Risky Gaze: Understanding and Mitigating Photosensitive Risks from User-Generated Content on Video Platforms

***Abstract*—Videos have rapidly become the dominant medium for online communication, with billions of users engaging daily through platforms such as TikTok and YouTube. Yet, this cultural shift introduces new security and accessibility challenges. In particular, attackers have begun exploiting the visual intensity of short videos to target users with photosensitive epilepsy (PSE) through seizure-triggering content. Despite the prevalence of such risks, moderation and detection mechanisms remain fragmented and largely ineffective.**

To understand this emerging threat landscape, we systematically analyze the attack surface of modern video platforms and identify how adversaries can exploit video pipelines—through video source manipulation, creative editing, platform processing, and playback. Using proof-of-concept videos overloaded with visual stimuli, we demonstrate the feasibility of such attacks across 15 popular platforms. Our study reveals that most platforms fail to consistently mitigate PSE-triggering risks, underscoring the urgent need for robust defense mechanisms.

This gap motivates PRISM (Photosensitive Risk Identification and Safety Moderator), an efficient and formally verified content moderating system designed to identify seizure-triggering videos. We conclude by discussing the broader implications for accessibility, safety, and responsible moderation in an era dominated by user-generated content.

1. Introduction

Video platforms have become a de facto form of communication, connecting individuals within and across communities, with short videos revolutionizing how content is made and consumed. It is estimated that 3.7 million videos are uploaded to platforms such as YouTube daily, amounting to more than 518,000 hours of new content [1]. While this brings social and economic benefits, it also expands the attack surface for malicious or harmful content. For most users, such harm takes textual or psychological forms, including hate speech or harassment [2]–[6]. However, for individuals with photosensitive epilepsy (PSE), exposure to certain visual patterns—particularly high-contrast flashes or rapid luminance changes—can trigger seizures, turning visual media into a vector of physical harm.

The Pathology of Photosensitive Epilepsy and the Physical Impact of Exploitation. Photosensitive epilepsy is a condition in which exposure to certain visual stimuli, usually ones that exhibit cyclical patterns in time and space, can trigger the onset of tonic-clonic seizures [7]. These seizures can lead to respiratory depression, cardiac arrhythmia, or cerebral depression, leading to sudden unexpected death in epilepsy (SUDEP) [8], [9]. The pathological root cause of these seizures is due to neurons becoming overly excited and synchronizing abnormally in response to visual stimulus in one part (focal onset) or the entire brain (generalized onset) [10]. Contrary to the popular belief that these conditions are rare, they impact 5% of the epileptic population, and can be triggered by flickering lighting or rapidly changing contrasting patterns [11].

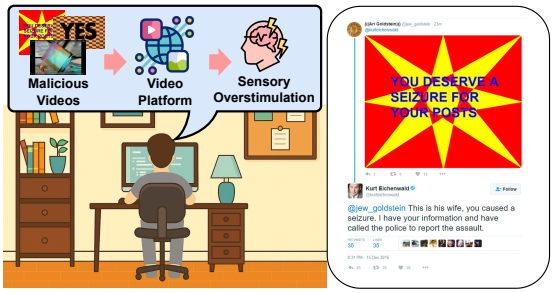


Figure 1: Overview of the risk of user-generated content.

As multimedia becomes the primary means of communication over the Internet, reports have emerged of attackers deliberately inflicting physical harm on susceptible individuals. In 2016, journalist Kurt Eichenwald was attacked online with a strobe GIF sent via Twitter (Figure 1). The deliberate act caused a real seizure, marking one of the first cases where a digital message was used as a physical weapon [12]. The attacker was charged with assault, setting a legal precedent for treating certain online actions as bodily harm [13]. Unfortunately, this was not an isolated case [14], [15]. These incidents highlighted the overlooked dangers of digital spaces for vulnerable individuals[1]. It also marked a turning point in recognizing the need for accountability and safety in online communication.

