Bystander Privacy in Video Sharing Era: Automated Consent Compliance through Platform Censorship

Si Liao*†‡
School of Information Science and Technology
ShanghaiTech University
Shanghai, Shanghai, China
liaosi@shanghaitech.edu.cn

Hanwei He
School of Information Science and Technology
ShanghaiTech University
Shanghai, Shanghai, China
hehw2024@shanghaitech.edu.cn

Huangxun Chen
IoT Thrust, Information Hub
Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, Guangdong, China
huangxunchen@hkust-gz.edu.cn

Zhice Yang
School of Information Science and Technology
ShanghaiTech University
Shanghai, Shanghai, China
yangzhc@shanghaitech.edu.cn

Abstract

Bystander privacy has become a critical concern amidst the widespread activities of video sharing, engaging billions of users daily. Concerns arise when individuals inadvertently appear in public videos without consent. Existing methods for determining bystander permissions require significant adaptation and modifications by videographers and video sharing platforms, potentially limiting their adoption. This study explores leveraging platform censorship capabilities to enforce bystander privacy. We introduce *SelfFlag*, a type of violative media signal designed to trigger automatic content flagging. Bystanders exhibiting such signals, captured in public videos, can be automatically identified and removed by platforms, thereby indirectly enforcing privacy preferences, primarily through the efforts of bystanders themselves. We conduct thorough measurements on current censorship practices, propose music-based triggering content, and develop an auxiliary tool for videographers to produce high-quality content with privacy compliance.

CCS Concepts

• **Security and privacy → Privacy protections**; *Social aspects of security and privacy*; • **Hardware → Sound-based input / output**.

Keywords

Bystander Privacy, Copyright, Censorship, Music

ACM Reference Format:
Si Liao, Hanwei He, Huangxun Chen, and Zhice Yang. 2025. Bystander Privacy in Video Sharing Era: Automated Consent Compliance through Platform Censorship. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3706598.3713391

*This work was done while Si Liao was a visiting student at HKUST (GZ).
†Also with Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Science.
‡Also with University of Chinese Academy of Sciences.

CHI '25, Yokohama, Japan

ACM ISBN 979-8-4007-1394-1/25/04
https://doi.org/10.1145/3706598.3713391

1 Introduction

Video-sharing platforms have become integral to daily life for billions of users, driven by rapid advancements in Internet technology. Platforms like YouTube, TikTok, Instagram, and Facebook attract countless individuals who spend significant time watching and sharing a wide range of video content [19, 40]. As of 2024, there are over 200 million content creators worldwide [19], and YouTube alone sees over 3.7 million videos uploaded daily [41]. Users typically watch an average of 17 hours of online video content per week, and nine out of ten businesses use video for marketing [40].

However, alongside the popularity of video sharing, concerns over bystander privacy have emerged. Video shoots often occur in public settings where individuals, *i.e.*, bystanders, may unintentionally appear in the background of footage without their consent or intention to be recorded. Sharing such videos publicly can potentially expose their personal information. Previous studies [29, 52] on bystanders' privacy concerns have shown that video capturing raises significant privacy issues, ranking just below biometric data collection in terms of bystanders' discomfort.

To address this privacy concern, one direct method is face-to-face interaction to seek confirmation from bystanders nearby. However, this approach can be inefficient, and bystanders may not even realize they are being recorded. Some video-sharing platforms offer a privacy complaint option, allowing users to request the removal of content related to them [10]. However, such measures are often not timely. Bystanders may not realize their video has been uploaded until it has already caused significant discomfort.

For automated methods, despite many concepts and prototypes proposed in the literature, practical tools have not yet emerged. Reasons for this vary. For example, the sharing platform can enforce individuals' privacy preferences by first recognizing identity and then protecting their personal information in uploaded content [15, 17, 43]. However, implementing identity recognition on uploaded content will not only add operating costs but also, due to the lack

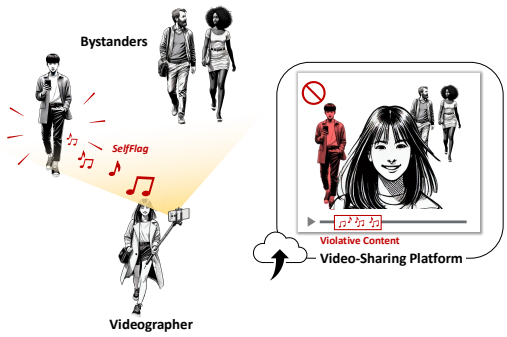


Figure 1: Utilizing Censorship Mechanisms to Enforce Bystander Privacy. Bystanders play *SelfFlag* signals, *i.e.*, music tracks with identifiable copyright features. Videographers record bystander images along with their *SelfFlag* signals. After videos are uploaded, the content censorship mechanisms of video-sharing platforms identify segments containing violative music and suggest edits. In this way, bystanders indirectly communicate and enforce their privacy preferences.

of a legal framework, potentially lead to battles over functionalities and privacy. For instance, Facebook halted its face recognition tagging feature years ago due to privacy law accusations [33].

On the other hand, it is the videographers who produce and upload the videos. However, enforcing bystanders' privacy preferences by videographers introduces its own complexities, and perhaps even more so. Our survey and existing studies indicate that some videographers do value bystander privacy and are willing to take steps to respect their privacy needs [2]. However, technically, there are no widely accepted methods for communicating privacy preferences with passersby, and processing the video itself involves additional costs. More importantly, videographers often face a trade-off between protecting privacy and maintaining the functional and aesthetic quality of the video [2, 14]. Given these challenges, the most we can critique is limited to "they should have asked me first" or "they should have screened me out".

Therefore, from an incentive perspective, we believe it is most reasonable and direct for bystanders to take the lead in initiating privacy protection measures for themselves. However, interdependent privacy protection cannot be a standalone effort [16, 46], so bystander-led initiatives should aim to minimize the need for software or hardware modifications by video-sharing platforms or videographers, and avoid negative impacts on videographers' creative content. This leads to a key question: how can the involved parties, with the bystander leading the effort, collaborate to effectively respect bystander privacy?

Interestingly, we observe that, while video platforms lack dedicated recognition capabilities for bystanders, they do possess powerful capabilities for screening content related to copyright, harassment, violence, *etc.*, to comply with laws and regulations. They deploy automated and mandatory censorship steps for uploaded content. Our idea is to associate an individual's identity with content that violates censorship guidelines. When this individual exhibits such content, videos inadvertently capturing it will automatically be flagged and blocked by platform censorship mechanisms prior to sharing. Such effort by the bystander indirectly communicates and enforces its privacy permissions by making use of the censorship mechanism.

In this paper, we will validate and realize this idea. Firstly, we introduce *SelfFlag*. As shown in Fig. 1, *SelfFlag* represents a type of media signals carried by bystanders with privacy protection needs. *SelfFlag* is designed to trigger automatic content flagging by video-sharing platforms' censorship. Our goal is for *SelfFlag* to serve as an efficient, elegant, and unobtrusive information indicator. To achieve this, we systematically study and measure the content censorship behaviors of major video-sharing platforms. Based on our findings, we select music as the primary form for *SelfFlag* and establish criteria for choosing suitable tracks. Additionally, we design ultrasonic speakers to play *SelfFlag* music in a human-inaudible frequency range, enabling the scheme to operate effectively even in quiet environments.

Furthermore, in many cases, videographers prefer that the captured *SelfFlag* content does not lead to heavy editing or muting of their primary content. To assist videographers in complying with bystander permission requirements while preserving the integrity and aesthetics of their original content, we develop an automated *SelfFlag* compliance tool. This tool utilizes a cloud service to link the bystander's identity information with the *SelfFlag* music, enabling the targeted removal of only the content associated with the bystander. To minimize the impact on video content during the removal, the tool employs image inpainting techniques [22, 24, 53] to seamlessly fill in the areas affected by the removal.

The contributions of this work are summarized as follows:

- We revisit the practice of bystander privacy from a new perspective and propose a method to leverage platform censorship to enforce bystanders' privacy preferences.
- We measure and characterize the current landscape of content censorship on major video-sharing platforms.
- We propose a music-based censorship triggering method, with detailed evaluation in real-world environments.
- We present a video editing tool that automates bystander privacy compliance for videographers when sharing videos.

In the following sections, we will first review the literature in Section 2 and introduce the content censorship background in Section 3. Then, to identify appropriate bystander-carried violative content, in Section 4, we will conduct a comprehensive test of censorship mechanisms on major video-sharing platforms. Subsequently, in Section 5, based on our findings, we will enforce bystanders' privacy preferences using copyrighted music as censorship triggers. To address videographers' needs, we will also propose a tool to help achieve automated privacy compliance while preserving the main video content.

2 Related Work

Previous efforts to enforce bystander privacy have primarily focused on the photo sharing scenario. The emerging video format, which generally carries more information and deserves careful privacy protection, has not been extensively studied so far. The following overview of existing efforts mainly derives from the photo sharing context. Bystander privacy involves three participants: the

Related Work	Videographer	Sharing Platform	Bystander	Protection
I-Pic [1]	sw+radio		sw+radio	on-capture
iRyP [47]	sw+radio		sw+radio	on-capture
DNC [34]	sw+radio		sw+radio	on-capture
PrivacyCamera [23]	sw+radio+loc		sw+radio+loc	on-capture
Offlinetags [31]	sw		tag	on-share
Privacy.Tag [4]	sw		tag	on-share
Cardea [42]	sw	sw	reg	on-capture
COIN [56]	sw+loc	sw	sw+loc	on-capture
PrivacyEye [45]	sw+hw			on-capture
SnapMe [17]	loc	sw	loc	on-share
HideMe [43]		sw	reg	on-share
Videre [57]		sw	reg*	on-share
Hasan et.al[15]		sw		on-share
Our approach	sw		hw*+reg*+speaker	on-share

*: optional
sw: software
hw: specific hardware
reg: registration
loc: localization capability

Table 1: Summary of Related Work.

videographer, the video-sharing platform, and the bystander. These methods can be broadly categorized based on the involved efforts. We summarize representative works in Table 1.

2.1 Recorder in Charge

Recorder-in-charge refers to the scenario where protection measures are mainly implemented by videographers. There are two main challenges to address in this context. Firstly, bystanders need effective ways to communicate their privacy preferences to photographers. The literature commonly utilizes short-range radio technologies such as Bluetooth [1, 34, 47], Wi-Fi Direct, and visible tags [4, 31] for this purpose.

