DHH viewers' access to and overall enjoyment of video content. For

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642162

example, researchers have investigated various factors influencing the perceived quality of captions, including transcription accuracy [13], ASR-generated caption accuracy [37], caption visibility/legibility [14, 66], caption timing [63], and caption location/position [9–11, 33, 36, 41, 58]. To facilitate this progress, researchers have investigated metrics for the evaluation of the speech caption quality [6].

However, captions are much more complex than just speech transcriptions. Captions also communicate non-speech audio information (NSI), and while the quality and implementation of captions of speech audio information have been greatly improved, the status of NSI captions on video-sharing and streaming platforms is far from adequate. NSI includes information about non-speech sounds such as environmental sounds, sound effects, incidental sounds, and music, as well as additional narrative information and extra-speech information (ESI), which gives context to spoken or signed language such as manner of speech (e.g. "[Whispering] Oh no") or speaker label (e.g. "[Juan] Oh no"). NSI seems to be often overlooked even though NSI is often critical for understanding video content. Current NSI captioning of online video seems scarce and insufficient for the needs of the DHH viewers compared to the quality and availability of the captions of speech information, but how scarce is it? Is NSI captioning becoming more prevalent? How is the rise of ASR affecting NSI captioning? Recent studies have begun investigating NSI [7], however, such studies have been limited to topics such as speaker identifier [12, 27, 42, 64], manner of speech, and prosodic and emotional element of the speech [22, 45] — no research currently exists to understand current NSI captioning practices of online video. In this paper, we seek to address this oversight and to understand the current state of NSI captioning practices on the most popular online video-sharing platform, YouTube. In particular, we aim to answer the following research questions:

(1) What is the current and historical prevalence of non-speech information captioning on YouTube?
(2) What factors may affect non-speech information captioning practices on YouTube?

To answer these questions and better understand the current state of NSI on YouTube, we created a dataset of YouTube videos and captions spanning a decade, estimated and manually annotated NSI within this dataset, and conducted a quantitive exploratory analysis of the data. Our analysis not only facilitates a deeper understanding of NSI captioning of online video, but it also identifies multiple paths forward to improve NSI captioning and thus increase the accessibility of online video content.

2 BACKGROUND AND RELATED WORK

2.1 Non-Speech Information (NSI)

Captions that contain non-speech information (NSI) are those that include information about all audio events except for the actual words used in verbal or signed communication. NSI includes sound effects (e.g. "[Lightsabers swooshing]"), environmental sounds (e.g. "[Crickets chirping]"), music (e.g. "[Upbeat jazz guitar]"), as well as captions that provide extra-speech information including who is speaking and the manner in which they are speaking (e.g. "[Thandi whispers] I love you").

The history and evolution of closed captions have given precedence to verbal or signed communication over NSI [23]. Sean Zdenek, a theorist of rhetoric and captioning, has pointed out the *logocentrism* in captioning, where speech is often given priority over NSI, providing many examples of cases where barely audible background speech is captioned instead of narratively important NSI, such as excluding "[Single gunshot]" and only captioning the words softly mumbled "Oh, no", even though the gunshots are more narratively important and that both could likely be captioned simultaneously. Despite numerous qualitative studies that reveal what DHH viewers determine to be relevant ambient sounds, existing closed captioning does not convey robust sound scenes [35, 48, 51]. *YouTube* first introduced automatic captioning at scale in 2009, with no capabilities to caption NSI [30].

Several automatic captioning systems have been developed and deployed since then. In 2013, Naim et al. tackled the shortcomings of combining real-time crowd-sourced captions by aligning multiple captions on a word-by-word basis, providing a more efficient and cheaper alternative to automatic captioning for many content creators at the time [56]. The impact of ASR systems on understanding for real-time classroom captioning was further studied at the University of Washington in 2016 [38]. Other models look to improve existing ASR systems; Wald discusses the advantage of crowd-sourcing to correct errors in speech and sound recognition while Liu et al. experiment with cross-modal audio-visual attention mapping to use relevant information from the visual scene to describe ambiguous sound events [49, 65]. Prior work has explored how to improve auto-generated captions, including adding punctuation based on audio context for increased readability [29].

While these systems have undoubtedly had positive impacts, the lack of human vetting and editing of the generated captions, which often contain inaccuracies, have earned them the nickname "autocraptions" by members of the DHH community [25]. *YouTube* expanded its automatic captioning system through the addition of NSI, what it called "sound effect information", in 2017 [18]. The NSI captioned was limited to the classes of "[Laughter], [Applause], and [Music]", providing the rationale that "while the sound space is obviously far richer and provides even more contextually relevant information than these three classes, the semantic information conveyed by these sound effects in the caption track is relatively unambiguous, as opposed to sounds like [RING], which raises the question of 'what was it that rang — a bell, an alarm, a phone?' " Upon further examination of these claims, the inclusion of music as a "relatively [semantically] unambiguous" class of NSI appears to be inaccurate. The infamous "[Music]" caption provides little more semantically meaningful information than "[RING]'. The ambiguity present in deciding which NSI should be captioned and what exact language should be used has contributed to the practice of captioning NSI lagging behind speech in both automatic and manual captioning contexts. Recently, models in the machine listening community have been developed for automated audio captioning (AAC) that aim to describe environmental sound scenes with natural language [24, 54]. While these models seem relevant for automated and semi-automated NSI captioning pipelines, current AAC models have not been designed to meet the NSI captioning needs of DHH audiences — current models caption audio independently of any other context, but the NSI needs of DHH audiences are dependent on complex relationships of sound, visuals, narrative, speech content, audience hearing abilities, and audience preferences that affect what sounds should be captioned and how they should be captioned [7].

Captions on YouTube include a mix of manually generated captions, edited automatically generated captions, and unedited automatically generated captions. Up until 28 September 2020, YouTube allowed community members to submit captions to the video creator who could review or edit the captions before adding them to the video. YouTube discontinued this service stating that it was "underutilized", although research found that 46.8% of surveyed creators leveraged community caption services when they were available [47, 50]. The captioning practices on YouTube greatly affect DHH viewer's experience on the platform. Li et al. identified several frustrations with captioning practices on YouTube among DHH viewers and creators [47]. These included caption tracks with severe grammatical and punctuation errors, incomplete captions such as "[Joke]" rather than the actual words of the joke, incomplete NSI information such as "[Music]" with no additional detail, and the practice of adding additional, potentially confusing in-jokes not related to the video content in the video's captions.

2.2 Current Captioning Best Practices

In the United States, television broadcasters and distributors are legally required by the Federal Communications Commission (FCC) to ensure content released contains captions that are "Accurate... Synchronous... Complete... [and] Properly placed" [2, 68]. While there are no explicit guidelines outlined by the FCC on best captioning practices for NSI, under the "Accurate" subheading, it is mandated that captions "...convey background noises and other sounds to the fullest extent possible." The current state of NSI captioning can perhaps be attributed to a lack of clear and extensive guidelines on how to caption NSI, as the FCC is the only administrative agency that can hold media providers responsible for the quality of their captions produced for public viewing.

Best practices for captioning of media on the internet and other user-generated content are relatively ambiguous. Under current FCC guidelines, "full-length internet video programming" that has previously been aired on TV in the US is required to be captioned. Yet, this is only regulated when clips from videos are distributed on the programming distributor's own website. Clips found on third-party websites, such as YouTube, are not legally required to provide closed captioning, limiting access to online media content to many individuals [1].

Despite the lack of legally binding guidelines for professional captioning online, there have been attempts at suggestions for captioning NSI and style guides for categories of NSI, such as sustained versus discrete sound effects. For instance, Zdenek's captioning theory and guideline book *Reading Sounds* provides nuanced guidance on how to enhance the quality and legibility of NSI captions. Zdenek highlights the complexities present in captioning NSI, stating that "captioning is not simply about the sounds themselves but about the relationships among sounds, images, and the audience's presumed knowledge (or cultural literacy). A sound should be analyzed not only in terms of its sonic and contextual salience within a scene but also in terms of its visual and cultural salience to the audience" [71]. The *World Wide Web Consortium (W3C)*, the main organization governing web standards, recommends that NSI be captioned when it states that captions should "include all dialogue and equivalents for non-dialogue audio information needed to understand the program content, including sound effects, music, laughter, speaker identification, and location" [17]. The *Described and Captioned Media Program's Captioning Key*, administered by the National Association of the Deaf, further outlines the importance of including background music and onomatopoeic descriptions when captioning non-speech sound [53].

Much like how the style of NSI captioning varies from one content creator to another, there are various opinions and personal preferences for customization in the visual display of captions among DHH users. It is largely agreed upon that the contrast between the captions and background should be high enough to be readable and that captions should not obstruct critical visual components of the screen. However, there is no clear consensus on the color and typeface of text, caption background color, and opacity, the number of lines of captioning, caption placement, as well as personal preference for genre-adaptive caption typeface and movement [8, 15].

2.3 Sound Communicating Technologies for Communicating Non-Speech Information

Previous work has explored communicating specific facets of NSI and ESI through modifications of captions, or through novel sound communication technologies (SCTs). Alonzo et al. explored NSI communication in the context of user-generated content through the addition of graphic captions and icons [7]. They highlighted that, while text is efficient at communicating information precisely and with potentially less visual distraction, it often does not include the desired level of temporal and contextual information. Dynamically altering properties of the text, such as size, placement, and typeface, have been shown to effectively communicate sonic attributes such as loudness [67]. Kushalnagar et al. emphasize the difficulty in communicating NSI through captions due to the simultaneous delivery of aural and visual content and lack of standardization in representing non-speech information. They explored the benefit of incorporating vibrotactile haptics as a way to augment communication of NSI [43].

Preferences for speaker identification in a fast-paced, multi-speaker setting have been researched to effectively communicate ESI in non-interactive audio-visual media [28]. Moreover, the communication of contextual vocal expression, such as sarcasm, and perceived emotional content based on a speaker's tone of voice has been previously explored through the formatting of text, the inclusion of additional punctuation and/or icons as indicators, and adjustments to the visual design of the text [22, 26, 31, 45]. Dynamic movement and positioning of letters to indicate sonic and affective qualities as well as speaker identity location have been extensively investigated to improve ESI captioning [16, 40, 61]. However, these technologies have yet to be evaluated for longer periods and have not yet seen large-scale implementation.

3 DATA COLLECTION AND ANNOTATION

3.1 Sourcing Video Captions and Metadata

To analyze the contemporary and historical state of NSI captioning on YouTube, we developed a dataset consisting of two different samples of videos: a *popular* video sample and a *studio* video sample. The popular sample aimed to understand the captioning practices in a broad spectrum of popular, impactful videos on YouTube. In contrast, the studio sample was more targeted. It sought to examine captioning practices among the top-tier production houses, often viewed as industry benchmarks due to their influence and vast resources available for accessibility. Furthermore, many of these production houses produce content that is legally mandated by the FCC to be captioned when broadcast on television in the United States, some of which may overlap with their online content. Therefore, given the differences in the production, captioning, and distribution processes, this distinction is made to more clearly understand the trends and possible accessibility improvements to be made in these two different systems of content production. For both samples, we retrieved videos from every month spanning the years 2013 to 2022 and limited our selection to videos primarily from the US, in English, and presented in a standard 2D format.