Gap in Existing Protection for Users with Photosensitive Epilepsy. The first attempt to address the issue of seizure-triggering videos was prompted by the well-known incident

1. (a) Twitter GIF attack: https://newatlas.com/computers/epilepsy-foundation-twitter-strobe-seizure-gifs-law/ (b) Pot Noodle ad: https://www.youtube.com/watch?v=Bs0WCUZqoJg (c) Citroën ad: https://www.youtube.com/watch?v=47PXfNwa8iA **(Warning: These videos may cause visual discomfort for viewers)**

involving Pokémon. In 1997, an episode of the Pokémon anime triggered seizures in hundreds of viewers in Japan due to rapidly flashing red and blue lights [16]. This eventually led to the international broadcasting guideline ITU-R BT.1702-2 in 2019 [17]. Later on, W3C has developed a new set of guidelines under its Web Accessibility Initiative [18]. These updated guidelines provide more detailed recommendations on how content should be moderated to avoid visuals that induce seizures. However, as user-generated video platforms, such as TikTok and YouTube, become an increasingly integral part of everyday communication, it is crucial to understand how they change the threat landscape and develop effective defenses to protect users with PSE.

Our Approach - PRISM. Recognizing this need, we developed PRISM, a content moderator that effectively and efficiently detects seizure-triggering videos. Several challenges in designing and implementing PRISM are:

(1) Analyzing Attack Surfaces and Testing Platform Defenses. To ensure that our defense remains robust amid the growing volume of user-generated video content, we first performed a security analysis to enumerate the attack surfaces that adversaries could exploit to harm viewers. We surveyed 15 popular video platforms to construct an abstract model of the content delivery pipeline, from video creation and editing to processing and playback. Using proof-of-concept attack videos, we found that defenses are either absent or insufficient on most platforms. Our analysis highlights both the mechanisms attackers could exploit to introduce epilepsy-triggering content and the constraints they must navigate. We responsibly disclosed our findings to the platforms and found the pressing need for effective defenses against this less-studied threat vector.

(2) Bridging the Gap between Human Physiology and Cyber Protection. Unlike malware, such overstimulation of the senses does not necessarily violate any conventional security policy—such as unauthorized file access or software exploitation—but instead aims to harm users physically by inducing neurophysiological responses. Consequently, detecting such violations must be grounded in an understanding of the underlying pathology. While our initial approach followed the Web Content Accessibility Guidelines (WCAG) [18], collaboration with medical experts revealed that the diversity of modern playback environments introduces additional challenges to enforcing these standards. Unlike prior implementations, PRISM considers variations in usage scenarios and playback conditions that alter a flashing video's brightness, color, area, and frequency.

(3) Using Formal Methods to Ensure Implementation Correctness. Due to the complexity of the guideline, we found that it is not straightforward to ensure the correctness of the implementation, as suggested by our evaluation of existing tools in § 7. To ensure the soundness of the bridge between the semantic gap from psychophysiology to software implementation, we formally specified the policies with Rocq and applied formal verification. This step ensures that the final program complies with the specification and that no misinterpretation was made during program implementation.

Evaluation and Outcome. We evaluated PRISM against an edge-case dataset, testing its robustness against videos with borderline triggering characteristics. We further evaluated existing detectors with the dataset, and found that their implementations present multiple inconsistencies with the most up-to-date guidelines, allowing an adaptive attacker to bypass the defense. We then conducted experiments on PRISM's runtime performance, and observed a per-frame processing time faster than the standard display frame rate. Our implementation with GPU codes achieved notable speedup compared to the CPU implementation, taking advantage of parallel processing under the video platform setting.

Contributions. To the best of our knowledge, we make the first step toward studying and detecting seizure-triggering content on video platforms while considering manipulations during content delivery. Our contributions are:

- We systematically analyze the attack surface of modern video platforms and empirically demonstrate the presence of vulnerability in 15 modern video platforms. We further study existing seizure-triggering video detection tools, and find that their implementations do not match the most up-to-date guidelines.
- We develop PRISM, a policy-based video content monitor that faithfully identifies potentially seizure-triggering videos. We also incorporate detection considerations essential to video platform scenarios, such as variations in playback conditions.
- To ensure the correctness of our implementation with respect to the guideline, policies are formally specified with Rocq. PRISM is then formally verified to ensure that the implementation remains consistent with the guideline. All of our codes and specifications are open-sourced to the community[2]. As datasets used in the experiments contain seizure-triggering materials, access will be granted upon individual requests.