The subsequent question is how photographers can accurately identify bystanders in their captured scenes who require protection. One approach involves bystanders sharing their identity information, such as facial features [1, 34, 47] or precise location [23, 56], which photographers can then use to locate them in their captured footage. Some methods aim to further keep bystander information entirely private from photographers, *i.e.*, enforcing privacy policies during the on-capture phase before the media is stored and shared. This is not straightforward since photographers must identify bystanders first to apply privacy measures. Therefore, these methods often involve cryptography [1] or third-party platforms [42, 56] to negotiate essentials without revealing bystanders' identities.

The above processes necessitate close collaboration between bystanders and photographers' devices and software. PrivacyEye [45] explored a different approach focused solely on the videographer's side. They propose that videographers wear specific eye-tracking devices to measure their eye movements, which could indicate whether captured bystander information should be protected.

2.2 Server Inspection

Another set of solutions is to deploy new cloud services that are able to automatically detect bystanders in uploaded media content and apply corresponding privacy protection measures. These services are not necessary but preferably integrated with the sharing platform.

For example, SnapMe [17] observes that photos' metadata often includes location information. They propose utilizing this data to identify registered users near the location and apply privacy measures based on these users' preferences if they appear as bystanders in the uploaded images. Similarly, Li et al. [43] proposed HideMe, where registered users can specify a privacy policy that decide to allows photo viewers to obtain their face information or not.

Server inspection often does not impose strong assumptions on the capabilities of videographers. Additionally, some designs do not even require prior knowledge, such as registered facial information, from bystanders. For example, Hasan et al. [15] uses machine learning to statistically analyze the features of bystanders in photos, based on which it determines what content to retain and display in general uploaded photos. Videre [57] integrates registered facial information with statistical features of bystanders. When it identifies a registered user in a photo, it adheres to that user's privacy policy; when no registered user is identified, it classifies individuals as bystanders based on their behavior and awareness.

3 Content Censorship Background

In this section, we review the content censorship mechanisms on video-sharing platforms from three perspectives:

- What type of content does the censorship mechanism target?
- What is the primary method for targeted content detection?
- What actions are taken after detecting violative content?

3.1 Censored Content

The primary objective of censorship on video-sharing platforms is to comply with laws and regulations, guide user behavior, and promote an inclusive and friendly environment. According to official documentation from YouTube [54], Instagram [20], and previous surveys [25, 32, 38, 44, 48], the content targeted by censorship can be classified into three main categories as shown is Table 2.

- **Sensitive and Harmful Content**. Sensitive content includes nudity, eroticism, child pornography, and ongoing harassment and bullying of others [51]. Harmful content includes hate speech targeting specific groups, material that promotes and supports terrorist activities, violence, self-harm, *etc.*
- **Spam and Misleading Content**. Spam refers to irrelevant or repetitive advertisements for promotion [49]. Misinformation involves the intentional spread of false or misleading information that can cause public panic or defraud individuals [27, 39].

Content Category	Specifics
Sensitive & Harmful Content	Nudity, Eroticism, Child Pornography Violence, Bullying, Harassment, Hate Speech, Terrorism, Self-Harm
Spam & Misleading Content	Misinformation, Spam, Fraud
Intellectual Property	Trademark, Copyright

Table 2: Censored Content on Video-Sharing Platforms.

- **Intellectual Property (IP)**. For video-sharing platforms, IP primarily includes trademarks and copyrights, which protect the interests of creators.

3.2 Content Detection Approach

To enforce censorship, video-sharing platforms primarily rely on two detection methods: manual review and detection algorithms. The actual choice generally depends on the clarity of the censored content and the state-of-the-art in technical development. In terms of sensitive, harmful, and spam content, platforms may utilize advanced AI models for image recognition or natural language processing to block videos with clear violations. However, some content like misleading ones often lacks a unified and clear specification, forcing platforms to rely on manual review to address it. For copyright issues, most video-sharing platforms are equipped with automated tools to detect copyright infringements. Additionally, these platforms often allow creators to add their copyrighted content to the platform's copyright pool, enabling easier management and protection of their works.

3.3 Post-detection Measures

Upon detecting targeted content, video-sharing platforms may adopt the following countermeasures:

- **Video Masking and Flagging.** Videos containing controversial content may be subjected to masking and flagging operations. For instance, Facebook places a warning screen [26] over potentially inappropriate content to inform viewers, allowing them to decide whether to view it.
- **Video Deletion or Prohibition.** Videos with severe content violations will be directly banned by the platform or automatically removed within a short time after publication, preventing access by viewers.
- **Account Restriction or Suspension.** If the number of violations exceeds a certain threshold, the uploader may be restricted from using certain functions of the account, such as uploading new videos or posting comments. For severe violations like terrorism or child pornography, the platform may immediately and permanently suspend the account.

4 Repurposing Content Censorship

4.1 Effectiveness of Different Censored Content

To use censored content to signal the involved parties about the bystander's privacy preferences, we must first determine which modality or type of censored content is most effective. Table 2 summarizes potential options, but their actual effectiveness remains unclear. We have not found any publicly available research covering this topic, so this subsection will focus on testing the censorship

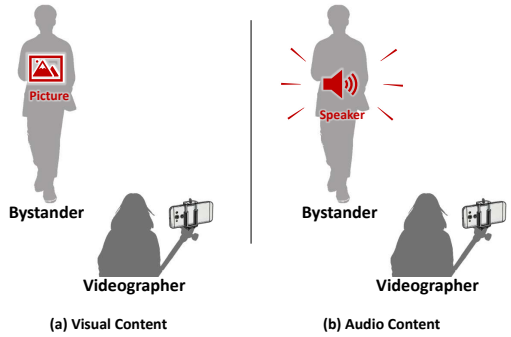


Figure 2: Example of Showing and Playing Censored Content.

mechanisms of video-sharing platforms. According to our application scenario, we are primarily concerned with the following two points: (1) whether the detection of specific types of content is automated, and (2) the general criteria that content must meet to trigger detection mechanisms.

Therefore, we create videos featuring various types of censored content and test them against five major video-sharing platforms to observe how these platforms' censorship mechanisms respond. Specifically, according to Fig. 2, we synthesize censored content into healthy videos in two modalities as follows:

• Visual Content: To emulate the scenario in Fig. 2(a), in an ordinary vlog video, we overlaid static images containing violence (sensitive content), trademarks (IP), fake news pages (misleading content), and muted copyrighted TV series clips (IP) onto the bystander's chest using video editing tools. These contents were collected through search engines, and in our view, each one clearly violates the platform's content regulations. They were overlaid in a picture-in-picture format, and we also adjusted their size to test the impact of the proportion of violative content.

• Audio Content: To emulate the scenario in Fig. 2(b), where the bystander can introduce audio into video recordings using a portable speaker (*e.g.*, a mobile phone), we overlaid new audio tracks onto the existing audio track of the same vlog video. These additional tracks included hate speech (harmful content) and copyrighted music (IP). For hate speech, we first selected textual examples from a social media corpus [9], and then utilized text-to-speech technology to convert them into audio format. As for the copyrighted music, we selected them from our own Apple Music subscription and ensured that they are not available in the free music libraries of video-sharing platforms.

We produce five test videos for each sub-types mentioned. Each video is five minutes long, with the censored content appearing within the first 1.5 minutes. The experimental results are shown in Table 3. We have the following findings:

• *For the majority of content types, content censorship is neither automated nor can it be triggered by simple complaint reports.* ○ symbols in Table 3 indicate that the videos can be successfully uploaded and shared. However, even after manually reporting these contents several times using a different account registered several months ago, they remain accessible on the platform. It can be understood that the criteria for detecting violations of certain censorship guidelines are highly subjective. For these types of content, it may be

Censored Modality	Specific Topic	Video-sharing Platform				
		YouTube	Instagram	X (Twitter)	TikTok	Facebook
Visual Content	Graphic Violence (Sensitive Content)	●	●	●	●	●
	Trademarks (IP)	○	○	○	○	○
	Copyrighted Video (IP)	●	○	○	○	○
	Fake News Page (Misleading Content)	●	○	○	○	●
Audio Content	Hate Speech (Harmful Content)	●	○	○	○	○
	Copyrighted Music (IP)	●	○	○	●	○

Table 3: Testing Censored Content against Video-Sharing Platforms. The symbol ○ indicates that the content is explicitly prohibited in the official guidelines and has a designated user reporting channel, but is not able to be auto-detected. The symbol ● indicates that the platform can detect the censored content when the visual content occupies the entire video frame or the audio content fills the entire audio track. The symbol ● indicates that the censored content is automatically detected even when the content does not occupy the entire video frame or audio track.

difficult to establish a quantitative algorithm for automated screening. What we did not anticipate is that the platforms do not respond to our complaint reports. This could be due to two reasons: firstly, reports from a single source may not be sufficient for the platform's decision-making; secondly, since our videos have not gained significant popularity, the platforms may prioritize resources to handle more influential content.

• *Different platforms vary in their automated censorship detection capabilities.* Generally, automated detection is available for violent content and content related to copyright. The first row of Table 3 suggests that platforms can automatically flag graphic violence to some extent. Additionally, YouTube and TikTok are capable of automatically detecting copyrighted videos and audio. Moreover, the more explicit the violative content in the video, the easier it is to detect. For instance, YouTube and TikTok flagged the test videos containing violent content even when the violative content was overlaid on only a portion of the video, while other platforms often require the content to occupy nearly the entire frame.

4.2 Choosing Copyrighted Music as a Carry-on Censorship Trigger

If bystanders aim to trigger platform censorship by signaling specific content, this content must, in practical terms, be able to effectively trigger the automated censorship detection mechanisms of video-sharing platforms. At the same time, it should also be suitable for everyday carrying, without overwhelming the bystander or causing discomfort to others.

Based on the empirical measurements described above, it is noted that within the visual content category, only violent content can be automatically blocked by the platform. It would be quite unreasonable to have people wear clothing containing such content in public and display it conspicuously. Considering that some platforms may internally use neural networks for detection, there might be methods to create images that do not appear to contain violence but are recognized as violent content by the neural network. However, we have not tested such methods because they typically are not effective when physically expressed in the real world.

For audio content, copyrighted music demonstrates outstanding effectiveness on YouTube and TikTok. Additionally, we find that including just a few seconds rather than the entire track is enough to trigger copyright flagging. Furthermore, playing music through portable devices like smartphones is quite a normal behavior and does not require installing any software. Therefore, we consider copyrighted music to be a good candidate to help bystanders express their privacy concerns and enforce their privacy preferences.