3.1.1 *Selection Criteria for Popular Sample.*

For the popular sample, we employed the `search().list` method from YouTube's Data API to source the most popular videos by view count for each of YouTube's 32 video categories[1]. Content creators assign one of these categories to each of their videos upon upload, thus in this categorization scheme, each video is associated with only one category, reflecting the creators' perception of their content. For each of these categories, up to 500 of the most popular videos were retrieved for each month for the 2013–2022 period.

3.1.2 *Selection Criteria for Studio Sample.*

Videos in this category were sourced from 25 YouTube channels owned by leading production companies each valued at over 1 billion USD, as estimated by Forbes [55]. The channels supplying captions for the studio sample include:

- Comcast: NBC, Universal Pictures, Peacock Kids, SyFy, MSNBC.
- Walt Disney: Disney, Pixar, ESPN, LucasFilm, Marvel Entertainment.
- Paramount Global: Paramount Pictures, Showtime, Comedy Central, Nickelodeon, BET.
- Warner Bros. Discovery: Warner Bros. Pictures, Discovery, DC Entertainment, HBO Max, CNN.
- Fox Corporation: Fox Sports, Fox News, Fox Weather, Fox Nation, Fox Business.

Up to 500 (but typically far fewer) of the most popular videos in each channel were retrieved month-wise from 2013 to 2022 using the YouTube Data API.

3.1.3 *Captions Retrieval.*

The extraction of captions from YouTube videos was conducted through two primary methods: the YouTube Data API (v3) and the YouTube-DL tool[2]. The YouTube Data API facilitates data extraction by querying the `search().list` method and allows for filtering and identification of videos with captions using YouTube's search API. On the other hand, the YouTube-DL tool permits the specification of the preferred language (English) and the option to download only the captions, excluding video content. Notably, the `extract_info` method in this tool aids in pinpointing videos with captions.

Both manually created captions (by human captioners) and automatically generated captions (by YouTube) were downloaded. In cases where videos offered multiple caption modes, redundancy was minimized by retaining only one. Preference was given in the order: en, en-US, en-[other countries], followed by caption modes CC1 or DTVCC1, which are primarily utilized for caption displays.

3.1.4 *Metadata Retrieval.*

An essential facet of our methodology was the acquisition of metadata. For both sample types, we utilized the YouTube videos API. By querying this API with individual video IDs from our curated lists, we were able to amass pertinent metadata for our study. This metadata includes high-level YouTube-assigned topics as retrieved by the YouTube Data API[3]. In contrast to the user-assigned categories that we used to construct the popular sample, these topics are automatically assigned by YouTube using topic tagging models and each video may have more than one topic associated with it. Since users may choose a category based on criteria other than the best match to their content (e.g., search engine optimization), we group videos based on YouTube-assigned topic in our analysis rather than user-assigned category. We simplified the 62 YouTube-assigned topics in the dataset to 24 high-level topics as defined in the mapping in Appendix B. In our analysis, we use this smaller set of topics to facilitate analysis and visualization.

3.2 Automatic Retrieval of Estimated NSI Captions

To minimize annotation efforts and to aid in analysis at scale, we defined an "estimated NSI criteria". This involved extracting all non-alphanumeric characters while excluding typical punctuation such as !?,'. Following this, we manually examined the captions of 500 videos from each sample to discern indicators that signified the captions as NSI.

We augmented our list of NSI criteria with a list of known indicators, constructed on best practice recommendations from the Described and Captioned Media Program[4] (DCMP) and the World Wide Web Consortium[5] (W3C), along with considering practices from leading studio creators:

- **NSI Descriptors**: These often include language identifiers, sound effects, paralanguage, and manner of speaking identifiers. They're frequently encapsulated within [ ] or ( ), in line with standards set by DCMP and W3C which suggest using lowercase text. Examples include:
  - Language markers, such as "(speaking French)".
  - Identifiers for the manner of speaking, e.g., "(whispers) Don't go!".
  - Standalone NSI indicators like "(grunts in alarm)" for paralanguage.

[1] https://developers.google.com/youtube/v3/docs/videoCategories/list
[2] https://ytdl-org.github.io/youtube-dl/
[3] https://developers.google.com/youtube/v3/docs/videos
[4] https://dcmp.org/captioningkey/print
[5] https://www.w3.org/WAI/WCAG21/Understanding/

- **Speaker Identifiers**: These highlight the speaker, especially when they're off-screen or when clarity is needed. Typically, these are in uppercase letters, followed by a colon, as in "Narrator: This is the island of New Penzance."
- **Musical Elements**: Musical markers can denote song titles, lyrics, descriptions of music, or even musical notes. For instance, "[♪♩♪]" or "♩ Searchin' for light in the darkness ♪".
- **Channel Identifiers**: These denote the medium of sound or communication. An example might be "[Woman over PA]: Your attention, please.". These often embed keywords such as 'over' or 'on' within brackets.
- **Formatting Best Practices**: Off-screen sound effects are conventionally italicized. Sustained sounds adopt the present participle form, like "[dog barking]", whereas sudden sounds use the third person verb form, i.e., "[dog barks]". Environmental sounds/effects sometimes amalgamate with onomatopoeia, leading to captions like "[doorbell ringing] ding-dong".

With these guidelines in place, we delineated indicators to assist in identifying estimated NSI captions. Our NSI criterion encompasses the following symbols: <, |, (,), [,], “”, >, ♩, , ♫, 𝅝, ♭, ♮, ♯, #, and :, and our 'NSI estimator' estimates that any caption containing these symbols has NSI.

3.3 Manual Annotation

To validate our NSI estimator and classify instances of NSI into different NSI types, we manually annotated a subset of the full dataset. We defined the subset by randomly selecting 300 videos with estimated NSI from each sample in 2013, 2018, and 2022.

3.3.1 Annotation Labels. Our NSI labels were designed to cover a spectrum of non-speech information (NSI) found within captions. The following list illustrates our labeling strategy, based on types of NSI previously identified in the literature [52]:

- **Not NSI:** Captions that do not contain any NSI.
- **Music:** Any genre of music, whether diegetic or not.
- **Environmental Sounds, Sound Effects, and Incidental Sounds:** Non-music and non-speech sounds. This includes non-verbal vocalizations like laughter, grunts, and crying, provided they aren't used to modify speech.
- **Extra-speech Information:** Text that gives added context to spoken or signed language.
- **Additional Narrative Information:** Descriptive text that doesn't pertain directly to sounds.
- **Quoted Speech:** Captions containing internal quotation marks. This label is used when there's uncertainty about its current NSI status, prompting a possible revisit.
- **Unsure, Misc, or Ambiguous:** For instances where the appropriate label is unclear or the caption doesn't fit current categories.
- **Non-English Captions:** Used for captions not written in English and subsequently excluded from further annotation.

3.3.2 Annotation Process. We randomly partitioned the video data into three equally sized sets. Two members of our team independently annotated the captions on each set using the labels defined in Section 3.3.1. Annotators could assign multiple labels to each caption when appropriate. Annotations were mainly based on the captions, but to ensure precision, links to the original videos along with start and end times for each caption were provided. This allowed the annotators to refer back to the video for auditory and visual confirmation when uncertainty arose, especially when deciphering complex sounds or understanding context.

Upon the completion of the primary annotations, the results from both annotators were revisited to resolve any discrepancies or disagreements. During instances of uncertainty about the interpretation of a sound, the video was consulted to validate the correct label. This rigorous approach minimized errors and ensured the highest level of consistency. Discussions were held throughout the annotation process, and our labeling criteria were refined as necessary to better accommodate the data.

4 YOUTUBE NSI CAPTIONING DATASET

The resulting YouTube NSI Captioning Dataset consists of ~715k videos with a total of ~273M lines of captions, ~6M of which are estimated instances of NSI. These videos span 10 years and 21 topics. The dataset consists of two samples: 1) *popular* which contains a broad spectrum of popular, impactful videos and 2) *studio* which contains videos from production houses with vast resources available for accessibility. Within the full dataset, NSI has been identified using our NSI estimator (see Section 3.2), and a subset of the dataset has NSI manually annotated by the research team. This annotated subset consists of 1799 videos with a total of ~36k annotated captions lines, ~114k of which are instances of NSI annotated on 7 different categories. These videos span 3 years (2013, 2018, and 2022) and 20 YouTube-assigned topics. Each video has annotations by 2 annotators along with the consensus annotation. The dataset contains the links to the YouTube videos, the captions, the video metadata from the YouTube API, and the annotations described above. The temporal and topic distributions of both the full dataset and annotated subset are in the Appendix in Figures 8 and 9. The dataset is available here: https://doi.org/10.5281/zenodo.10681804.

5 MEASURES OF NON-SPEECH INFORMATION IN CAPTIONS

We define four measures of NSI in captions: 1) *NSI presence*, 2) *NSI count per minute if present (NSI CPMIP)*, 3) *Estimated NSI presence* and 4) *Estimated NSI count per minute if present (Estimated NSI CPMIP). NSI presence* is defined as the proportion of videos in which we have manually identified the presence of NSI captions. *Estimated NSI presence* is defined as the proportion of videos in which our NSI estimator has identified the presence of NSI captions. *NSI CPMIP* and *Estimated NSI CPMIP* are measures of the density of NSI captions. For a set of videos $V = \{(v_{i,\text{count}}, v_{i,\text{duration}}) \mid i = 1, 2, \ldots, n\}$, CPMIP is defined as:

$$\text{CPMIP} = \operatorname*{Median}_{v_i \in V'} \left(\frac{v_{i,\text{count}}}{v_{i,\text{duration}}} \right) \tag{1}$$

where $V' = \{(v_{i,\text{count}}, v_{i,\text{duration}}) \in V \mid v_{i,\text{count}} > 0\}$ is the subset of V where at least one instance of NSI captioning is present in each video, $v_{i,\text{count}}$ is the count of NSI for video v_i, and $v_{i,\text{duration}}$ is the duration in seconds for video v_i. NSI CPMIP uses counts of manually

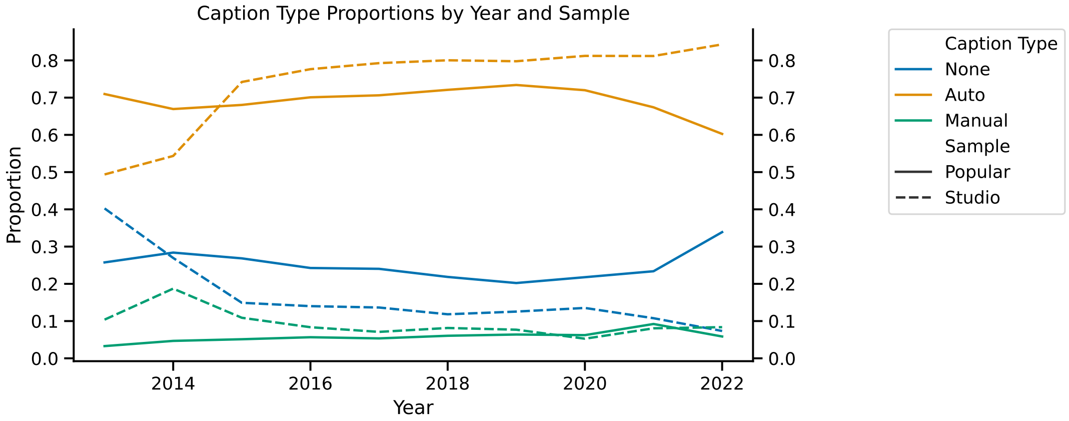

Figure 1: The proportion of caption generation types (including the *None* type for videos without any captions) by year and sample.

identified NSI and Estimated NSI CPMIP uses the NSI estimator defined in Section 3.2. (Estimated) NSI presence is intended to be a measure of the prevalence of NSI captions, and (Estimated) NSI CPMIP is intended to be a measure of the quality of NSI captions based on the assumption that more frequently captioned NSI within a video is an indicator of better NSI captioning.