2. Background and Motivation

Photosensitive Epilepsy (PSE). As a type of reflex epilepsy, PSE can be triggered by specific visual stimuli, such as lights flashing at particular brightness, color, or frequency [10]. Common trigger sources include video games, television, computer screens, and certain types of light sources, such as sunlight or strobe lights from emergency vehicles. According to the World Health Organization, a record of 50 million people worldwide suffer from epilepsy [19], with approximately 5% of patients being photosensitive [20]. Symptoms include headaches, migraines, nausea, staring, loss of consciousness, tonic-clonic seizures, and even death [7], [9]. Many patients may not be aware of their condition until symptoms manifest, and most do not outgrow the condition [21].

Emerging Popularity and Safety Concerns of Video Platforms. Consumer habits in content viewing have changed

2. https://anonymous.4open.science/r/PRISM-E7A9

drastically in recent years. A 2021 survey showed that more than 71% of the respondents watch video content on video platforms with their mobile devices instead of traditional media [22]. Major video platforms, such as TikTok, were estimated to reach over 2.35 billion global users by 2029 [23]. The shift to video platforms as the main source of information consumption has prompted content creators to produce eye-catching videos for increased social media influence or to generate short videos as advertisements to promote products. However, these videos can sometimes contain content that is inappropriate or harmful to certain viewers. Particularly related to PSE, seizure-triggering content may either be created and uploaded by a malicious actor or produced by content creators who are simply unaware of the issue. For instance, Conti et al. [24] first discussed such cases in which attackers use sensory and perceptual stimuli to specifically harm individuals with PSE. Later, Mcnutt et al. [25] regarded strobe visualization as capable of inflicting direct sensory violence when discussing design spaces for villainy visualizations. As a result, video platforms should assume the responsibility to perform content moderation for photosensitive epileptic users and prevent harmful sensory overload. It is imperative to examine the attack surface of modern video platforms - as we do in this paper.

Prior Regulatory Efforts on Seizure-Triggering Content. As incidents involving photosensitive epilepsy due to digital media began to surface, the medical community has since been conducting clinical studies to identify risk factors [7], [10], [11], [26]–[28]. Binnie et al. [11] first revised the UK Independent Television Commission (ITC) Guidance Note [29], which details the characteristics of visual stimuli that can have devastating effects on patients, eventually adopted internationally as ITU-R BT.1702-2 [17]. However, existing guidelines largely focus on the protection of TV broadcast content. As content consumption moved to the Internet, few consumer protection authorities have focused on the moderation of seizure-triggering content online. For instance, the US Federal Communications Commission regulates broadcast content, but not online content [30]. While the W3C proposed the Web Content Accessibility Guidelines (WCAG) to advise developers on generating accessible web pages [18], patients are still encountering triggering videos online. Meanwhile, existing guidance are also not fully consistent nor easy to adapt [21], leading to the concerning outcome of video platforms not correctly identifying seizure-triggering contents.

Challenges in Implementing Medical Guidelines and Secure Defense. Historically, many programs and advertisements that can trigger seizures have been televised even with broadcast guidelines in place [12], [31]–[35]. Though there have been existing efforts to develop tools that help users avoid seizure-triggering content [36], [37], the safety guarantee is either up to the developer's discretion or only heuristically evaluated. Incorrect interpretation and oversimplification of the guidelines are often critical reasons that lead to questionable moderation quality [32]. In addition, triggers are highly dependent on the way content is being received by the viewers and are not just limited to the content itself. Factors such as display hardware and playback settings (e.g., video orientation, user viewing distance, etc.) can significantly affect whether a video is triggering to the viewer or not. Furthermore, the interaction between users and the displayed materials, which is common on video social platforms, may also lead to harmful outcomes.