4.3 In-depth Testing of Using Music as a Censorship Trigger

To use music as triggering content, this subsection will conduct more detailed tests and analysis of video-sharing platforms' ability to detect copyrighted music, thereby guiding the selection criteria. We consider various factors, including:

- **Music Genre** categorizes different types of music based on a shared tradition or set of conventions.
- **Music Duration** is the duration of music recorded by videographers.
- **Music Section** describes the structural elements of music recorded by videographers.
- **Signal to Noise Ratio (SNR)** quantifies the audio interference recorded by videographers; when noise power is constant, SNR reflects the volume of the playback.
- **Overlapping Occurrence** indicates whether there are multiple pieces of copyrighted music playing simultaneously.

The testing protocol involves first creating a series of music clips from full music tracks. These tracks are explicitly included in the copyright lists of video-sharing platforms. By default, the clips are from the middle chapter of the music track. They last for 65 seconds and have a uniform average volume with no added noise. Then, we use a typical 90-second vlog video with no sound as a carrier. We overlay one of the produced audible music clips onto the audio track of this carrier video using video editing tools, aligning the start of the music clip with the start of the video's audio track. This emulates the scenario where a videographer records music played by bystanders. Next, we upload this edited video to video-sharing platforms to test whether it triggers copyright flagging. The detection ratio is the ratio of triggered instances to non-triggered instances. The followings describe these tests and results:

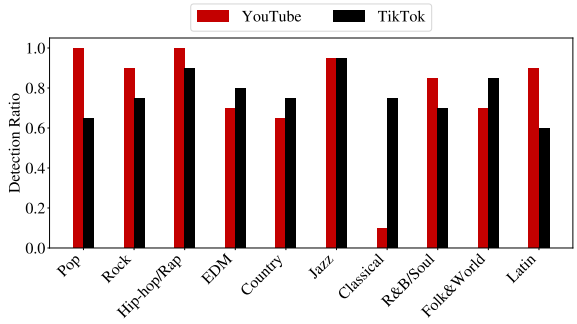


Figure 3: Impact of Music Genre.

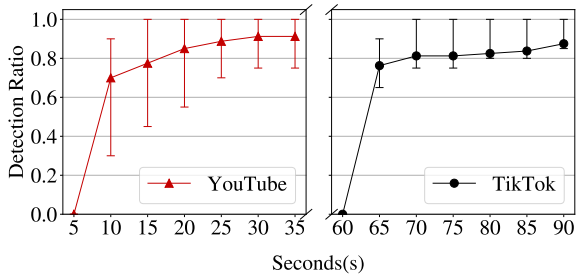


Figure 4: Impact of Music Duration.

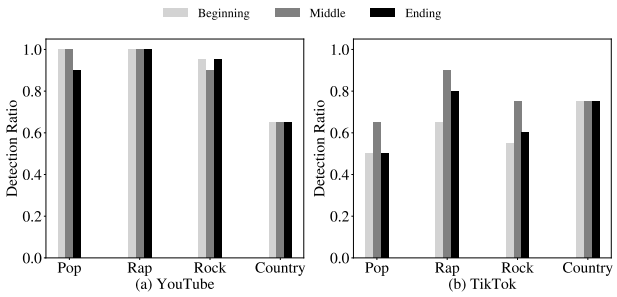


Figure 5: Impact of Music Section.

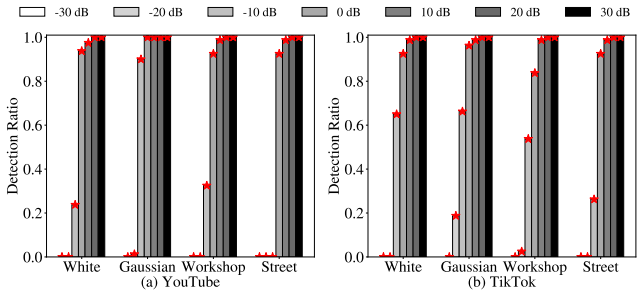


Figure 6: Impact of Noise.

4.3.1 Music Genre. We refer to a mainstream music software [12] for music genres classification and select 10 music genres for testing: *Classical*, *Countryside*, *EDM*, *Folk/World Music*, *Jazz*, *Latin*, *Popular*, *R&B/Soul*, *Rap*, and *Rock*. We choose 20 tracks from each category on our Apple Music library. To ensure linguistic diversity, 50% of them is in English, while the remaining includes German, French, Spanish, Portuguese, Korean, Chinese, and instrumental ones without lyrics.

Fig. 3 illustrates the detection ratio for 10 music genres on both platforms. Jazz and Rap demonstrate quite high detection ratios on both platforms, both exceeding 90%. Pop music shows a notable discrepancy in triggering copyright flagging between two platforms, with 100% detection ratio on YouTube but only 65% on TikTok. Classical music is less effective on YouTube, achieving only 10% detection ratio.

Guideline 1): Jazz and Rap are good choices of music genre for triggering platform censorship flagging.

4.3.2 Music Duration. A complete music track typically lasts for 3 to 4 minutes. As video-sharing platforms do not require the entire track to assert copyright violation, the necessary duration determines the temporal sensitivity of the triggering method. To determine the minimal length of music clips required to ensure consistent copyright detection, we select 20 different music tracks from each of the four most popular genres—*Pop*, *Rock*, *Rap*, and *Country*. The music clips made from these tracks range from 5 seconds to 90 seconds in length.

The results are shown in Fig. 4. The overall trend is a gradual increase in the detection ratio as the duration of the music clip increases. On YouTube, none of the 5-second clips triggers copyright flagging, while 91.25% of the 30-second clips are recognized. Extending the duration beyond 30 seconds does not further improve the detection ratio. On TikTok, music clips of 60 seconds or less can not trigger copyright detection, while 70-second clips achieve a detection ratio of 81.25%.

Guideline 2): Different platforms have varying levels of sensitivity in term of music duration. When the duration of the music clip exceeds a certain threshold (YouTube - 10 seconds, TikTok - 65 seconds), the detection ratio tends to stablize.

4.3.3 Music Section. Most music pieces typically start with a simple melody, while the middle section usually features a more complex melody and more lyrics, and the ending section often has a lower pitch. To investigate which music section is more effective in general, we extract different sections from the music tracks. To avoid subjective partition, our extraction is not based on musical content, *e.g.*, intro, verse, chorus, outro, *etc.*, but rather on quantifiable positions: beginning: clip from the first second; middle: clip from (total duration - 65)/2 seconds; ending: clip from (total duration - 65) seconds.

Fig. 5 illustrates the impact of different music sections on detection ratios. Overall, the effect of different sections on the detection ratio is minimal. On YouTube, the detection ratios across different sections are quite similar. On TikTok, however, the middle section demonstrates better performance compared to the other sections. Another observation from the figure is that the overall detection ratio on TikTok is lower, which is due to the fact that the 65-second music clips we used do not fully reach its optimal detection duration. According to the results in Fig. 4, extending the duration of the music clip to 90 seconds can increase the detection ratio to 81%.

Guideline 3): The middle part of a music track, *e.g.*, the verse section, is easier to detect by the platforms.

(a) TikTok

Volume(A-B)	-10dB					0dB					10dB				
Overlap Radio	**0**	**0.25**	**0.5**	**0.75**	**1.0**	**0**	**0.25**	**0.5**	**0.75**	**1.0**	**0**	**0.25**	**0.5**	**0.75**	**1.0**
Pop&Country	AB	AB	AB	AB	B	AB	AB	AB	A	A	AB	AB	A	A	A
Pop&Pop	AB	AB	AB	AB	B	AB	AB	AB	A	A	AB	AB	A	A	A
Pop&Rap	AB	AB	B	B	B	AB	AB	AB	A	A	AB	A	A	A	A
Pop&Rock	AB	AB	AB	AB	B	AB	AB	AB	A	A	AB	AB	AB	A	A

(b) YouTube

Volume(A-B)	-10dB					0dB					10dB				
Overlap Radio	**0**	**0.25**	**0.5**	**0.75**	**1.0**	**0**	**0.25**	**0.5**	**0.75**	**1.0**	**0**	**0.25**	**0.5**	**0.75**	**1.0**
Pop&Country	AB	AB	AB	AB	B	AB	AB	AB	AB	B	AB	AB	AB	AB	A
Pop&Pop	AB	AB	AB	AB	B	AB	AB	AB	AB	A	AB	AB	AB	AB	A
Pop&Rap	AB	AB	AB	AB	B	AB	AB	AB	AB	A	AB	AB	AB	AB	A
Pop&Rock	AB	AB	AB	AB	B	AB	AB	AB	AB	A	AB	AB	AB	AB	A

Table 4: Impact of Overlapping. The table shows the detection results of overlapping music tracks A and B, considering different combinations of genres, percentage overlap ratios, and volumes. The symbol A denotes that only A's copyright is detected; the symbol B denotes that only B's copyright is detected; the symbol AB denotes that both are detected.

4.3.4 Music Volume and Environmental Noise. Previous tests all used ideal tracks without any noise, but in actual video recordings, the music played by the bystanders recorded by the videographer is inevitably affected by other environmental sounds. To assess the platforms' performance under such conditions, we conducted this test. We follow the previous default test setup and select 10 music tracks that are relatively easy to detect according to previous tests. We then add four representative types of environmental noise to the audio tracks: Gaussian noise, white noise, workshop noise, and street noise. Gaussian noise is generated by scaling a normal distribution with a mean of 0 and a variance of 1. Both white noise and workshop noise are sourced from Google's AudioSet public dataset [13]. However, this audio set lacks a diverse range of street sounds, such as street conversations and vendor calls. Therefore, we sourced alternative sounds from other online resources [28] to represent street noise. Since noise audio clips are typically short, we looped the segments to match the length of the videos.

To test the impact of noise, we scale the values of the noise signals to different levels while keeping the music signals unchanged. Since the noise is not physically played, we cannot measure its absolute sound intensity. Instead, we use the ratio of the power of the music to the power of the noise, called the Signal-to-Noise Ratio (SNR), to characterize the impact of different noise levels. The SNR measures the relative intensity of the music compared to the noise. We test an SNR range from -30 dB to 30 dB, where a lower SNR indicates that the music is weaker compared to the noise, and 0 dB indicates that the intensities of the two signals are approximately equal.

Fig. 6 shows the results at different SNR levels. We use red stars to reveal the invisible bars. Overall, it can be observed that the clearer the music, the higher the detection ratio. When the music power exceeds the noise power by 10 dB, the detection ratio approaches 100% and becomes relatively stable. When the music intensity is 20 dB lower than the noise, the copyright detection mechanism essentially stops working. Our subjective feeling is that -20 dB is a very strong noise level where the music is barely audible. From this, we also personally feel that their detection algorithms are quite robust to noise.