6 ANALYSIS

6.1 Trends in General Captioning Practices

We begin our analysis by investigating how content creators typically generate captions in 2022 and how this practice has changed over the past decade. Recall that the popular sample aims to understand the captioning practices in a broad spectrum of popular videos, while the studio sample seeks to examine captioning practices among the top-tier production houses, often viewed as industry benchmarks due to their influence and vast resources available for accessibility. In Figure 1, we plot the proportion of captions that are automatically generated (Auto), manually generated (Manual), or absent (None) in both our popular and studio samples from our full dataset. We find that content creators predominately use automatic captioning algorithms in their contemporary video captioning practice (60% and 84% for the popular and studio samples respectively), followed by not captioning videos at all (34% and 7% respectively), and lastly followed by manually generated captions, which make up just 6 and 8 percent of captions for the popular and studio samples respectively in 2022. When we look at the temporal trends, we see that in the studio sample, the use of automatically generated captions has increased significantly during the decade of analysis (from 49% in 2013 to 84% in 2022), primarily corresponding to a reduction in uncaptioned videos (from 40% in 2013 to 7% in 2022) but also with a smaller reduction in manually captioned videos (from a peak of 19% in 2014 to 8% in 2022). The distribution of caption generation methods was relatively more stable in the popular sample, exhibiting a decrease in automatically generated captions over time (from 71% in 2013 to 60% in 2022), an increase in videos without captions (from 26% in 2013 to 33% in 2022), and an in increase in manual captioning (from 3% in 2013 to 6% in 2022).

Prior to 2017, automatically generated captions did not contain any NSI. However, starting in 2017, YouTube's automatically generated captions have included some NSI captions, but they are limited to *[laughter]*, *[music]*, and *[applause]*. For richer, more informative NSI captions, content authors currently must manually generate or edit captions. Thus, while overall captioning in the studio sample has dramatically increased over the past decade, the number of manually-created captions that more likely have informative NSI has actually decreased. Whereas for the popular sample, overall captioning has slightly decreased, but manually captioning increased by a factor of 2. In the remainder of the paper, we focus our analysis only on videos with manually created captions since only those videos have the rich NSI information that we are interested in.

6.2 Trends in NSI Captioning Practices

To investigate the current state of NSI captioning and recent trends, we first computed estimated NSI presence and estimated NSI CPMIP on the full dataset. We find that the estimated NSI presence and estimated NSI CPMIP are roughly the same for automatically generated captions in both the popular and studio samples in 2022 (see Figure 2a and 2b), as we would expect since both samples use the same algorithm. However, for manually generated captions in 2022, we find that the popular video sample has both greater estimated NSI presence and estimated NSI CPMIP. Looking at this trend in manual captions over time (see Figure 3), we find that the difference in estimated NSI presence between the popular and studio samples has persisted for the past decade, with an oscillating pattern occurring in both samples, but we find the trend in estimated NSI CPMIP is more complicated — the popular sample exhibits an oscillating pattern, but the studio sample has had a decline by a factor of 3.2 over the past decade (from 2.79 in 2013 to 0.87 in 2022).

The analysis of NSI thus far has been dependent on our NSI estimator. To validate our estimator and investigate NSI captioning practices in greater detail, we computed NSI presence and NSI

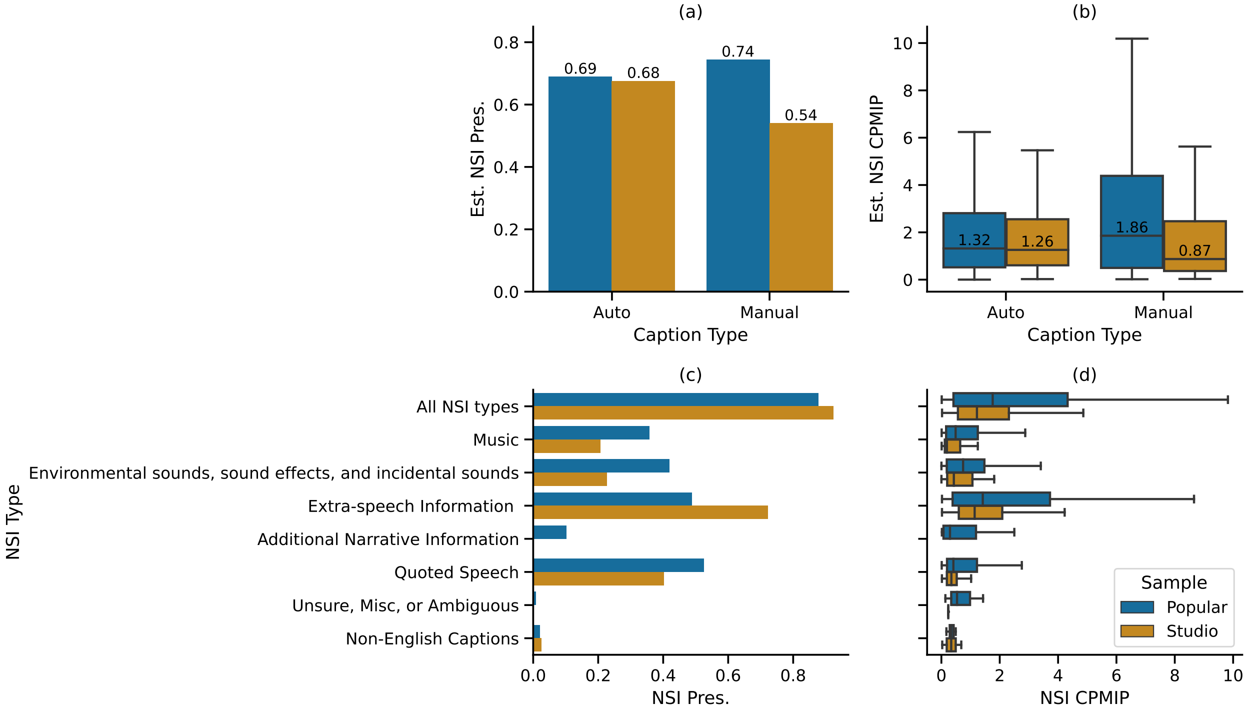

Figure 2: Top: The (a) mean estimated NSI presence and (b) median estimated NSI CPMIP for auto and manual caption generation types for both the *popular* and *studio* 2022 samples in the full dataset. Bottom: The (c) mean NSI presence and (d) median NSI CPMIP for all annotated NSI types in both the *popular* and *studio* 2022 samples of the annotated subset.

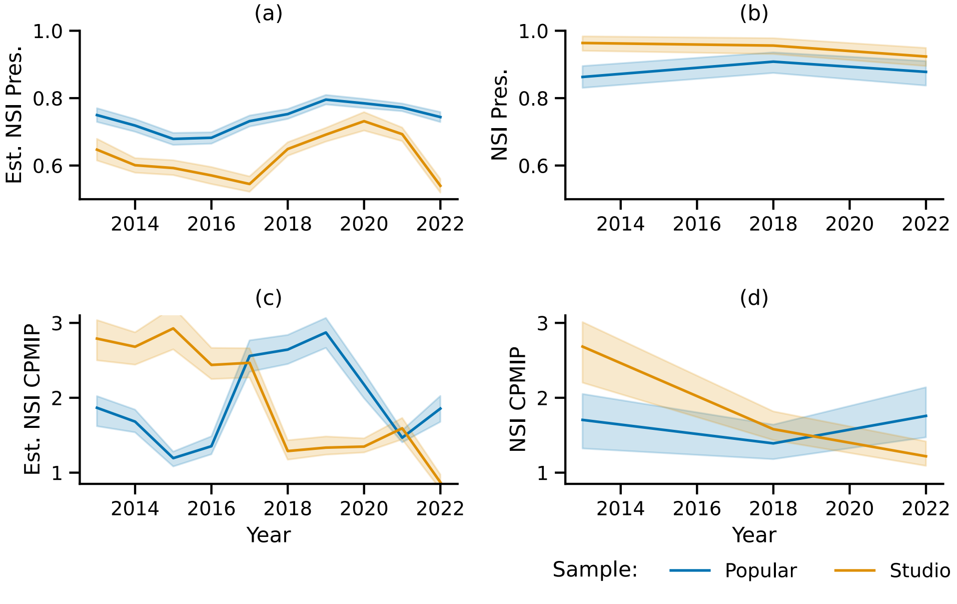

Figure 3: Left: The (a) mean estimated NSI presence and (c) median estimated NSI CPMIP by year for manual captions in both the *popular* and *studio* samples in the full dataset. Right: The (b) mean NSI presence and (d) median NSI CPMIP by year for both the *popular* and *studio* samples in the annotated subset (i.e., 2013, 2018, 2022).