3. Threat Model

Attacker Motivation & Capability. The goal of the attacker is to cause physical discomfort or harm to the targeted users who suffer from photosensitive epilepsy. Such attacks can be motivated by hatred or Internet trolling [12], [15]. From the perspective of attacker capability, since the focus of our work is defense, we assume a strong attacker who already has a mechanism to deliver the malicious video to the target users. This can be done using direct communication to a specific user, such as in the case of *Eichenwald v. Rivello* [13], or through platform-specific group sharing, such as the attack on Epilepsy Foundation [35].

Security Goal and Assumptions. The goal of PRISM is to detect videos that may trigger photosensitive epilepsy. We assume that, as the owner of the video distribution platform, the defender has the ability to inspect any video that is delivered to users. We also assume that faithfully following established guidelines is sufficient to reduce the risk of photosensitive epilepsy for most viewers.

Because the onset of photosensitive epilepsy can also depend on the viewing device, the screen size, surrounding lighting conditions, and other environmental factors, we further assume that the platform operates within a predictable deployment environment. In other words, the platform provider knows the set of devices it supports and the typical viewing conditions. This is consistent with standard product design practices, where systems are evaluated based on an expected operating environment.

4. Video Platform Content Delivery Pipeline

To characterize the attack surface, we conducted a systematic survey of 15 popular video platforms and analyzed their content delivery pipelines in detail. Our goal was to understand how an attacker could inject seizure-triggering stimuli into videos and deliver them to potential victims. During this process, we also identified several practical constraints that an attacker must consider when attempting to exploit the system. These insights further informed and refined the design of our detection mechanism.

Pipeline Overview. To reason effectively about attacks and defenses, we first abstract the end-to-end video processing pipeline used by modern multimedia platforms, as shown in Figure 2. A video begins at the source stage, where it is produced or uploaded. Once received by the platform, the video undergoes a series of internal transformations such as transcoding, resizing, and frame-rate conversion; these operations normalize videos into formats suitable for

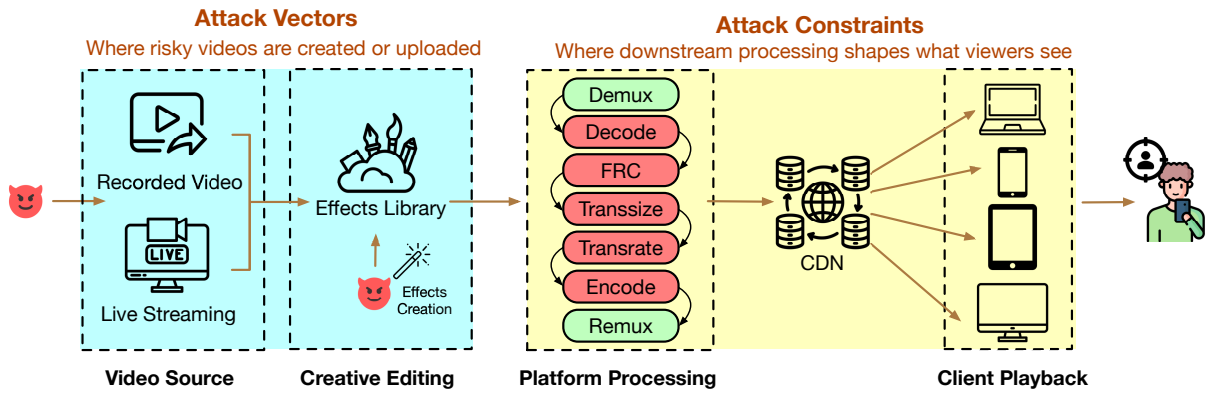


Figure 2: Video platform content delivery pipeline. The first two stages show where risky videos can be created or uploaded, while later stages illustrate how processing and playback transformations affect their visual properties—factors that both attacks and defenses must account for.

distribution. After processing, the platform distributes the transformed video to viewers. Finally, the video is rendered on client devices, where factors such as screen size, playback speed, and viewing environment influence the visual output experienced by the user. With this pipeline defined, we can now reason about both attacker capabilities and attacker limitations. The first two stages—the video source and the creative editing —expose the attack surface: a malicious actor can directly upload a malicious video or apply harmful visual effects. The latter stages, however, impose constraints that the attacker has to consider carefully during the exploitation process, since platform processing may alter or discard visual patterns, while client-side rendering introduces additional variability through device characteristics, playback settings, and the viewing environment.