Furthermore, the copyright detection algorithms of different platforms exhibit varying sensitivities to different types of noise. Except in scenarios involving Gaussian noise, TikTok is more likely to detect copyright when the music volume is below the noise level (less than 0 dB). However, as the music volume surpasses the noise volume, YouTube's detection ratio quickly increases to match, and in some cases, exceeds TikTok's detection performance.

Guideline 4): To ensure that the platform can effectively detect copyrighted music clips, the bystander should try to keep the music volume equal to or higher than the environmental sounds during playback.

4.3.5 Overlapping Occurrence. In a real vlogging scenario, it is possible that multiple bystanders might be playing different copyrighted music tracks simultaneously. To understand the performance in this situation, this test will measure the detection ratio under different combinations of overlaid genres, ratios, and volumes. Specifically, we selected two pieces of music from different genres, denoted as A and B, to serve as a test pair. They were combined and then played. All the selected pairs can be correctly detected by the two platforms. To further analyze different overlap ratios, all music clips are cut to a length of 180 seconds. For each test pair, we overlay B onto A, while adjusting the volume of B to −10 dB, 0 dB, and +10 dB relative to the volume of A, and also delaying the start time of B to vary the overlap ratio of A and B's tracks from 0% (no overlap) to 100% (full overlap).

The results are shown in Table 4. When the two music clips are non-overlapped, both YouTube and TikTok can detect them (AB). When the two music clips are completely overlapped, usually only the clip with the larger volume can be detected. If the volumes of the two clips are similar, *i.e.*, 0 dB, at least one of the clips can still be detected, but the results may be random. When the two music clips are partially overlapped, in most cases both platforms can correctly identify the two clips. In comparison, TikTok requires a

greater non-overlap ratio, which is reasonable given that TikTok is less sensitive in time domain as suggested by our previous *Music Duration* test.

Guideline 5): When the videographer's recorded audio track contains multiple overlapping music tracks, the detection mechanism may miss some of the information.

5 Design and Implementation

5.1 Minimum Viable Implementation

Based on the tests and guidelines described in §4.3, individuals concerned about being inadvertently filmed in situations where they need privacy protection can choose to use their phone's loudspeaker to play the middle sections of some copyrighted Jazz or Rap music tracks on loop mode at a volume louder than the ambient sounds. Then, any nearby videographer devices capturing the music-playing individual will also record the music being played. Once the video is uploaded to a mainstream sharing platform, the platform's content censorship mechanism will flag the video portions containing copyrighted music and prohibit the posting of the whole video. The uploader will be given the option by the platform to either mute the video completely or trim the relevant audio and video portions to comply with copyright regulations (§3.3). As a result, the bystander's privacy preference is indirectly communicated to the videographer, and the platform's auto-censorship detection helps the bystander limit the public sharing of the video content containing it.

Above is the basic and core design of *SelfFlag*. It does not require any dedicated programs or hardware; users only need to be able to play music. However, in practice, there are two aspects that need improvement. First, using a loudspeaker to play music in public to convey a privacy preference may not be suitable for all scenarios, as surrounding people might not appreciate the music and could even find it annoying. Second, the platform's content flagging and coarse-grained editing suggestions might not always align with the videographer's intentions and could potentially affect the content they wish to share. For example, when copyrighted music is playing, the videographer might also be providing meaningful commentary or narration. In the next two subsections, we address them by implementing *SelfFlag* Player and *SelfFlag* Cloud.

5.2 "Inaudible" Music Player

The *SelfFlag* Player is used by the bystanders. Its hardware component is responsible for emitting music signals. The phone's built-in loudspeaker is a convenient option, as it allows bystanders to use their existing devices. Additionally, we implement an external speaker capable of playing music that is inaudible to humans but audible to a normal microphone. As shown in Fig. 8(a), it mainly consists of an ultrasonic speaker, an OPA541 power amplifier, and a 192 kHz sound card[1].

This design is largely based on a phenomenon widely explored by existing studies [7, 30, 36, 55]: ultrasonic signals, typically above 20 kHz, are beyond the human hearing range but can still be detected by microphones. This is because, in practice, due to imperfections in some microphone units, such as non-linearity in amplifiers, ultrasonic signals cannot be completely filtered out and instead create low-frequency aliases that fall within the microphone's receivable range.

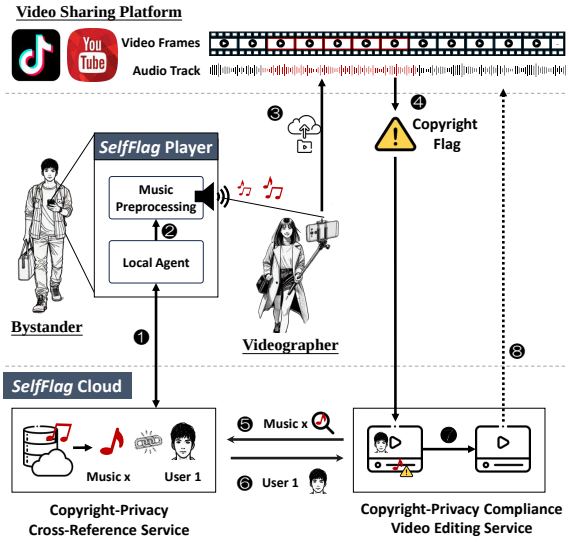


Figure 7: *SelfFlag* Overview.

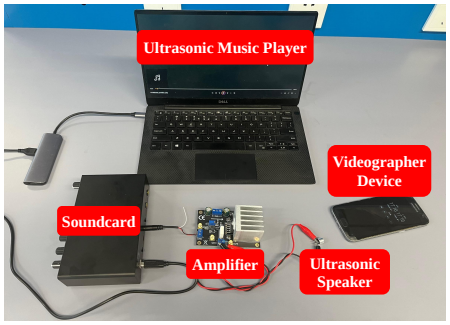
As shown in Fig. 7, the software component of the *SelfFlag* Player is a mobile application that serves two main functions: First, ❶ an agent module that works in conjunction with the *SelfFlag* Cloud Service to associates the bystander's identity with specific copyrighted music through the enrollment service. Second, ❷ a preprocessing module that converts the original music signals into ultrasonic signals by modulating the amplitude of an ultrasonic carrier according to the original music signal, as shown in Fig. 8(b).

5.3 Cloud Service for Automated Privacy Compliance

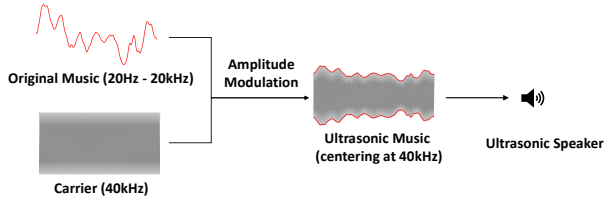
The *SelfFlag* Cloud provides a cloud-based tool for videographers to meet the privacy preferences of bystanders. As we analyze in §1, this is a concern for bystanders, but for videographers, who respect bystanders' privacy, they might be willing or, to some extent, be compelled to address it—especially after video sharing platforms indirectly communicate bystanders' preferences by rejecting uploads. Nevertheless, we aim for such a tool to seamlessly integrate into the videographer's workflow without bringing too much overhead. We expect a third-party service funded by bystanders to provide such services for videographers. Technically, it helps videographers remove bystander information from the footage while keeping the rest of the content intact. Also, the copyrighted music should be removed to allow successful sharing.

As shown in Fig. 7, the *SelfFlag* Cloud offers two main services. The first is the Copyright-Privacy Cross-Reference Service, which connects copyrighted music with the privacy preferences of specific accounts. It maintains a music-identity database that associates privacy information of one account, such as its facial features, with

[1]The speaker costs less than $0.5, and some of the latest smartphones [55] are already equipped with sound card that support sampling rates above 128 kHz for high-fidelity audio playback.



(a) Ultrasonic Speaker Hardware



(b) Software Processing to Produce Ultrasonic Music

Figure 8: Playing "Inaudible" Music. The original music signal is frequency-shifted to the ultrasonic band by amplitude-modulating a 40 kHz ultrasonic carrier wave. The corresponding sound emitted by the ultrasonic speaker is inaudible to human ears but can be received by ordinary microphones.

one copyrighted music. During enrollment, users receive a list of copyrighted music options and can select one based on their preferences and their permission to play the music. The second service is the Copyright-Privacy Compliance Video Editing Service, which renders a seamless privacy compliance editing workflow for videographers.

Specifically, ❸ when a videographer captures a bystander using the *SelfFlag* Player, the recorded video will include the copyrighted music associated with the bystander. ❹ The content will fail to meet censorship regulations of sharing platforms, and the videographer can upload the video to the *SelfFlag* Cloud for automated privacy content cleaning. ❺ The Editing Service uses the name of the music indicated by the video sharing platform[2] to ❻ retrieve the associated bystander's account and its privacy information, ❼ and employs image and audio removal techniques to produce video content that complies with both copyright and privacy requirements while preserving fluency and quality. ❽ After processing, the service returns a clean video to the videographer for review and reuploading. The above processing is hosted on a server with Intel Xeon Platinum 8378A processor and NVIDIA RTX A600 GPUs.

5.3.1 Privacy Compliance Video Inpainting. The video processing workflow of step ❼ is detailed in Fig. 9: First, ① the cloud determines the bystander account and video frame range that needs to be processed based on the flagged music information returned by the sharing platform. Then, ② it applies image segmentation [21]

[2]Alternatively, we also use SoundFingerprinting [8] for music fingerprinting.

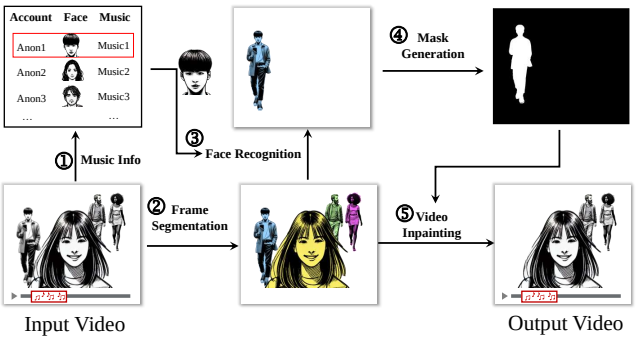


Figure 9: Workflow of Privacy Compliance Video Inpainting.

to each video frame to divide it into multiple parts based on their semantic properties. After that, ③ it uses recognition algorithm [11] to locates the segmented part containing the bystander according to the bystander's information, *e.g.*, facial features. ④ These segmented parts provide a highlighting mask for the video inpainting algorithm [58] to ⑤ remove the bystander from all video frames. According to existing study [22, 24, 53], this algorithm can intelligently fill in the blank areas left, ensuring that the cleaned video looks natural and coherent.