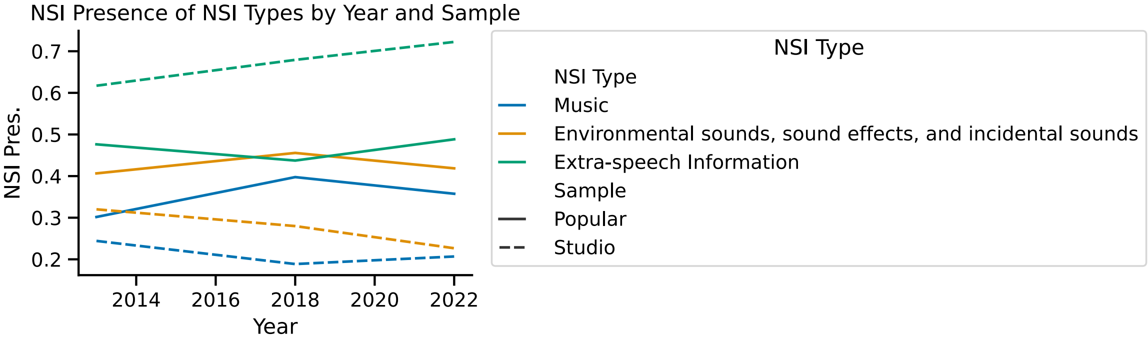

Figure 4: NSI Presence of NSI types by year for both the *popular* and *studio* samples in the annotated dataset (i.e., 2013, 2018, 2022)

CPMIP on our annotated dataset, which is limited to manually captioned videos with estimated NSI. We find that *All NSI types* (i.e., presence of any annotated NSI) has an NSI presence value of 0.88 and 0.92 for the popular and studio samples respectively, which is equivalent to the precision of our estimator. Thus, the simple estimator is of relatively high quality, but it slightly overestimates NSI in the popular sample more so than the studio sample. This results in the studio sample exhibiting a slightly higher NSI presence than the popular sample in 2022, a trend which holds over the past decade as shown in Figure 3b. However, the observed trends in estimated NSI CPMIP seem to still hold true in NSI CPMIP, with the studio sample NSI CPMIP still dropping significantly over the past decade (from 2.68 in 2013 to 1.22 in 2022) as shown in Figure 3d. Furthermore, when we break NSI presence and NSI CPMIP down by different NSI types (see Figure 2c and 2d) in 2022, we find that the increase in NSI presence in the studio sample over the popular sample is due to a large difference in *Extra-speech Information* presence (popular: 0.49; studio: 0.72), rather than from *Music* (popular: 0.36; studio: 0.21) or *Environmental sounds, sound effects, and incidental sounds* (popular: 0.42; studio: 0.23), which are actually much lower in the studio sample. We find that NSI CPMIP for all types of NSI is lower in the studio sample than the popular sample and that for both samples, *Extra-speech information* has the highest NSI CPMIP. When we look at these trends over time (see Figure 4), we find that *Extra-speech information* has been increasing (from 0.62 in 2013 to 0.72 in 2022) in the studio sample, while *Environmental sounds, sound effects, and incidental sounds* and *Music* has been decreasing (Music: from 0.24 in 2013 to 0.21 in 2022; Environmental sound: from 0.32 in 2013 to 0.23 in 2022).

6.3 Factors Affecting NSI Presence and Density

To understand what factors affect the presence and density of NSI captioning, we investigated the relationship between our measures of NSI and video popularity, video topic, and video duration.

6.3.1 Video popularity. In Figure 5a, we plot video view counts for manually-captioned videos with and without estimated NSI and find that the popularity of a video does not seem to have a relationship to the presence of NSI in a video. However, when we look at the relationship of view count to caption generation type (see Figure 5b), we find that videos with manually-generated captions tend to be more popular, followed by videos with automatically-generated captions and, lastly, videos without captions.

6.3.2 Video topic. We also computed the NSI measures on both samples of the full and annotated dataset broken down by our simplification of YouTube-assigned topic. We found that all topics seem to have some videos with NSI in them, but the proportion with NSI (i.e., NSI presence) does vary by topic, as does the NSI CPMIP (see Appendix Figures 10 and 11 for details). Thus, the presence and density of NSI is dependent on the topic. We looked into the effect of topic further by computing the entropy of the NSI type distribution for each topic in the annotated dataset (see Figure 6a) — this provides an indication of the diversity of NSI types within each topic, e.g., if there is primarily one NSI type present in a particular topic (an example of low entropy) or if all NSI types were equally present in a topic (an example of high entropy). We found that there was considerable entropy variation among topics, with the *Lifestyle* (2.08), *Music* (1.98), and *Food* (1.93) topics having the highest entropy, and *Sports* (0.60), *Military* (0.72), and *Humour* (0.99) having the lowest entropy. When we looked at the mean entropy averaged over topics that were common to both the popular and studio samples in the three years of analysis (see Figure 6b), we found that mean entropy was quite similar for the two samples in 2013 and 2018, but has begun to diverge slightly in 2022 with the studio sample having lower entropy than the popular sample (popular: 1.58; studio: 1.23).

While NSI presence showed slight variation between topics, NSI CPMIP differed substantially with topics such as 'Religion' and 'Music' having well over four times the CPMIP compared to 'Military' and 'Politics' videos. While the distribution of types of NSI captions (e.g. music, sound effects) has not appeared to change substantially over time, this distribution differs greatly by topic. Additionally, the entropy of NSI caption types varied substantially by topic where, for example, 'Lifestyle' and 'Music' videos had substantially more NSI caption type diversity when compared to 'Sports' and 'Military' videos. Given these factors impacted by topic, future captioning analysis research and development of novel technologies must therefore consider the influence of topic/genre.

6.3.3 Video duration. To investigate the relationship between video duration and NSI caption density, we analyzed the manually captioned videos in the full dataset that had at least one NSI caption.

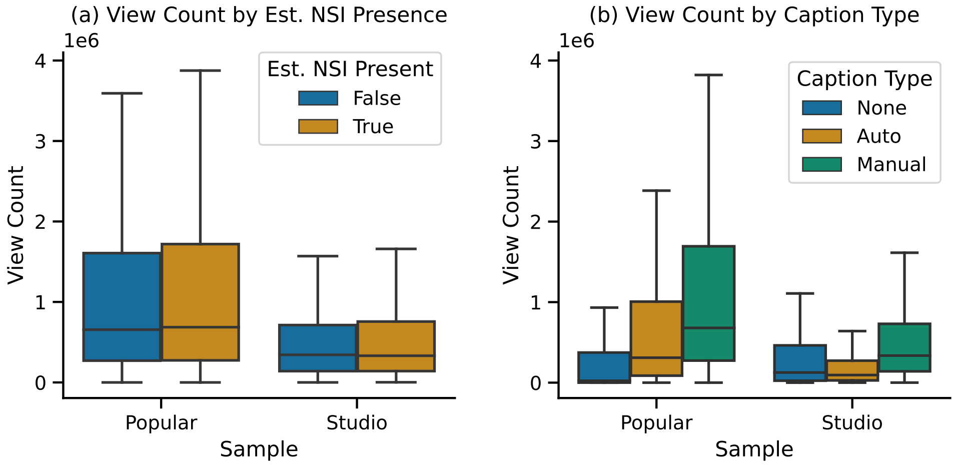

Figure 5: (a) YouTube video view counts by estimated NSI presence for both the *popular* and *studio* 2022 samples of videos with manually-generated captions. (b) YouTube video view counts by caption generation type for both the *popular* and *studio* 2022 samples.

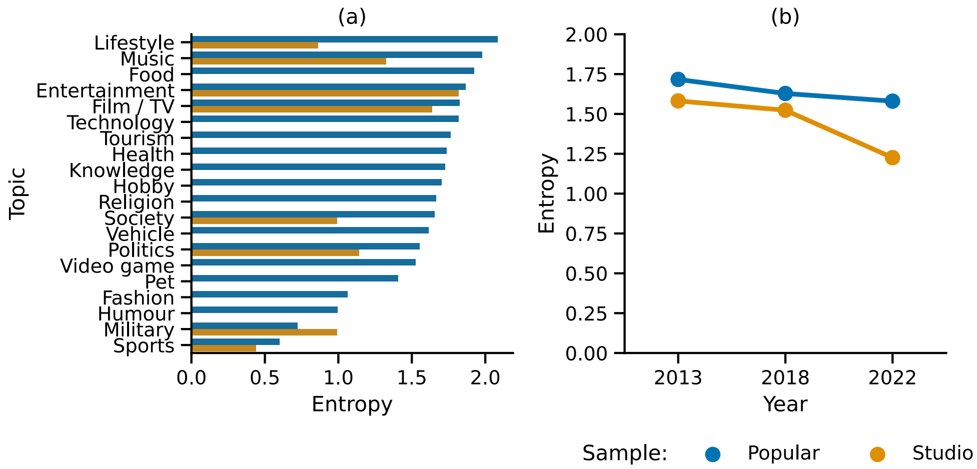

Figure 6: (a) Entropy of NSI type distribution by YouTube-assigned topic for both *popular* and *studio* 2022 samples in the annotated subset. (b) The mean entropy of NSI type distributions by year for both *popular* and *studio* samples in the annotated subset, with each year averaged over the 6 YouTube assigned topics common to both samples in all years. Low entropy indicates a focus on fewer NSI types per topic and thus less diversity of NSI types per topic, whereas high entropy indicates a more distributed focus across NSI types per topic and thus more diversity of NSI types per topic.

We construct three distinct video duration groups for our analysis: '< 1 min', '1 - 10 min', and '> 10 min'. Notably, at the time of dataset construction, none of the videos tagged as *YouTube Shorts* in the YouTube API contained manual captions. Therefore all videos analyzed here are regular YouTube videos and not *YouTube Shorts*, which are typically displayed in a vertical 9:16 ratio and have their own placement and dedicated sub-systems within YouTube's search and recommendation systems. We found that estimated NSI CPMIP tends to decrease as video duration increases for both samples (see Figure 7a). We performed a two-way aligned rank transform (ART) ANOVA [69] to analyze the effect of the sample and duration group on estimated NSI CPMIP. It revealed a statistically significant main effect for duration group (F(2,33859)=627.76, p<0.001), but did not show a statistically significant main effect for sample (F(1,33859)=0.04, p=0.85) nor a statistically significant interaction between sample and duration group (F(2,33859)=2.04, p=0.13). In a post hoc ART comparison test, we found all contrasts of duration groups to be statistically significant (p<0.001). We also found

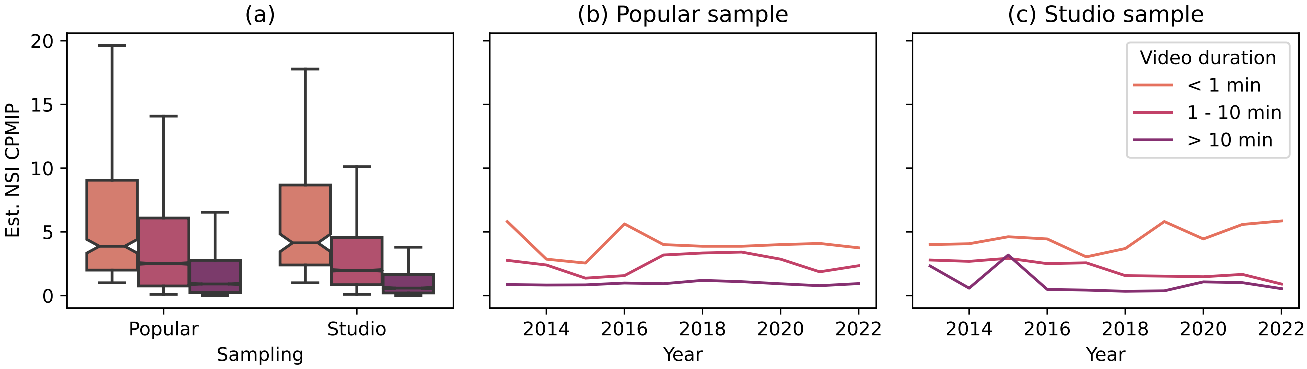

Figure 7: (a) Estimated NSI CPMIP for all years by duration group and sample (notches are 95% CIs). Median estimated NSI CPMIP by year and duration group for videos in the (b) popular sample of the full dataset and (c) the studio sample of the full dataset. All three subfigures are limited to manually captioned videos that contain at least one NSI caption.)

a statistically significant negative Spearman rank correlation between estimated NSI CPMIP and video duration, both in the studio (r=-0.51, p<0.001) and popular samples (r=-0.35, p<0.001). Lastly, when we look at the median estimated NSI CPMIP over time for each duration group (see Figure 7b and c), we see the trend that the rank order of the duration groups by estimated NSI CPMIP is largely maintained for both samples in all years except for in 2015 when the two longest duration groups were about equal.