4.1. Attack Vector - Video Source

An attacker can directly manipulate a source video at the pixel and frame level to induce seizures. By altering pixel values, the number of concurrently flashing pixels, and the count of flashing frames, an attacker can exceed the thresholds that define seizure-triggering content. After manipulating the content of the source video, an attacker can directly propagate it on the Internet via both targeted and untargeted channels, realizing the attack.

Sharing Recorded Videos. One basic function all video platforms have is the ability for users to upload videos to the platforms. An attacker can upload a harmful video and deliver it to the victim viewer browsing the platform. Note that this can either be untargeted or targeted. In an untargeted attack, the attacker can simply upload the harmful video, set it to publicly viewable, and wreak havoc. In a targeted attack, the attacker can upload the video and then send the video link via direct message to the victim, or try to influence the recommendation of the platform, such as including the victim's username in the video description.

Live Streaming. Another untargeted vector that can be taken advantage of is live streaming. Given the real-time capability and some platforms that require a certain number of followers to enable streaming, the impact of attacks delivered through live videos is profound and immediate. An attacker can directly broadcast triggering content with livestreams to reach a significant number of viewers.

4.2. Attack Vector - Creative Editing

Platforms promoting short-form videos allow users to apply special effects to pre-recorded or live-recorded videos. While these features promote creativity and the diversity of user-generated content, they can also turn benign videos into seizure-triggering ones. Specifically, attackers can exploit creative editing features in two ways: (i) by uploading harmful effects that directly trigger seizures when used by victims during editing, or (ii) by submitting seemingly benign effects that remain harmless in isolation but become harmful when applied to other benign videos. Note that the benign video can either be a truly benign video owned by the victim, or a seemingly benign one offered by the attacker (such as in the case of re-posting). While there are many variations of special effects, those can be further characterized by whether they depend on the original content and by the nature of their modification, as illustrated in Figure 3.

Leveraging Content-Agnostic Effects. This type of effect modifies the original video regardless of its content. These have visual elements that overlay the original video and do not interact with the objects or people in the video. An overlayed content-agnostic effect is characterized as an animation of arbitrary 2D shapes or virtual 3D objects overlaying either the original video or a real-world scene captured on camera, similar to artificial reality. An attacker may directly overlay flashing components with this effect. Besides these 2D and 3D overlay effects, a transformative content-agnostic effect alters the original video entirely. For instance, it may duplicate the content to create a collage or transform the video to be played in slow motion. Even so, the way the effects modify the video is not dependent on its content but is universal across videos.

Leveraging Content-Dependent Effects. This type of effect modifies the original video based on the objects that

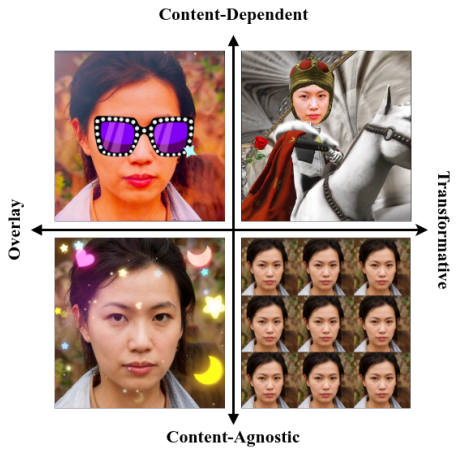


Figure 3: Examples of effects from creative editing.

are detected in the video. Most of these would first need to obtain the presence and location of a face. With an overlayed content-dependent effect, the face may then be augmented with virtual accessories. An attacker may overlay flashing components targeting the face area with this effect. With a transformative content-dependent effect, the face may be cartoonized, transplanted into other scenes, or the person may be left intact while the background is replaced. This kind of effect offers the most manipulative freedom to the attackers, as it alters the original content extensively.

4.3. Attack Constraints - Processing and Playback

Video platforms perform transcoding to convert uploaded files into formats suitable for efficient storage, streaming, and playback. This process enables adaptive bitrate streaming, which dynamically switches between resolutions and bitrates based on network conditions [38]. While necessary for usability, each stage in processing and playback can alter pixels, frames, and timing properties in ways that affect both attack feasibility and defense reliability.