5.3.2 Copyright Compliance Audio Cleaning. In step ❼, after ensuring the bystander's privacy, we employ audio source separation techniques [18, 35] to remove the copyrighted music from the video's audio track to allow successful uploading.

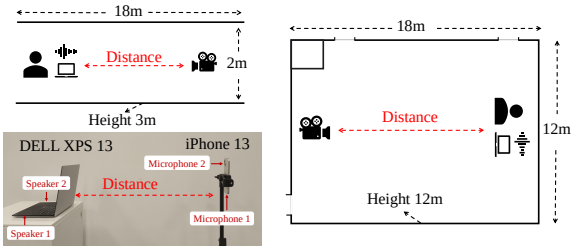
When the videographer records multiple bystanders using *SelfFlag* Player to output music signals, in most cases the compliance process is similar to that of a single bystander. This is because the platform can separately detect and return the start positions of multiple tracks as long as they do not completely overlap[3]. However, in cases where the music signals from bystanders happen to be completely overlapped, the platform might miss detecting some of the music tracks. In such situations, the video after one round of cleaning may still be rejected by the platform, as the overlapped music signals might be revealed after the cleaning process. Therefore, *SelfFlag* will perform multiple rounds of clean processing. This ensures that all overlapping signals and bystanders are addressed before the final output, eliminating the need for videographers to re-upload and face repeated rejections.

6 Experiments and Results

Our experiments aim to understand the practical implications of using *SelfFlag* in the following aspects:

- The Effectiveness of Copyright Detection: We assess how the copyright detection systems of video-sharing platforms perform when music is played and recorded in real-world scenarios.
- The Effectiveness of Bystander Privacy Protection: We evaluate *SelfFlag*'s overall effectiveness of bystander privacy protection, specifically its ability to accurately remove a bystander from the video footage.

[3]It seems to be the case for YouTube. While we did not measure the exact limit, it is capable of detecting at least 25 copyrighted pieces within a single video. In contrast, TikTok's limit appears to be around 4 or 5 pieces.



(a) Corridor (b) Badminton Hall

Figure 10: Test Sites Layout and Settings.

Figure 11: Videographer Hardware.

- Users' Willingness of Using *SelfFlag*: We aim to understand whether bystanders and videographers are willing to adopt *SelfFlag* in practice for privacy protection, as well as the potential concerns associated with using or not using *SelfFlag*.

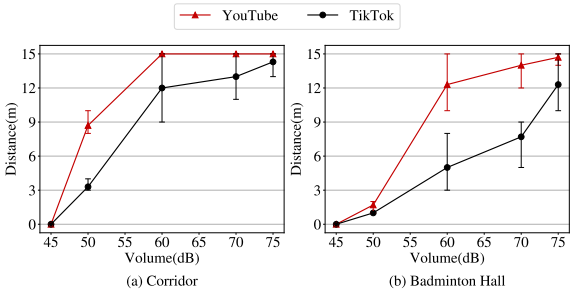
6.1 Effectiveness of Copyright Detection

6.1.1 Method. We played music in real-world settings while simultaneously recording to emulate the situation where a bystander is captured on video. We uploaded the recorded results to YouTube and TikTok, two mainstream sharing platforms, to understand under what conditions copyright flags might be triggered.

We used a DELL XPS 13 laptop as the music playback device for the bystander. The laptop offers two audio output options: its built-in speakers, located along the bottom edges near the keyboard, which produce sounds audible to the human ear, and the ultrasonic speaker (Fig. 8(a)), which emits sounds inaudible to humans. In the experiments presented below, unless explicitly stated otherwise, we used the non-ultrasonic (audible) speakers for music playback.

We selected three devices—Samsung S7 (smartphone), Sony ZV1 (camera), and Dji Action 4 (sport camera)—as the default videographer hardware. Following the guidelines detailed in §4.3, we chose the middle section of a piece of Rap music as the audible copyrighted music for testing. We set the music duration to meet the requirements for both YouTube and TikTok copyright detection algorithms.

The setup for the recording and playback devices is shown in Fig. 10(a). Below, we consider the impact of different test sites (Fig. 10), recording distances, videographer hardware (Fig. 11), ultrasonic speakers (§5.2), and music factors (§4.3).



(a) Corridor (b) Badminton Hall

Figure 12: Copyright Detection Performance at Different Sites. Values are the maximum distance at which the music pieces can effectively trigger the platform's copyright flagging across various music volumes and test sites.

6.1.2 Test Sites. We conducted tests in two indoor environments: a narrow corridor and a relatively spacious badminton hall. The environmental noise levels in the testing sites were measured to be approximately 38 dB in the corridor and 33 dB in the badminton hall. We varied the volume of the emitted music from 45 dB to 75 dB and measured the maximum distance at which each platform can effectively trigger a copyright flag. Since facial features may not be clearly recognizable beyond a distance of 15 meters in video recordings under typical conditions [3, 50], we set the maximum test distance at 15 meters. Video recordings were conducted at one-meter intervals, with 10 shots taken at each distance and uploaded for copyright detection. We then calculated the detection ratio. If the radio exceeds 80%, we considered that distance a valid position for effective copyright detection.

The results, as depicted in Fig. 12, indicate that audio collected in the corridor scenario reach greater distances than those recorded in the badminton hall at the same music volume level. This difference is attributed to the walls alongside the corridor, which help reflect sound further, whereas audio energy is not that focused in the large open space of the badminton hall. In the corridor site, when the audio volume is set at 60 dB, all devices can trigger copyright detection on the platforms at a distance of more than 12 meters. However, this distance decreases to around 5 meters for TikTok in the badminton hall. The results indicate that under YouTube's detection capabilities, a bystander's privacy preference can be effectively conveyed even from a distance of more than 10 meters, covering the vast majority of use cases. Although TikTok has not yet achieved such capabilities, we believe that as technology evolves, it will improve in the future.

6.1.3 Videographer Hardware. We tested all eight devices shown in Fig. 11, rather than the default three, to assess the impact of videographer hardware. They are three common types of videographer devices: mobile phones, cameras, and sports cameras. For each type, we chose products from 2-4 major vendors. We still used the middle section of an audible Rap music for testing and set the music duration of 35 seconds for YouTube and 90 seconds for TikTok. We fixed the music volume to 60 dB and the test site to the corridor. Additionally, we also evaluated the music fingerprinting implementation based on the open-source SoundFingerprinting tool [8] for comparison. Similarly, we measured the maximum distance at which copyright flagging can be triggered.

Videographer Device	Device Brand	Bystander's Audible Speaker			Bystander's Inaudible Speaker		
		YouTube	TikTok	Sound-Fingerprinting	YouTube	TikTok	Sound-Fingerprinting
Mobile Phone	iPhone 13	15 m	12 m	10 m	-	-	0.1 m
	Samsung S7	15 m	12 m	10 m	10 cm	0.1 m	0.2 m
Camera	Sony ZV1	15 m	9 m	8 m	0.6 m	1 m	1 m
	Canon G7X3	15 m	9 m	8 m	-	0.6 m	0.7 m
Sports Camera	DJi Action4	15 m	15 m	14 m	0.8 m	1 m	1 m
	DJi Packet3	15 m	13 m	10 m	-	0.1 m	0.1 m
	Insta360 X4	15 m	8 m	10 m	0.2 m	0.3 m	0.5 m
	GoPro 12	15 m	12 m	12 m	-	-	-

Table 5: Copyright Detection Performance with Different Player and Videographer Hardware. Values in the table indicate the maximum distance at which the evaluated music pieces trigger copyright flagging across various videographer devices, copyright detection systems, and bystander's music playing devices.

The results in Table 5 show that the detection capabilities of different videographer devices vary. The two DJI devices and the GoPro sports camera achieve the farthest detection distances. This is likely due to their relatively better sensitivity and recording quality, which preserve the necessary audio details and facilitate copyright detection. By comparing the performance of different detection platforms and algorithms, we find that YouTube's copyright detection is more robust than the others. SoundFingerprinting exhibits a relatively shorter maximum detection distance. This may be attributed to our use of the default threshold for fingerprint matching, which is not optimized for our scenarios.

6.1.4 Inaudible Speaker. We then used the inaudible speaker in Fig. 11 to act as the bystander's device. As shown in Fig. 11(a), we pointed the microphone of the videographer device towards the ultrasonic speaker to ensure the recording quality. The other settings were the same as in the previous tests in §6.1.3.

Table 5 shows that the most noticeable trend is that the detection distance using an inaudible speaker is significantly lower compared to using an audible speaker. Through detailed investigation, we found that the volume of the recorded music is significantly lower than in cases involving non-ultrasonic speakers. This discrepancy may have caused certain videos, such as those from the iPhone 13, to fail in triggering copyright flagging. Additionally, some recordings contain device-specific noise from the microphone components. For example, in the Canon G7X3, this noise is more noticeable, which likely prevented it from successfully triggering YouTube's copyright flagging. Furthermore, the GoPro 12's microphone cannot record ultrasonic audio. We suspect this is due to extra shielding around the microphone module, designed for water resistance, which reduces its sensitivity to ultrasonic sounds.

In addition to the volume factor, we replayed some of the recorded audio to our ears and found that, subjectively, the music was clear with minimal noise interference. However, the detection results on YouTube and TikTok were poor. We suspect this may be due to the captured inaudible music exhibiting different frequency characteristics compared to regular music. This difference could cause it to fall outside the optimization zones that these platforms apply to regular music. Interestingly, SoundFingerprinting outperforms the two platforms in identifying the recorded ultrasonic music. We speculate that this is due to the relatively small size of music dataset we created for it. In contrast, YouTube and TikTok target large-scale music samples, requiring a balance between precision and recall, which likely involves additional filtering of abnormal interference factors present in the recorded ultrasonic music.

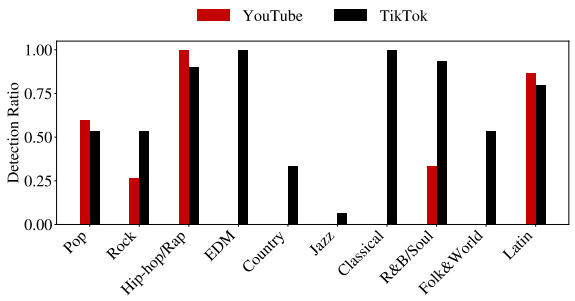


Figure 13: Copyright Detection Ratio v.s. Music Genre.