7 DISCUSSION

7.1 Manually Captioned Videos Have More Views

The view count for manually captioned videos from the popular sample had higher view counts than those with automatic or no captions, as seen in Figure 5. Additionally, the presence of NSI in the captions appeared to have no clear relationship to view count, further suggesting that the mode of captioning, manual or automatic, has more impact than the quality or quantity of NSI captioning. This trend may have been explained by creators before, such as creators choosing to manually caption only videos that are receiving a high number of views or creators only manually captioning videos they believe are of a high quality. However, the causality of this trend may also operate in the reverse direction. Many non-YouTube affiliated resources online claim that manually adding captions or a video transcript to YouTube improves search engine optimization (SEO) for the video, claiming that this additional text resource allows search engines and recommendation algorithms to more accurately index the video and therefore recommend the video [46].

High-quality manual captions may also be an attractive feature to some viewers, and essential to some viewers having access to the videos at all, and therefore videos with, presumably, higher-quality manual captions are simply accessible to and enjoyed by more people, notably DHH audiences who have expressed dissatisfaction with the quality of captioning on YouTube [59]. YouTube additionally allows users to filter search results based on whether they have a caption track that was not automatically generated by YouTube. Some content creators additionally add “[CC]” or “#Captioned” to their video titles or descriptions to communicate that they believe their video possesses high-quality captioning [47]. On the reverse side, videos that use sensitive non-advertising-friendly words, such as racist or sexist language, slurs, or swear words, can be flagged and demonetized by YouTube and may not be recommended to as many viewers. This flagging process can be done on the basis of automatically generated captions, which may falsely identify and caption sensitive words or phrases in a video that does not contain any [47]. Therefore, content creators who manually caption their videos may avoid this flagging if they ensure their caption tracks do not contain sensitive words or phrases, potentially resulting in more views for content with manual captions.

While overall captioning rates have decreased in the popular sample, the rate of manual captioning has increased two-fold. However, the percentage of videos in the popular sample with no captions at all increased by nearly 10% from 2021 to 2022. With the rise of automatic caption use in the videos in the studio sample, it is unclear why such a significant amount of content creators would not opt to use automatic captions, instead, explicitly opting out of having automatically generated captions on their videos. Content creators may view automatic captioning as insufficiently accurate or distracting to their videos, and some may have additional concerns about how these potentially inaccurate representations of their video content may impact their videos' SEO. While currently, creators have to actively opt out of having automatically generated captions appear as an option on their videos, it is unclear if accounts from different regions or of different ages go through this same process. For example, an account that existed long before automatic captioning was a feature on YouTube may have this new feature turned off by default, and these creators have simply not opted-in to this. Future research could explore the various factors that influence video creators to not include automatically generated closed captions in their videos.

7.2 Differences in Large Studio Captioning Practices

When manually captioned, videos from the popular sample had a 20% higher rate of NSI presence compared to videos from the studio sample using the NSI estimate. While this trend was narrowly reversed in the annotated dataset, the overwhelming majority of NSI captioning in the studio sample consisted of ESI, such as speaker labels. Additionally, the median CPMIP rate in the popular sample videos was over twice as high as the large studio sample in both the estimate and annotated analyses. Large studios are increasingly using YouTube's automatic captioning system over adding manual captions, as seen in Figure 1. While this has led to a notable decrease in uncaptioned videos, the quality of automatic captions is notably lower, particularly with regard to NSI. However, even within large studios' manual captions, the CPMIP of NSI captions has exhibited a 3-fold decline over the past ten years. Lastly, the diversity of types of NSI captioning seen in each video, captured by entropy in Figure 6, is consistently higher in the popular sample when compared to the studio sample across all annotated years as well as all topics with the exception of 'Military' videos. Extra-speech information (e.g. speaker labels and manner of speech indicators) had the highest presence and CPMIP in both the popular and studio samples. Interestingly, some videos in the popular sample used captions in creative capacities that did not describe audio events but rather used the captions to add additional commentary, contextual information, narration, or short summary reminders. This use of creative captions was not seen in videos in the studio sample.

These trends in the large studio sample may be a result of increased reliance on post-editing of ASR captioning tools in the process of manual captioning. In a 2023 survey on captioning with over 300 respondents across a variety of industries, over 40 percent of respondents reported using post-edited ASR outputs for captions [5] — this is up from 27 percent in 2017 as reported in a similar survey with over 1400 participants [4]. Notably these automated tools may detect NSI at a much lower rate, if at all, which may lead captioners who are editing ASR captions to focus only on speech content and caption NSI at a lower rate as well. Therefore, automated captioning tools that more frequently and accurately detect and caption NSI may help increase the density of NSI captions, even in manually captioned videos. Further research into current professional captioners' practices is needed to more deeply understand the role of current ASR captioning tools in the trend of decreasing NSI CPMIP in manual captions on large studios' YouTube videos, as well as how future automated NSI captioning tools could be meaningfully incorporated into the manual captioning process.

Several notable patterns of inaccuracies and errors in captions on large studios' YouTube videos were also noted. There were timing errors and delays present in many of the professional studio captions, with the captions being delayed by over ten seconds in some cases [6]. This may be because live CART-style captions are used when the content is being broadcast live on TV, and these captions are not time-corrected before being uploaded to YouTube. Additionally, some captioning tracks on the large studio videos appear to be gibberish, containing a series of random, uncorrelated letters [7]. This may be due to a technical error with digital TV (DTV) captioning protocols, caption display commands, or raw input from a stenographic keyboard not being processed correctly. Both the timing errors and gibberish captioning tracks illustrate that the captioning practices and general production and distribution processes of large studios are complex.

Therefore these studios may require caption processing solutions and quality assurance strategies that are quite different from other content creators. Additionally, the majority of content produced by these studios is legally required to be proficiently captioned while being broadcast on TV or large streaming platforms. However, the clips that are selected by the studios to be edited and uploaded to YouTube may have their captioning files either lost, as evidenced by the over 90% of large studio videos with no manual captions, or introduce errors in the timing and quality of the captions during this process. Additionally, video content with scripted dialogue may be more common in media produced by large studios and can leverage auto-aligning features based on the dialogue script or video transcript [47].

7.3 Video Duration Influences NSI Caption Density

There is a weak but significant correlation between NSI CPMIP and video duration, in both the studio and popular samples. Notably, in the full 2022 dataset, videos shorter than one minute had a median CPMIP approximately 5 times higher than videos longer than 10 minutes.

This may be due to the fact that shorter videos simply require less time and effort to be captioned compared to longer videos, therefore captioners, particularly amateur captioners, are less likely to experience fatigue while captioning, which may lead to more NSI being captioned. Additionally, NSI may be more important to the narrative and intelligibility in shorter videos that have less time to establish context. Therefore creators may find it more necessary or advantageous to caption NSI in these videos. Even though none of the analyzed videos were tagged *YouTube Shorts*, the rise in popularity of short-form video and the resulting stylistic trends established for these videos may impact content creators' captioning practices. Lastly, the shorter video content itself may be relatively more NSI-dense as a result of stylistic trends in short-form video content.

7.4 Presence and Density of Manually Annotated NSI Is Generally Low

Based on our analysis, only 6-8% of videos in our sample have manual captions, and of those 88-92% have any form of NSI, resulting in just 5-7% of videos containing any NSI outside of the 3 NSI labels that YouTube's automatic captioning algorithm provides. Of the videos that have NSI, the density also seems low (and is decreasing for the studio sample, from 2.68 in 2013 to 1.22 in 2022). For an approximate reference of expected CPMIP in "high-quality" captions, we refer to Zdenek's book *Reading Sounds* [71] in which he published NSI statistics for four Hollywood films released on DVD. From his published data, we found the mean CPMIP for these

[6] Large timing delay example: https://www.youtube.com/watch?v=zYh6Rw5DSbY

[7] Gibberish caption example: https://www.youtube.com/watch?v=b9A4Js4G2q0

releases was 3.03. In addition, much of the NSI is extra-speech information, particularly speaker labels. Music and environmental sound NSI are present in only 3-6% and 2-3% of videos respectively and at quite a low density as well — at less than 1 per minute when present. At face value, these values appear low for most content, but further research is needed to understand how that compares to "high-quality" NSI captions by respected captioners in the community, and how we can design tools to aid and inform captioners in their NSI captioning practices. Additionally, these findings highlight the need for improved captioning tools to encourage manual captioning or editing of automated captions, as well as improved tools for automatic captioning of NSI.

7.5 Implications of Findings on Improving NSI Captioning

After analyzing the data, several insights emerged that indicate possible points of improvement in captioning system design as well as avenues for future HCI research related to NSI captioning.

Improved Evaluation of NSI Captioning Quality: In this study, even videos from the large studio sample were often not manually captioned, and upon manual inspection and annotation, it became clear that these videos could not be used as a meaningful benchmark for NSI captioning best practice. While several metrics exist to evaluate the accuracy of speech captioning such as word error rate and the NER metric [57, 62], no such metrics exist for NSI captioning. NSI captioning is often contextual and opinions regarding what NSI should be captioned and the desired level of detail varies among DHH viewers [52], increasing the difficulty of establishing an NSI captioning quality metric. Combined with little research systematically exploring the effect of NSI captioning practices on DHH viewers' experience, the need for frameworks or tools to evaluate NSI captioning quality that centered around DHH experiences is clear. While the development of a standardized metric would be greatly beneficial to the field, video-sharing sites such as YouTube present an additional opportunity for community editing and feedback tools for NSI captions that allow viewers to rate NSI captions as well as suggest edits.

The utilization of community-based collaboration on information accessibility has been studied and shown to be effective in other marginalized communities and is an ongoing area of research and design exploration. For example, Audio Description (AD) researchers working with viewers who are blind or low-vision (BLV) around the accessibility of video media have investigated collaborative AD creation tools, utilizing community-supported editing and evaluating approaches, and their findings suggest this is an effective way for creating AD that better meets the needs of BLV folks [3, 39]. Similarly, researchers also studied the effectiveness of crowd workers' collective evaluation of visual descriptions of images by making suggestions and voting for the best descriptions [44]. Therefore, similar community-based editing and authoring synergies can be leveraged to improve the quality of the automatically generated speech and NSI captions for DHH viewers.