Based on the results above, to further improve the performance and usability of the *SelfFlag* Player, we have two possible approaches in mind. First, since the inaudible speaker does not disturb people nearby, its music volume could be further increased to improve the detection ratio. Existing research has also proposed methods for increasing the power efficiency and coverage of such speakers [7, 36, 37], which provide valuable references. Second, when converting a music track to ultrasonic music, the conversion can be tailored more precisely to the detection mechanisms. For instance, features that the detection algorithms are particularly sensitive to can be enhanced to improve detection.

6.1.5 Music Factors. §4 explores the impact of music factors using synthesized music tracks. In this experiment, we will revisit these factors in actual recordings. The settings for this experiment is similar to those in §6.1.2. The test site was located in the corridor, and tests were conducted with the non-ultrasonic speakers, Samsung S7, Sony ZV1, and DJI Action 4. Since the detection performance depends on the sensitivity of microphone, according to Table 5, the three recording devices were positioned at their maximum detection distances: 12 m, 9 m, and 15 m, respectively. The middle section was used in the music genre test. Each music piece, lasting 90 seconds, was played and recorded 10 times.

Fig. 13 shows that the detection ratios for the Hip-hop/Rap, Classical, and Latin genres are generally consistent with previous results

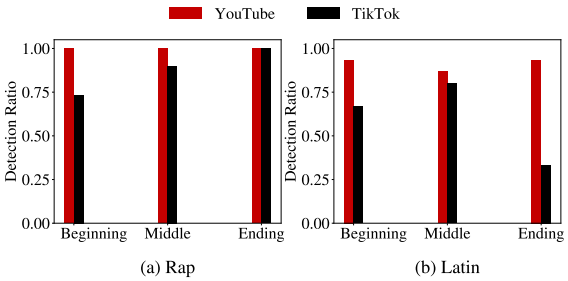


Figure 14: Copyright Detection Ratio v.s. Music Section.

in §4, though some genres differ slightly. Specifically, EDM, Country, Jazz, and Folk&World are not detected by YouTube. Additionally, the detection ratio for the Jazz genre on TikTok is also lower. These differences may be due to the fact that different music genres exhibit varying frequency distributions, and different frequencies experience different levels of propagation attenuation. According to the results in Fig. 13, Rap and Latin are selected to investigate the impact of different music sections in Fig. 14. The findings are consistent with previous conclusions, indicating that the middle section generally provides better and more stable detection performance.

6.2 Effectiveness of Bystander Privacy Protection

6.2.1 Method. In this experiment, we aim to assess whether the proposed solution can effectively protect privacy in real-world scenarios. In a complete video recording and sharing process, the overall effectiveness of *SelfFlag* depends both on the music detection capabilities evaluated in the previous experiment and on the effectiveness of the video editing solutions provided by *SelfFlag* Cloud. We invited two participants to act as bystanders in the following two scenarios, and we acted as videographers to record their activities:

- **Restaurant Scenario**: The videographer sat in front of a dining table and used a tripod to hold the Samsung S7 smartphone for video recording. Two participants, Bystander-1 and Bystander-2, appeared in the video. Bystander-1 sat at a nearby table for a while and played audible copyrighted music on his mobile phone, with the speaker turned on to protect his privacy. Bystander-2 walked across the footage and did not play any music.
- **Outdoor Corridor Scenario**: The videographer was holding a Samsung phone and walking through an outdoor corridor, taking selfies with the front camera, much like many vloggers do. Two participants, Bystander-1 and Bystander-2, were captured in the footage. During the recording, Bystander-1 continuously played audible copyrighted music on his phone, with the speaker facing forward (though not always directly towards the recorder). Unlike the restaurant scenario, Bystander-1's distance and relative direction from the videographer gradually changed, causing the recorded music volume to vary. Bystander-2, who was closer to the videographer, did not play any music. This scene is shown in Fig. 15(a), which was captured by another camera.

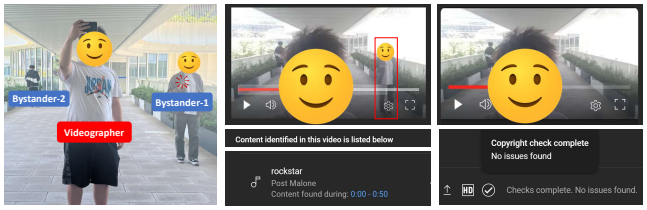


Figure 15: Experiment Settings and Results for Bystander Privacy Protection. (a) Video shooting scenario in an outdoor corridor. (b) Video-sharing platform flags copyright of an uploaded video. (c) After edited with the *SelfFlag* Cloud, the reuploaded video triggers no copyright flag, and Bystander-1, who is playing the music, is removed from the video frames.

Scenarios	Overall Effectiveness		
	Correct Editing	No Editing	Incorrect Editing
Restaurant	100%	0%	0%
Outdoor Corridor	80%	20%	0%

Table 6: Overall Effectiveness of Bystander Privacy Protection. In these videos, the bystanders use the *SelfFlag* player to output their music, and the percentages represent the proportion of cases in which bystanders' privacy is (not) successfully protected.

For each scenario, we recorded 10 videos and uploaded them to YouTube for copyright detection. If the platform identified copyrighted music, as shown in Fig. 15(b), we used the detected music name and specific time range to initiate the *SelfFlag* Cloud for privacy cleaning. After completing the processing steps, we reuploaded the edited video to YouTube for a second round of copyright detection. If *SelfFlag* functions correctly, the results should appear as shown in Fig. 15(c), where the bystander signaling copyrighted music is correctly removed from the video frames, and the copyrighted music is also removed from the audio tracks.

6.2.2 Results. According to the output of the *SelfFlag* Cloud, the results fall into three categories:

- **Correct Editing**: This represents cases where the bystander who activated the *SelfFlag* Player is accurately identified and removed.
- **No Editing**: This represents cases where the bystander who activated the *SelfFlag* Player is not removed.
- **Incorrect Editing**: This counts situations where a bystander who does not activate the *SelfFlag* Player is mistakenly removed.

In the outdoor corridor case, due to the unstable volume of the recorded music, the platform was unable to successfully identify copyrighted music in 20% of the cases, resulting in the associated bystander not being correctly removed. However, it is noteworthy that there were no instances of incorrect editing, which suggests that the face recognition and video inpainting models adopted in *SelfFlag* Cloud are well-developed. This indicates that the system can accurately identify and edit the correct subjects as long as the *SelfFlag* signal is properly detected.

6.3 User Survey

In our previous experiments, we focused on quantifiable performance metrics. In this experiment, we use a questionnaire to explore whether bystanders are willing to play music to protect their privacy and whether videographers are willing to process videos with privacy concerns. This study has been approved by the ethics review committee of our institute.

6.3.1 Participants. We randomly recruited 70 participants, aged between 20 and 60, with a nearly equal gender ratio. The participants were recruited online and included students, working professionals, and freelancers. Considering that younger users make up the majority of current video platform users [5, 6], 62 participants (89%) were between the ages of 20 and 40.

Before filling out the questionnaire, each participant received a detailed informed consent form outlining the background and purpose of the study. They were explicitly informed of their right to withdraw from the study at any time. Additionally, their personal information, such as names, was anonymized. We strictly adhere to data privacy protocols, ensuring that all responses remain confidential and are used solely for academic research. On average, each participant spent approximately 5-10 minutes completing the questionnaire and received a $5 gift card.

6.3.2 Survey Design. We used an electronic questionnaire to understand participants' preferences. The participants completed the questions under our supervision. While we did not disclose the specific details of *SelfFlag* design, we clearly conveyed the concept of privacy protection through the use of music prior to answering the questions.

The questionnaire is shown in the Appendix. It has two main sections: questions (1-6) are for bystanders, and questions (7-10) are for videographers. In the first part, we first confirm whether participants care about their privacy when being filmed by others. Next, we gather information on the methods they commonly use to protect their privacy. We then introduce the concept of using music as a privacy protection measure and ask if they would be willing to use this method. Subsequently, we ask them to compare the music-playing method with the methods they currently use. In the second part, we first filter out participants who have previously shared videos. We then inquire about their willingness to protect bystanders' privacy and their willingness to use the tools provided by *SelfFlag* to facilitate the privacy-cleaning task.

6.3.3 Results. 62 participants (89%) believe that when others post videos containing their facial information, it should be anonymized through methods like blurring, mosaics, or removal (Q1). However, more than half of the participants have never taken any measures to protect their privacy. One-third have proactively asked the videographer to blur their face, while some chose to wear masks or simply walked away to protect their privacy (Q2).

When participants who were concerned about their privacy were informed that playing music could notify videographers to protect privacy, 75% expressed a willingness to use this method (Q3): They thought this method could protect their privacy without requiring direct interaction with the videographer. Furthermore, it can also safeguard them in situations, where they were unaware of being recorded. (Q4) The remaining 25% reject to play music, because they concerned about: Playing music in public could potentially disturb others. Even with ultrasonic devices that emit inaudible music, some participants found the method too cumbersome (Q5). Regarding the ranking of various privacy protection methods (Q6), the results, as shown in Table 7, indicate that participants preferred notifying the videographer upon discovering their appearance in an online video. Compared to wearing physical tags on clothing, participants showed a stronger preference for continuously playing ultrasonic music.

Rank	Option
1st	File a complaint on the sharing platform
2nd	Inform the videographer upon discovering being filmed
3rd	Play inaudible music with an ultrasonic speaker
4th	Wear physical labels (*e.g.*, "Blur Me", "No Photos") on your clothing
5th	Play audible music

Table 7: Bystanders' Choices for Privacy Protection

Rank	Option
1st	Self-discipline, blur all bystanders by default
2nd	Platform notifies of bystander privacy issues when uploading
3rd	Bystander explicitly informs of their request in person

Table 8: The Most Motivating Factors for Videographers to Protect Bystanders' Privacy

In the feedback from Q2 and Q5, some participants mentioned that if they realized they appeared in someone else's recordings, they would choose either wear a mask or leave immediately. However, when they were unaware of being recorded or unable to leave, there were no better alternatives available. Therefore, a key advantage of using *SelfFlag* to protect privacy is its ability to address situations where the bystander is unaware that their privacy has been compromised. We will discuss this in more detail in §7.