Additionally, if automatic models are used in the process of caption generation, errors flagged by the community and user ratings could be used to improve the models as well as give feedback to the content creators themselves. Again, a similar pattern is seen in the AD space where authoring and automatic editing tools for AD make use of feedback [34, 60], and automatic speech and NSI captioning models could leverage the same design patterns.

More Nuanced Automatic NSI Captioning: Given the widespread use of automatic captioning in both popular and large studio videos identified by this study and the severe limitations of current automatic NSI captioning on YouTube, it is clear that automatic NSI captioning requires improvement to provide DHH viewers more equitable access to video content that is automatically captioned. Specifically, there is a clear need for the incorporation of automated audio captioning with increased NSI labeling capabilities designed to meet the needs of DHH audiences. This should include additional classes of NSI beyond music, applause, and laughter, and should include additional descriptive attributes, such as the genre of music, the amount of laughter, etc. Additionally, these systems could provide viewers with increased agency to select if they would like automatically generated NSI captions to be displayed so that if the system is not meeting their needs, it will not increase distraction.

Increase Ease of Manual Editing of Automatically Generated Captions: The evidence that shorter videos tended to have more NSI captions and were more likely to be manually captioned supports previous findings that manual captioning and editing can cause fatigue in content creators [47]. Therefore, tools that generate automatic captions could make the process of manual editing more efficient to reduce captioning fatigue. To this end, these tools could communicate the confidence the model has in the accuracy of each of the generated speech or NSI captions, or other tools could be created to highlight areas of the video in which automatic speech and NSI captioning systems may be most inaccurate. These metrics could then be used to improve the manual editing process of automatic captions by inviting captioners to spend time on editing the most potentially problematic areas first. Asking content creators to correct the captions that have the highest likelihood of being incorrect may lead to higher-quality captions on more videos, particularly in videos of longer duration.

Additional Transparency of How Captions Affect Video Recommendations: Video distribution platforms, like YouTube, could improve transparency around how manual and automatic captioning factors into their recommendation and demonetization systems. This clarity might encourage more creators to enable automatic captioning on their videos if they knew how that might affect the video's view counts. Additionally, platforms could educate creators about the importance of captioning and accessible content creation practices, as well as the limitations of automatic captioning systems, such as difficulty with certain accents, limited NSI captioning, etc.

Improved Media Processing Pipelines and Caption Alignment for Large Studios: This study highlighted that over 80% of videos produced by large studios leverage automatic captioning instead of using manual caption tracks, which likely already exists for most media shared on these channels such as short excerpts of clips that were previously broadcast on television in the USA, and were therefore legally required to be captioned. Therefore, high-quality, legally compliant captions likely exist for these videos but are somehow separated from the video files during the process of editing and sharing these videos on YouTube [8]. Given the difference in scale and media processing pipelines, large studios require different strategies to improve the quality of captioning of videos on their YouTube channels. The first improvement would be a reworking of media processing pipelines so that captioning tracks are not separated from their original videos while being edited for YouTube distribution. Additionally, automated time-alignment tools may be helpful to correct the delays between dialogue and captions often seen in videos where the captions were originally generated in real-time, such as sports commentary clips. While the delays at the time of live-broadcaster are understandable, once these clips are shared in a fixed format, such as a YouTube video, time correcting the captions to align with the video's sound appears to be a relatively easy way to greatly increase the quality of captions on YouTube channels owned by large studios.

8 LIMITATIONS

While this work provides a first step toward understanding the accessibility of NSI information in YouTube videos, our approach has some limitations. First of all, we are limited by the retrieval capabilities of the YouTube API which restricts our analysis to videos that are easily retrievable by popularity within a given YouTube channel or category. Thus, our sample may not be representative of NSI captioning practices on all of YouTube, but it should be representative of the most watched videos across categories. However, given this sampling, our results may be overly optimistic about the current state of NSI captioning on YouTube and the true presence of NSI captions may be even lower. Furthermore, many of the categories that were sampled from are user-assigned and may not always be reflective of the true content of the video, and YouTube's definition of "popular" when querying via their API is not transparently defined. Second, even in our sample of videos, there may be NSI that was undetected. This is because we only manually annotated videos that had already been estimated to contain NSI by our NSI estimator — thus we know the precision of this estimator, but we don't know the recall. Thus, despite our efforts to catch as much NSI as possible with our estimator, we may be unknowingly neglecting some videos in this analysis. Notably, our sample focused exclusively on English language captions from videos released primarily in the United States and the trends found hear may not generalize outside of this cultural context. Similarly, this study focused solely on captioning practices on YouTube and the findings presented here may not translate to other video-sharing platforms, or even to similar video content created by the same studio distributed on different platforms.

The annotation process itself may have introduced unintended bias, even with the process of independent labeling and rectifying disagreements between annotators. The annotators were not trained captioners and did not always consider the full context of the entire video while annotating the captions that were flagged by the filter as plausibly containing NSI. Additionally, given the exploratory nature of this work, the causality and underlying mechanisms of some trends, such as why manually captioned videos tend to have more views, could be partially explained by current literature [46, 47], other trends, such as why over 80% of large studio videos do not contain manual captions, could not currently be sufficiently explained by the current dataset or available literature. We therefore recommend the underlying mechanisms driving these trends be explored in future work.

Lastly, only closed captions were analyzed. Open captions, which render the caption onto the video, would require video content analysis using computer vision, which was out of the scope of this work. While presumably less common than closed captions, this accessible NSI information is left out of our analysis.

9 CONCLUSION & FUTURE WORK

In this work, we performed an initial study of NSI captioning practices on YouTube to understand the contemporary state and historical trends, and to identify technological gaps that could improve the accessibility of the vast collection of videos on YouTube. The key contribution of this research is the compilation, annotation, and exploratory analysis of the captions of close to 715k YouTube videos published from 2013-2022. Approximately 36k lines of captions were manually annotated with the specific sub-type(s) of NSI they contained, and the resulting dataset is publicly available at: https://doi.org/10.5281/zenodo.10681804.

We conducted an exploratory analysis of the dataset. Notable trends in this exploratory analysis included the relationship between higher view counts on videos with manual captions, notable differences in captioning practices between the large studio and popular video samples, as well as the effect of video duration and topic on NSI captioning.

Future work may explore the underlying mechanisms of the trends uncovered by this paper. These include the captioning practices and video content distribution pipelines used at larger studios that result in over 80% of videos on these channels not having manual caption tracks available on YouTube, even though manual caption tracks for the same content are available on other platforms. Additionally, the trend of over 30% of popular videos not having manual or automatic captioning available is a pressing concern. As imperfect as automatic captioning often is, providing absolutely no captioning is arguably worse. Given that, at the time of publication, channels have to actively opt out of automatic captions on their videos, it is important to more deeply understand the motivations behind content creators choosing to do this. Similarly, a deeper, authoritative understanding of how captioning, be it automatic or manual, impacts a video's recommendation performance and monetization eligibility would be very beneficial for content creators.

While best-practice guides and small studies have investigated the impact of inadequate NSI captioning on DHH viewers' experiences, more research is needed in this area to undercover how specific factors, such as video genre, on-screen redundancy of NSI, and poorly described NSI, impact the viewing experience. Additional work is also needed to investigate the details of captioning practices and trends within different categories of NSI, such as music or sound effects, on a more granular level to understand the nuances behind each category to better inform the design of new captioning tools.

[8] Such as this video which contains no access to captioning https://www.youtube.com/watch?v=s58yfxvAvqU. This is a short clip of the *ESPN* show *The Jump*, which aired on American TV.

REFERENCES

[1] 2021. Closed Captioning of Internet Video Programming. https://www.fcc.gov/consumers/guides/captioning-internet-video-programming

[2] 2021. Closed Captioning on Television. https://www.fcc.gov/consumers/guides/closed-captioning-television

[3] 2023. YouDescribe - Audio Description for Youtube Videos. https://www.youdescribe.org/.

[4] {3Play Media}. 2017. *2017 State of Captioning Report*. Technical Report. 3Play Media.

[5] {3Play Media}. 2023. *2023 State of Captioning Report*. Technical Report. 3Play Media. https://go.3playmedia.com/soc-2023

[6] Akhter Al Amin. 2020. Audio-Visual Caption Evaluation Metric for People who are Deaf and Hard of Hearing. (2020).

[7] 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*. 1–12.

[8] 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 *Universal Access in Human-Computer Interaction. Access to Media, Learning and Assistive Environments: 15th International Conference, UAHCI 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II*. Springer, 189–201.

[9] 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.

[10] 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.

[11] Akhter Al Amin, Saad Hassan, Sooyeon Lee, and Matt Huenerfauth. 2022. Watch It, Don't Imagine It: Creating a Better Caption-Occlusion Metric by Collecting More Ecologically Valid Judgments from DHH Viewers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[12] Akher Al Amin, Joseph Mendis, Raja Kushalnagar, Christian Vogler, Sooyeon Lee, and Matt Huenerfauth. 2022. Deaf and Hard of Hearing Viewers' Preference for Speaker Identifier Type in Live TV Programming. In *International Conference on Human-Computer Interaction*. Springer, 200–211.

[13] Tom Apone, Brad Botkin, Marcia Brooks, and Larry Goldberg. 2011. Caption Accuracy Metrics Project Research into Automated Error Ranking of Real-time Captions in Live Television News Programs. *The Carl and Ruth Shapiro Family National Center for Accessible Media, Boston* (2011).

[14] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred appearance of captions generated by automatic speech recognition for deaf and hard-of-hearing viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[15] Larwan Berke, Khaled Albusays, Matthew Seita, and Matt Huenerfauth. 2019. Preferred appearance of captions generated by automatic speech recognition for deaf and hard-of-hearing viewers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[16] Andy Brown, Rhia Jones, Mike Crabb, James Sandford, Matthew Brooks, Mike Armstrong, and Caroline Jay. 2015. Dynamic subtitles: the user experience. In *Proceedings of the ACM international conference on interactive experiences for TV and online video*. 103–112.

[17] Ben Caldwell, Michael Cooper, Loretta Guarino Reid, Gregg Vanderheiden, Wendy Chisholm, John Slatin, and Jason White. 2008. Web content accessibility guidelines (WCAG) 2.0. *WWW Consortium (W3C)* 290 (2008), 1–34.

[18] Sourish Chaudhuri. 2017. Adding sound effect information to YouTube captions. https://ai.googleblog.com/2017/03/adding-sound-effect-information-to.html

[19] U.S. Congres. 2010. Twenty-First Century Communications and Video Accessibility Act of 2010.

[20] U.S. Congress. 1990. Television Decoder Circuitry Act of 1990. https://www.congress.gov/bill/101st-congress/senate-bill/1974 Pub. L. No. 101-431.