Only 19 participants (27%) had previously recorded and shared videos on video sharing platforms (Q7). Among these, 8 participants reported that they had proactively blurred the faces of all bystanders in their videos, while 4 said they blurred the faces of bystanders who had notified them about privacy concerns (Q8). Table 8 presents the ranking of video-handling methods for protecting bystanders' privacy (Q9). Nearly half of the participants ranked proactive privacy protection, such as blurring all bystanders' faces, as the most likely approach. Others indicated that they would prefer to address privacy issues after receiving alerts from sharing platforms. Nonetheless, all participants agreed that they would use additional tools to anonymously handle their videos if privacy issues were identified (Q10).

7 Application Scenarios

As suggested by the survey results in §6.3.3, although directly notifying the videographer or filing a complaint to the platform are viable options, their feedback also reflects the necessity of using music in certain scenarios.

For individuals with physical disabilities, such as blindness or visual impairment, it can be difficult for them to become aware in time that someone nearby is recording a video and notify them to enforce their privacy preferences. As a result, they may be unintentionally captured in video footage. In such cases, *SelfFlag* could

provide an effective way to protect their privacy. Specifically, they could play copyrighted music when outdoors to signal that they want their identity to be blurred or masked in the recorded video.

Another possibility arises when someone is occupied with ongoing tasks and, as a result, may be unaware that they are being recorded, or it may simply be inconvenient to notify the videographer. In a gym scenario, for example, some videographers may share their daily exercise routines but inadvertently capture others in the background. However, people in the gym are often focused on their workouts and may not notice others or have the opportunity to pause and notify them. Similarly, in restaurants, customers may share their food and dining experiences on social media but unintentionally capture waiters in the video. The waiter may feel that notifying the customer about their privacy concerns is inappropriate or awkward. In such circumstances, *SelfFlag*allows individuals to express their privacy intentions without allowing their activities to take a break, without drawing much attention or causing potential problems.

8 Limitation and Discussion

In this section, we will discuss the limitations of *SelfFlag* and outline directions for future work.

Copyright Music Pool. The music tracks used in our experiment were all licensed through third-party music platforms and are not included in the free music libraries of YouTube or TikTok. In real-world scenarios, if a coffee shop plays the same copyrighted music and a videographer records it, uploading the video to the *SelfFlag* Cloud Service could result in a failure in face recognition. Therefore, our system requires that the copyrighted music used be unique to prevent it from being played by others. To address this, we offer a selection of free and unique music through the *SelfFlag* database. When bystanders register and bind copyrighted music, they can choose from these tracks.

Momentary Occurrence. The duration of the captured copyrighted music directly affects whether the platform can detect it. According to the experimental results, music needs to be played for a sufficiently long duration to be accurately detected. For bystanders who appear in the video for less than 10 seconds, it is unlikely that the censorship mechanism will detect them. Therefore, for bystanders who appear briefly in the video and still wish to protect their privacy, it may be necessary to combine existing methods proposed in related research.

Multiple Occurrence. In real-world scenarios, the same bystander may appear multiple times in a video, and not every appearance lasts long enough for detection. To more effectively protect the bystander's privacy, the *SelfFlag* Cloud Service can be improved to identify bystanders not only in video segments with copyrighted music but also in the frames before and after the detected segments. This ensures that the bystander's privacy information is completely removed.

Object Bystanders. In addition to facial information, the content to be protected can also include general objects, such as items in an exhibition or special products. In these cases, copyrighted music can be associated with these items. By leveraging more powerful recognition algorithms in the *SelfFlag* Cloud Service, a more general form of privacy protection can be enabled.

Other Applications of Platform Censorship. Using copyrighted music to trigger censorship is just one of many possibilities we have explored. Building on censorship mechanisms, there are numerous other potential applications, such as embedding images or watermarks that could trigger censorship in online chats or video conferences to prevent unauthorized sharing.

9 Conclusion

In this paper, we explore a method of repurposing the content censorship mechanisms of video-sharing platforms to enable bystander privacy protection. Additionally, we have conducted thorough testing and analysis of the censorship mechanisms employed by major video-sharing platforms, with the hope that these insights will benefit others with similar interests. Despite the varying attitudes of different platforms toward content censorship, there are significant performance and functional differences in their implementations. We believe that a transparent and efficient censorship mechanism is crucial for trustworthy information distribution and effective copyright protection. As such, we foresee the core concept of *SelfFlag* becoming increasingly relevant as content censorship mechanisms continue to evolve and improve in accuracy and efficiency.

Acknowledgments

We sincerely thank the anonymous reviewers for their comments and suggestions. This work is supported by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007) and Guangzhou-HKUST(GZ) Joint Funding Program (SL2024A03J01192).

References

[1] Paarijaat Aditya, Rijurekha Sen, Peter Druschel, Seong Joon Oh, Rodrigo Benenson, Mario Fritz, Bernt Schiele, Bobby Bhattacharjee, and Tong Tong Wu. 2016. I-Pic: A Platform for Privacy-Compliant Image Capture. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services* (Singapore, Singapore) *(MobiSys '16)*. Association for Computing Machinery, New York, NY, USA, 235–248. https://doi.org/10.1145/2906388.2906412

[2] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett I. Bertenthal, and Apu Kapadia. 2020. Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*. IEEE, 1350–1366. https://doi.org/10.1109/SP40000.2020.00006

[3] Meng Ao, Dong Yi, Zhen Lei, and Stan Z Li. 2009. Face recognition at a distance: system issues. *Handbook of Remote Biometrics: For Surveillance and Security* (2009), 155–167.

[4] Cheng Bo, Guobin Shen, Jie Liu, Xiang-Yang Li, YongGuang Zhang, and Feng Zhao. 2014. Privacy.tag: privacy concern expressed and respected. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems* (Memphis, Tennessee) *(SenSys '14)*. Association for Computing Machinery, New York, NY, USA, 163–176.

[5] Laura Ceci. 2024. Distribution of TikTok users worldwide. https://www.statista.com/statistics/1299771/tiktok-global-user-age-distribution/

[6] Laura Ceci. 2024. Distribution of YouTube users worldwide. https://www.statista.com/statistics/1287137/youtube-global-users-age-gender-distribution.

[7] Yuxin Chen, Huiying Li, Shan-Yuan Teng, Steven Nagels, Zhijing Li, Pedro Lopes, Ben Y. Zhao, and Haitao Zheng. 2020. Wearable Microphone Jamming. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376304

[8] Sergiu Ciumac. 2024. Audio/Video fingerprinting and recognition in .NET. https://github.com/AddictedCS/soundfingerprinting.

[9] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the Eleventh International Conference on Web and Social Media, ICWSM 2017, Montréal, Québec, Canada, May 15-18, 2017*. AAAI Press, 512–515.

[10] Fackbook. 2024. Photos or Videos That Violate Your Privacy. https://www.facebook.com/help/428478523862899/?helpref=related_articles.

[11] Adam Geitgey. 2018. Face Recognition. https://github.com/ageitgey/face_recognition.

[12] Robert O. Gjerdingen and David Perrott. 2008. Scanning the Dial: The Rapid Recognition of Music Genres. *Journal of New Music Research* 37, 2 (2008), 93–100. arXiv:https://doi.org/10.1080/09298210802479268 https://doi.org/10.1080/09298210802479268

[13] Sound Understanding group. Jul. 06, 2024. AudioSet. https://research.google.com/audioset/dataset/index.html.

[14] Rakibul Hasan, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2021. Your Photo is so Funny that I don't Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 556, 14 pages. https://doi.org/10.1145/3411764.3445258

[15] Rakibul Hasan, David J. Crandall, Mario Fritz, and Apu Kapadia. 2020. Automatically Detecting Bystanders in Photos to Reduce Privacy Risks. In *2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020*. IEEE, 318–335. https://doi.org/10.1109/SP40000.2020.00097

[16] Rakibul Hasan, Rebecca Weil, Rudolf Siegel, and Katharina Krombholz. 2023. A Psychometric Scale to Measure Individuals' Value of Other People's Privacy (VOPP). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023*, Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (Eds.). ACM, 581:1–581:14. https://doi.org/10.1145/3544548.3581496

[17] Benjamin Henne, Christian Szongott, and Matthew Smith. 2013. SnapMe if you can: privacy threats of other peoples' geo-tagged media and what we can do about it. In *Proceedings of the Sixth ACM Conference on Security and Privacy in Wireless and Mobile Networks* (Budapest, Hungary) *(WiSec '13)*. Association for Computing Machinery, New York, NY, USA, 95–106. https://doi.org/10.1145/2462096.2462113

[18] Romain Hennequin, Anis Khlif, Felix Voituret, and Manuel Moussallam. 2020. Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software* 5, 50 (2020), 2154. https://doi.org/10.21105/joss.02154 Deezer Research.

[19] Josh Howarth. 2024. 30+ Incredible Creator Economy Statistics (2024). https://explodingtopics.com/blog/creator-economy-stats.

[20] Instagram. 2024. Instagram's Community Guidelines. https://www.facebook.com/help/instagram/477434105621119.

[21] Lei Ke, Martin Danelljan, Henghui Ding, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. 2023. Mask-Free Video Instance Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 22857–22866. https://doi.org/10.1109/CVPR52729.2023.02189

[22] Mohamed Khamis, Habiba Farzand, Marija Mumm, and Karola Marky. 2022. DeepFakes for Privacy: Investigating the Effectiveness of State-of-the-Art Privacy-Enhancing Face Obfuscation Methods. In *Proceedings of the 2022 International Conference on Advanced Visual Interfaces* (Frascati, Rome, Italy) *(AVI '22)*. Association for Computing Machinery, New York, NY, USA, Article 21, 5 pages. https://doi.org/10.1145/3531073.3531125

[23] Ang Li, Qinghua Li, and Wei Gao. 2016. PrivacyCamera: Cooperative Privacy-Aware Photographing with Mobile Phones. In *13th Annual IEEE International Conference on Sensing, Communication, and Networking, SECON 2016, London, United Kingdom, June 27-30, 2016*. IEEE, 1–9. https://doi.org/10.1109/SAHCN.2016.7733008

[24] Yifang Li, Nishant Vishwamitra, Bart P. Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and Users' Experience of Obfuscation as a Privacy-Enhancing Technology for Sharing Photos. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 67 (Dec. 2017), 24 pages. https://doi.org/10.1145/3134702

[25] J. Nathan Matias. 2016. Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 1138–1151. https://doi.org/10.1145/2858036.2858391

[26] Meta. 2024. Providing context on sensitive or misleading content. https://transparency.meta.com/zh-cn/enforcement/taking-action/context-on-sensitive-misleading-content/.