[21] U.S. Congress. 1996. Telecommunications Act of 1996.

[22] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 831, 15 pages. https://doi.org/10.1145/3544548.3581511

[23] Gregory J Downey. 2008. *Closed captioning: Subtitling, stenography, and the digital convergence of text with television*. JHU Press.

[24] Konstantinos Drossos, Sharath Adavanne, and Tuomas Virtanen. 2017. Automated audio captioning with recurrent neural networks. In *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 374–378.

[25] Meryl K Evans. 2019. Here's how automatic captions earned their nickname. https://www.youtube.com/watch?v=N7MfajxyWDY

[26] David Fourney and Deborah Fels. 2008. "Thanks for pointing that out." Making sarcasm accessible for all. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. SAGE Publications Sage CA: Los Angeles, CA, 571–575.

[27] Andrew Gallagher, Terrance McCartney, Zhonghua Xi, and Sourish Chaudhuri. 2017. Captions based on speaker identification. (2017).

[28] Benjamin M Gorman, Michael Crabb, and Michael Armstrong. 2021. Adaptive Subtitles: Preferences and Trade-Offs in Real-Time Media Adaption. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.

[29] Michael Gower, Brent Shiver, Charu Pandhi, and Shari Trewin. 2018. Leveraging pauses to improve video captions. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 414–416.

[30] Ken Harrenstien. 2009. Automatic captions in YouTube. https://googleblog.blogspot.com/2009/11/automatic-captions-in-youtube.html

[31] Saad Hassan, Yao Ding, Agneya Abhimanyu Kerure, Christi Miller, John Burnett, Emily Biondo, and Brenden Gilbert. 2023. Exploring the Design Space of Automatically Generated Emotive Captions for Deaf or Hard of Hearing Users. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 125, 10 pages. https://doi.org/10.1145/3544549.3585880

[32] Shawn Henry. 2022. Captions/Subtitles. https://www.w3.org/WAI/media/av/captions/

[33] Yongtao Hu, Jan Kautz, Yizhou Yu, and Wenping Wang. 2015. Speaker-following video subtitles. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 11, 2 (2015), 1–17.

[34] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. AVscript: Accessible Video Editing with Audio-Visual Scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[35] Dhruv Jain, Angela Lin, Rose Guttman, Marcus Amalachandran, Aileen Zeng, Leah Findlater, and Jon E. Froehlich. 2019. Exploring Sound Awareness in the Home for People who are Deaf or Hard of Hearing. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (2019). https://api.semanticscholar.org/CorpusID:86865534

[36] Bo Jiang, Sijiang Liu, Liping He, Weimin Wu, Hongli Chen, and Yunfei Shen. 2017. Subtitle positioning for e-learning videos based on rough gaze estimation and saliency detection. In *SIGGRAPH Asia 2017 Posters*. 1–2.

[37] Sushant Kafle and Matt Huenerfauth. 2019. Predicting the understandability of imperfect english captions for people who are deaf or hard of hearing. *ACM Transactions on Accessible Computing (TACCESS)* 12, 2 (2019), 1–32.

[38] Saba Kawas, George Karalis, Tzu Wen, and Richard E Ladner. 2016. Improving real-time captioning experiences for deaf and hard of hearing students. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. 15–23.

[39] Daniel Killough and Amy Pavel. 2023. Exploring Community-Driven Descriptions for Making Livestreams Accessible. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13.

[40] JooYeong Kim, SooYeon Ahn, and Jin-Hyuk Hong. 2023. Visible Nuances: A Caption System to Visualize Paralinguistic Speech Cues for Deaf and Hard-of-Hearing Individuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 54, 15 pages. https://doi.org/10.1145/3544548.3581130

[41] Kuno Kurzhals, Fabian Göbel, Katrin Angerbauer, Michael Sedlmair, and Martin Raubal. 2020. A view on the viewer: Gaze-adaptive captions for videos. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[42] Raja Kushalnagar, Gary Behm, Kevin Wolfe, Peter Yeung, Becca Dingman, Shareef Ali, Abraham Glasser, and Claire Ryan. 2019. RTTD-ID: Tracked captions with multiple speakers for deaf students. *arXiv preprint arXiv:1909.08172* (2019).

[43] Raja S Kushalnagar, Gary W Behm, Joseph S Stanislow, and Vasu Gupta. 2014. Enhancing caption accessibility through simultaneous multimodal information: visual-tactile captions. In *Proceedings of the 16th international ACM SIGACCESS conference on Computers & accessibility*. 185–192.

[44] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. 2013. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 151–162.

[45] Daniel G Lee, Deborah I Fels, and John Patrick Udo. 2007. Emotive captioning. *Computers in Entertainment (CIE)* 5, 2 (2007), 11.

[46] Elisa Lewis. 2018. 7 ways captions and transcripts improve video SEO. https://www.3playmedia.com/blog/7-ways-video-transcripts-captions-improve-seo/

[47] 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. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–26.

[48] F. Wai ling Ho-Ching, Jennifer Mankoff, and James A. Landay. 2002. From Data to Display: the Design and Evaluation of a Peripheral Sound Display for the Deaf.

https://api.semanticscholar.org/CorpusID:10800692

[49] Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H Tang, et al. 2022. Visually-aware audio captioning with adaptive audio-visual attention. *arXiv preprint arXiv:2210.16428* (2022).

[50] Kim Lyons. 2020. YouTube is ending its community captions feature and deaf creators aren’t happy about it. https://www.theverge.com/2020/7/31/21349401/youtube-community-captions-deaf-creators-accessibility-google

[51] Tara Matthews, Janette Fong, F Wai-Ling Ho-Ching, and Jennifer Mankoff. 2006. Evaluating non-speech sound visualizations for the deaf. *Behaviour & Information Technology* 25, 4 (2006), 333–351.

[52] Lloyd May, So Yeon Park, and Jonathan Berger. 2023. Enhancing Non-Speech Information Communicated in Closed Captioning Through Critical Design. In *Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–14.

[53] Described Media Program and Captioned. 2017. Captioning key - sound effects and music. https://dcmp.org/learn/602-captioning-key---sound-effects-and-music

[54] Xinhao Mei, Xubo Liu, Mark D Plumbley, and Wenwu Wang. 2022. Automated audio captioning: an overview of recent progress and new challenges. *EURASIP journal on audio, speech, and music processing* 2022, 1 (2022), 1–18.

[55] Andrea Murphy and Hank Tucker. 2023. The Global 2000. https://www.forbes.com/lists/global2000

[56] Iftekhar Naim, Daniel Gildea, Walter Lasecki, and Jeffrey P Bigham. 2013. Text alignment for real-time crowd captioning. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 201–210.

[57] Somang Nam, Deborah I Fels, and Mark H Chignell. 2020. Modeling closed captioning subjective quality assessment by deaf and hard of hearing viewers. *IEEE Transactions on Computational Social Systems* 7, 3 (2020), 621–631.

[58] Andrew D Ouzts, Nicole E Snell, Prabudh Maini, and Andrew T Duchowski. 2013. Determining optimal caption placement using eye tracking. In *Proceedings of the 31st ACM international conference on Design of communication*. 189–190.

[59] Ellie Parfitt. 2016. Auto-generated captions are often wrong. https://www.hearinglikeme.com/nomorecraptions/

[60] Amy Pavel, Gabriel Reyes, and Jeffrey P Bigham. 2020. Rescribe: Authoring and automatically editing audio descriptions. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 747–759.

[61] Raisa Rashid, Quoc Vy, Richard Hunt, and Deborah I Fels. 2008. Dancing with words: Using animated text for captioning. *Intl. Journal of Human–Computer Interaction* 24, 5 (2008), 505–519.

[62] Pablo Romero-Fresco and Juan Martínez Pérez. 2015. Accuracy rate in live subtitling: The NER model. *Audiovisual translation in a global context: Mapping an ever-changing landscape* (2015), 28–50.

[63] James Sandford. 2015. The impact of subtitle display rate on enjoyment under normal television viewing conditions. (2015).

[64] Quoc V Vy and Deborah I Fels. 2009. Using avatars for improving speaker identification in captioning. In *Human-Computer Interaction–INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part II 12*. Springer, 916–919.

[65] Mike Wald. 2011. Crowdsourcing correction of speech recognition captioning errors. In *Proceedings of the International Cross-Disciplinary Conference on Web Accessibility*. 1–2.

[66] James M Waller and Raja S Kushalnagar. 2016. Evaluation of automatic caption segmentation. In *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility*. 331–332.

[67] Fangzhou Wang, Hidehisa Nagano, Kunio Kashino, and Takeo Igarashi. 2016. Visualizing video sounds with sound word animation to enrich user experience. *IEEE Transactions on Multimedia* 19, 2 (2016), 418–429.

[68] Tom Wheeler, Rosenworcel, Clyburn, Pai, and O’Rielly. 2014. Federal Communications Commission FCC 14-12 before the federal ... https://docs.fcc.gov/public/attachments/fcc-14-12a1.pdf

[69] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[70] YouTube. 2023. YouTube for Press. https://blog.youtube/press/

[71] Sean Zdenek. 2015. Reading sounds. In *Reading Sounds*. University of Chicago Press.

A ADDITIONAL FIGURES

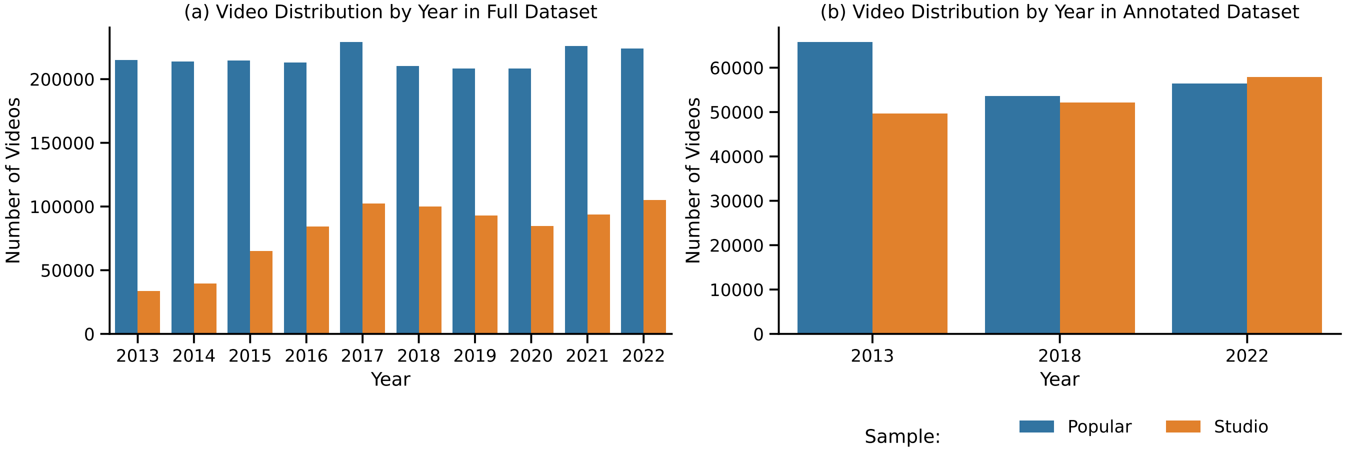

Figure 8: Video distribution by year within the YouTube NSI Captioning Dataset for both the full (a) and annotated datasets (b)

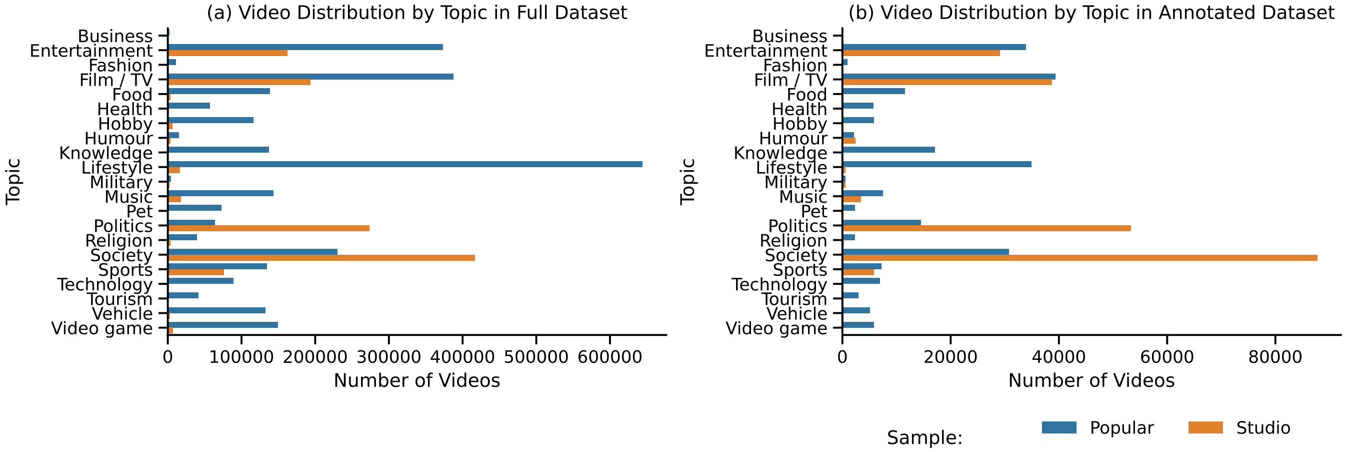

Figure 9: Video distribution by YouTube-assigned topic within the YouTube NSI Captioning Dataset for both the full (a) and annotated datasets (b)

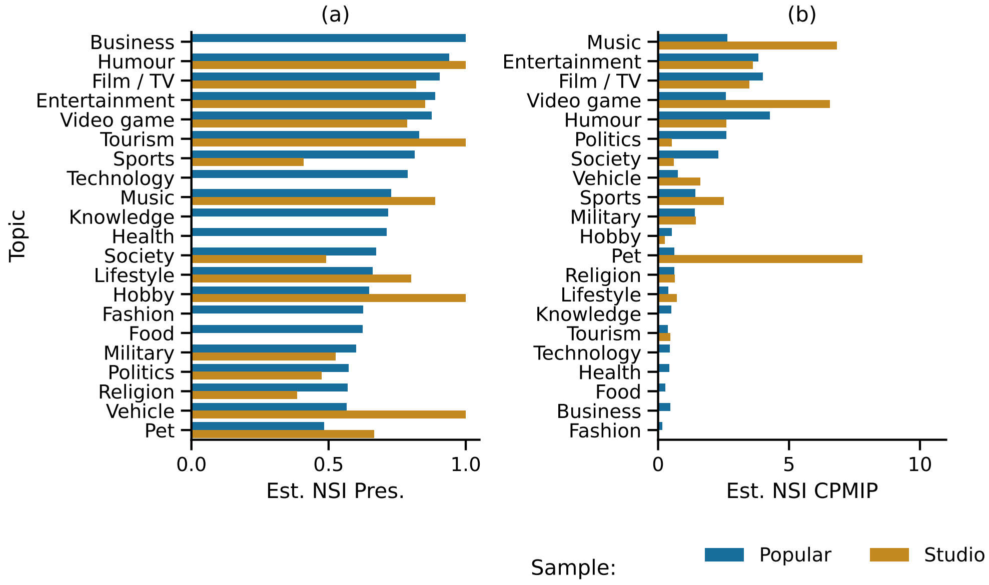

Figure 10: The (a) mean estimated NSI presence and (b) median estimated NSI CPMIP by YouTube assigned topic for both the *popular* and *studio* 2022 samples in the full dataset.

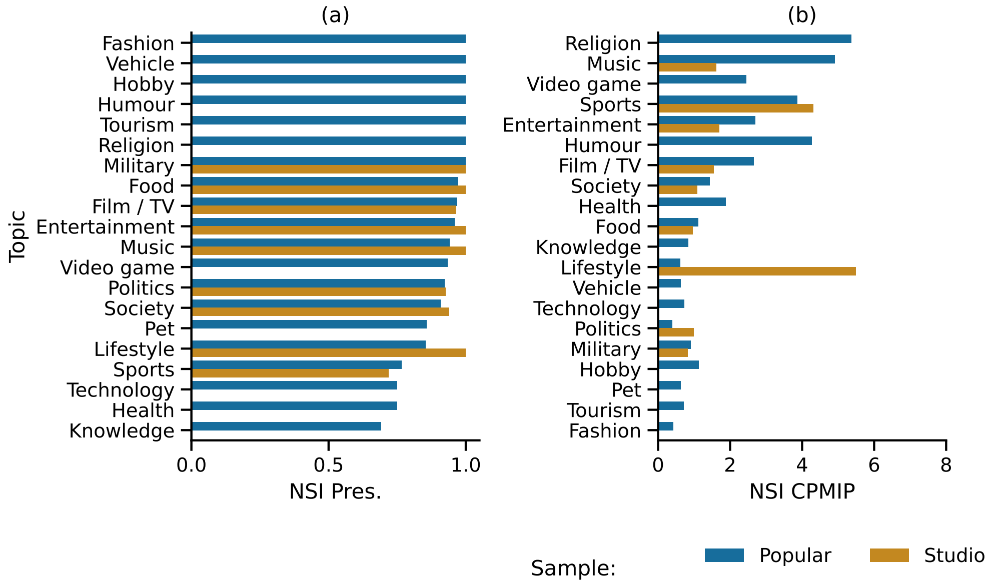

Figure 11: The (a) mean NSI presence and (b) median NSI CPMIP by YouTube assigned topic for both the *popular* and *studio* 2022 samples in the annotated subset.

B YOUTUBE TOPIC SIMPLIFICATION

YouTube-assigned topics from the YouTube Data API were simplified and mapped to higher-level topics (e.g., *baseball* → *sports*) using the following table:

YouTube Topic	Simplified Topic
https://en.wikipedia.org/wiki/Action-adventure_game	Video game
https://en.wikipedia.org/wiki/Action_game	Video game
https://en.wikipedia.org/wiki/American_football	Sports
https://en.wikipedia.org/wiki/Association_football	Sports
https://en.wikipedia.org/wiki/Baseball	Sports
https://en.wikipedia.org/wiki/Basketball	Sports
https://en.wikipedia.org/wiki/Boxing	Sports
https://en.wikipedia.org/wiki/Business	Business
https://en.wikipedia.org/wiki/Casual_game	Video game
https://en.wikipedia.org/wiki/Christian_music	Music
https://en.wikipedia.org/wiki/Classical_music	Music
https://en.wikipedia.org/wiki/Country_music	Music
https://en.wikipedia.org/wiki/Cricket	Sports
https://en.wikipedia.org/wiki/Electronic_music	Music
https://en.wikipedia.org/wiki/Entertainment	Entertainment
https://en.wikipedia.org/wiki/Fashion	Fashion
https://en.wikipedia.org/wiki/Film	Film / TV
https://en.wikipedia.org/wiki/Food	Food
https://en.wikipedia.org/wiki/Golf	Sports
https://en.wikipedia.org/wiki/Health	Health
https://en.wikipedia.org/wiki/Hip_hop_music	Music
https://en.wikipedia.org/wiki/Hobby	Hobby
https://en.wikipedia.org/wiki/Humour	Humour
https://en.wikipedia.org/wiki/Ice_hockey	Sports
https://en.wikipedia.org/wiki/Independent_music	Music
https://en.wikipedia.org/wiki/Jazz	Music
https://en.wikipedia.org/wiki/Knowledge	Knowledge
https://en.wikipedia.org/wiki/Lifestyle_(sociology)	Lifestyle
https://en.wikipedia.org/wiki/Military	Military
https://en.wikipedia.org/wiki/Mixed_martial_arts	Sports
https://en.wikipedia.org/wiki/Motorsport	Sports
https://en.wikipedia.org/wiki/Music	Music
https://en.wikipedia.org/wiki/Music_of_Asia	Music
https://en.wikipedia.org/wiki/Music_of_Latin_America	Music
https://en.wikipedia.org/wiki/Music_video_game	Video game
https://en.wikipedia.org/wiki/Performing_arts	Performing arts
https://en.wikipedia.org/wiki/Pet	Pet
https://en.wikipedia.org/wiki/Physical_attractiveness	Physical attractiveness
https://en.wikipedia.org/wiki/Physical_fitness	Physical fitness
https://en.wikipedia.org/wiki/Politics	Politics
https://en.wikipedia.org/wiki/Pop_music	Music
https://en.wikipedia.org/wiki/Professional_wrestling	Sports
https://en.wikipedia.org/wiki/Puzzle_video_game	Video game
https://en.wikipedia.org/wiki/Racing_video_game	Video game
https://en.wikipedia.org/wiki/Reggae	Music
https://en.wikipedia.org/wiki/Religion	Religion
https://en.wikipedia.org/wiki/Rhythm_and_blues	Music
https://en.wikipedia.org/wiki/Rock_music	Music
https://en.wikipedia.org/wiki/Role-playing_video_game	Video game
https://en.wikipedia.org/wiki/Simulation_video_game	Video game
https://en.wikipedia.org/wiki/Society	Society
https://en.wikipedia.org/wiki/Soul_music	Music

YouTube Topic	**Simplified Topic**
https://en.wikipedia.org/wiki/Sport	Sports
https://en.wikipedia.org/wiki/Sports_game	Video game
https://en.wikipedia.org/wiki/Strategy_video_game	Video game
https://en.wikipedia.org/wiki/Technology	Technology
https://en.wikipedia.org/wiki/Television_program	Film / TV
https://en.wikipedia.org/wiki/Tennis	Sports
https://en.wikipedia.org/wiki/Tourism	Tourism
https://en.wikipedia.org/wiki/Vehicle	Vehicle
https://en.wikipedia.org/wiki/Video_game_culture	Video game
https://en.wikipedia.org/wiki/Volleyball	Sports