[27] Muhammad Shujaat Mirza, Labeeba Begum, Liang Niu, Sarah Pardo, Azza Abouzied, Paolo Papotti, and Christina Pöpper. 2023. Tactics, Threats & Targets: Modeling Disinformation and its Mitigation. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society. https://www.ndss-symposium.org/ndss-paper/tactics-threats-targets-modeling-disinformation-and-its-mitigation/

[28] Reliable Music. 2022. Crowded Streets - People Walking - Noisy Environment. https://www.aigei.com/item/ren_lei_xing_4.html.

[29] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman M. Sadeh. 2017. Privacy Expectations and Preferences in an IoT World. In *Thirteenth Symposium on Usable Privacy and Security, SOUPS 2017, Santa Clara, CA, USA, July 12-14, 2017*. USENIX Association, 399–412. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/naeini

[30] Rajalakshmi Nandakumar, Alex Takakuwa, Tadayoshi Kohno, and Shyamnath Gollakota. 2017. CovertBand: Activity Information Leakage using Music. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 87 (sep 2017), 24 pages. https://doi.org/10.1145/3131897

[31] Frank Pallas, Max-Robert Ulbricht, Lorena Jaume-Palasí, and Ulrike Höppner. 2014. Offlinetags: a novel privacy approach to online photo sharing. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI EA '14)*. Association for Computing Machinery, New York, NY, USA, 2179–2184.

[32] Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. 2023. Lambretta: Learning to Rank for Twitter Soft Moderation. In *44th IEEE Symposium on Security and Privacy, SP 2023, San Francisco, CA, USA, May 21-25, 2023*. IEEE, 311–326. https://doi.org/10.1109/SP46215.2023.10179392

[33] Jerome Pesenti. 2021. An Update On Our Use of Face Recognition. https://about.fb.com/news/2021/11/update-on-use-of-face-recognition/.

[34] Moo-Ryong Ra, Seungjoon Lee, Emiliano Miluzzo, and Eric Zavesky. 2017. Do Not Capture: Automated Obscurity for Pervasive Imaging. *IEEE Internet Computing* 21, 3 (2017), 82–87.

[35] Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid Transformers for Music Source Separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096956

[36] Nirupam Roy, Haitham Hassanieh, and Romit Roy Choudhury. 2017. BackDoor: Making Microphones Hear Inaudible Sounds. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services* (Niagara Falls, New York, USA) *(MobiSys '17)*. Association for Computing Machinery, New York, NY, USA, 2–14. https://doi.org/10.1145/3081333.3081366

[37] Nirupam Roy, Sheng Shen, Haitham Hassanieh, and Romit Roy Choudhury. 2018. Inaudible voice commands: the long-range attack and defense. In *Proceedings of the 15th USENIX Conference on Networked Systems Design and Implementation* (Renton, WA, USA) *(NSDI'18)*. USENIX Association, USA, 547–560.

[38] Brennan Schaffner, Arjun Nitin Bhagoji, Siyuan Cheng, Jacqueline Mei, Jay L Shen, Grace Wang, Marshini Chetty, Nick Feamster, Genevieve Lakier, and Chenhao Tan. 2024. "Community Guidelines Make this the Best Party on the Internet": An In-Depth Study of Online Platforms' Content Moderation Policies. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 486, 16 pages. https://doi.org/10.1145/3613904.3642333

[39] Filipo Sharevski, Amy Devine, Emma Pieroni, and Peter Jachim. 2022. Folk models of misinformation on social media. *arXiv preprint arXiv:2207.12589* (2022).

[40] Mahnoor Sheikh. 2024. 50+ Social Media Video Marketing Statistics for 2024. https://sproutsocial.com/insights/social-media-video-statistics/.

[41] Rohit Shewale. 2024. YouTube Statistics. https://www.demandsage.com/youtube-stats/.

[42] Jiayu Shu, Rui Zheng, and Pan Hui. 2018. Cardea: context-aware visual privacy protection for photo taking and sharing. In *Proceedings of the 9th ACM Multimedia Systems Conference* (Amsterdam, Netherlands) *(MMSys '18)*. Association for Computing Machinery, New York, NY, USA, 304–315. https://doi.org/10.1145/3204949.3204973

[43] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. In *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023, Delft, Netherlands, July 3-7, 2023*. IEEE, 868–895. https://doi.org/10.1109/EuroSP57164.2023.00056

[44] Mohit Singhal, Chen Ling, Pujan Paudel, Poojitha Thota, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2023. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. In *8th IEEE European Symposium on Security and Privacy, EuroS&P 2023, Delft, Netherlands, July 3-7, 2023*. IEEE, 868–895.

[45] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. 2019. PrivacEye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 26, 10 pages. https://doi.org/10.1145/3314111.3319913

[46] Jose M Such and Natalia Criado. 2018. Multiparty privacy in social media. *Commun. ACM* 61, 8 (2018), 74–81.

[47] Yuanyi Sun, Shiqing Chen, Sencun Zhu, and Yu Chen. 2020. iRyP: a purely edge-based visual privacy-respecting system for mobile cameras. In *Proceedings of the 13th ACM Conference on Security and Privacy in Wireless and Mobile Networks* (Linz, Austria) *(WiSec '20)*. Association for Computing Machinery, New York, NY, USA, 195–206. https://doi.org/10.1145/3395351.3399341

[48] Ram Sundara Raman, Prerana Shenoy, Katharina Kohls, and Roya Ensafi. 2020. Censored Planet: An Internet-wide, Longitudinal Censorship Observatory. In

Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (Virtual Event, USA) *(CCS '20)*. Association for Computing Machinery, New York, NY, USA, 49–66. https://doi.org/10.1145/3372297.3417883

[49] Siyuan Tang, Xianghang Mi, Ying Li, XiaoFeng Wang, and Kai Chen. 2022. Clues in Tweets: Twitter-Guided Discovery and Analysis of SMS Spam. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Los Angeles, CA, USA) *(CCS '22)*. Association for Computing Machinery, New York, NY, USA, 2751–2764. https://doi.org/10.1145/3548606.3559351

[50] TP-Link. 2024. About Face Monitoring. https://security.tp-link.com.cn/service/detail_article_4513.html.

[51] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. 2021. Towards Understanding and Detecting Cyberbullying in Real-world Images. In *28th Annual Network and Distributed System Security Symposium, NDSS 2021, virtually, February 21-25, 2021*. The Internet Society.

[52] Maximiliane Windl and Sven Mayer. 2022. The skewed privacy concerns of bystanders in smart environments. *Proceedings of the ACM on Human-Computer Interaction* 6, MHCI (2022), 1–21.

[53] Anran Xu, Shitao Fang, Huan Yang, Simo Hosio, and Koji Yatani. 2024. Examining Human Perception of Generative Content Replacement in Image Privacy Protection. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 777, 16 pages. https://doi.org/10.1145/3613904.3642103

[54] YouTube. 2024. YouTube's Community Guidelines. https://support.google.com/youtube/answer/9288567?sjid=15820223809265265066-AP.

[55] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. DolphinAttack: Inaudible Voice Commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17)*. Association for Computing Machinery, New York, NY, USA, 103–117. https://doi.org/10.1145/3133956.3134052

[56] Lan Zhang, Kebin Liu, Xiang-Yang Li, Cihang Liu, Xuan Ding, and Yunhao Liu. 2016. Privacy-friendly photo capturing and sharing system. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Heidelberg, Germany) *(UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 524–534.

[57] Tengfei Zheng, Tongqing Zhou, Qiang Liu, Kui Wu, and Zhiping Cai. 2022. Characterizing and Detecting Non-Consensual Photo Sharing on Social Networks. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (Los Angeles, CA, USA) *(CCS '22)*. Association for Computing Machinery, New York, NY, USA, 3209–3222. https://doi.org/10.1145/3548606.3560571

[58] Shangchen Zhou, Chongyi Li, Kelvin C. K. Chan, and Chen Change Loy. 2023. ProPainter: Improving Propagation and Transformer for Video Inpainting. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 10443–10452. https://doi.org/10.1109/ICCV51070.2023.00961

Appendix

The questionnaire aims to understand your preferences regarding bystander privacy protection:

(1) Do you think people should anonymize (*e.g.*, blur, pixelate, remove) your face and/or body when they publicly share a video that includes your information on the Internet?
   - Yes
   - No (GOTO 7)

(2) If you want to prevent videographers from publicly sharing videos that include you, what actions have you taken? (Multiple choices)
   - Immediately inform the videographer when you discover you are being filmed (requesting that they stop filming or blur you)
   - After encountering a video that includes you, file a complaint on the sharing platform
   - Wear physical labels (*e.g.*, "Blur Me", "No Photos") on your clothing
   - Take no actions
   - Other

(3) If playing music on your phone could notify the videographer to avoid sharing videos that contain your face, would you be willing to use this method?
   - Yes, using the phone's speaker to play audible music
   - Yes, using an ultrasonic speaker to play inaudible music
   - Yes, both audible or inaudible music are acceptable
   - No (GOTO 5)

(4) What are the main reasons you would be willing to use this method? Please specify its advantages and disadvantages. (GOTO 6)

(5) What are the main reasons you would be unwilling to use this method?

(6) If the following methods could prevent videographers from publicly sharing videos that include you, which one would you prefer? (Rank in order)
   - Immediately inform the videographer when you discover you are being filmed (requesting that they stop filming or blur you)
   - After encountering a video that includes you, file a complaint on the sharing platform
   - Wear physical labels (*e.g.*, "Blur Me", "No Photos") on your clothing
   - Continuously play audible music
   - Continuously play inaudible music with an ultrasonic speaker

(7) Have you ever shared a recorded video on sharing platforms?
   - Yes
   - No (GOTO 11)

(8) As a videographer, what measures have you taken to protect the privacy of the bystanders in your videos?
   - Blurred the faces of all bystanders
   - Blurred the faces of those who informed me
   - Took no actions

(9) As a videographer, what factors most motivate you to protect bystanders' privacy? (Rank in order)
   - Strict self-discipline: by default, blur all bystanders' information, regardless of whether they have such a request
   - The bystander explicitly informs you of their request in person
   - When you upload a video, the video sharing platform informs you that your video may contain private information about other people that they do not wish to be filmed

(10) If you know that the video you shared includes individuals who do not wish to be filmed, would you be willing to edit and reupload the video to address this issue (the editing process may require software, time, and computational resources)?
   - Yes
   - No

(11) Your gender:
   - Male
   - Female

(12) Your age:
   - 20-29
   - 30-39
   - 40-49
   - 50 or above