Codec Re-encoding. Re-encoding applies compression that changes luminance, contrast, and fine-grained pixel patterns. This codec-induced transformation may destroy finely tuned attack cues crafted by the attacker. They may also shift pixel intensities in ways that trigger safety thresholds. Consequently, defenders must account for codec variability by validating content across representative codec profiles.

Frame Rate Conversion (FRC). FRC alters a video's temporal structure and can change perceived flicker frequency. Dropped frames can remove transient flashes; duplicated frames introduce artificial temporal repetition that may amplify low-frequency flicker; and interpolation can introduce new intermediate frames and temporal frequencies that are hazardous even when the source is safe. Common playback rates include 24, 30, and 60 fps, and conversions between them should be considered when validating safety.

Transsizing & Transrating. Resizing and bitrate adjustments change spatial resolution and pixel fidelity, which can affect the proportion and intensity of flashing regions. The chosen bitrate usually depends on the video resolution that is transsized (e.g., YouTube recommends 1080p videos to be processed at 8Mbps for 30fps).

TABLE 1: Attack Vectors, Constraints, and Flash Properties

			Flash Properties Impacted			
			Luminance	Color	Area	Frequency
Attack Vector	Video Source	Content Pixel Value	✓	✓		
		# of Flashing Pixels			✓	
		# of Flashing Frames				✓
	Creative Editing	Content-Agnostic Effects	✓	✓	✓	✓
		Content-Dependent Effects	✓	✓	✓	✓
Attack Constraint	Platform Processing	Codec	✓	✓	✓	✓
		Framerate				✓
		Resolution			✓	
		Bitrate	✓	✓		✓
	Client Playback	Display Type	✓			
		Viewing Distance			✓	
		Display Rotation			✓	
		Playback Speed				✓
		Screen Brightness	✓	✓		

Viewing Conditions. ITU-R BT.1702-2 [17] suggested that HDR content may be more problematic than SDR, and a device supporting HDR is potentially more troublesome. Viewing distance, device resolution, and screen size affect the viewer's visual field on the screen, which in turn affects the area and intensity of flashes received by the visual cortex [39]. On the other hand, a video can be triggering when played horizontally on a desktop, but appear safe when oriented vertically on a mobile phone due to scaling and letterboxing, which affect the flash area.

Playback Settings. Varying the playback speed and brightness may lead to harmful results, as these affect the flash frequency, luminance, and color. For instance, the perceived saturation of the same displayed color may depend on the brightness display setting of the screen, with a visual loss of saturation at low luminance (Hunt effect) [40].

5. Empirical Analysis of Attack Surface

Building on top of our understanding of the attack vectors and constraints on the content delivery pipeline, we systematically validate the feasibility of each attack vector and constraint on 15 popular video platforms, which are shown in Table 2. Since presenting epilepsy-triggering videos to human subjects may bring potential bodily harm, we leverage existing video auditing guidelines to develop PRISM to help determine if a video exhibits seizure-triggering capabilities. More specifically, according to Binnie et al. [11] and existing guidelines [17], [18], [29], epileptic triggers are characterized by four major properties: luminance, color, area, and frequency of flashes. The detailed parameters and implementation are discussed in §6. We also employed the Photosensitive Epilepsy Analysis Tool (PEAT) [36] to validate the proof-of-concepts graphically in Appendix A. We refer to §11 for the ethics considerations.

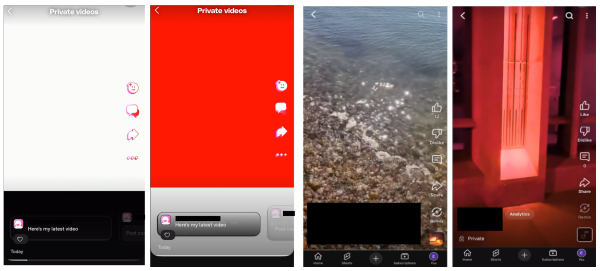
5.1. Evaluating Video Sources Protection

Experiments. To demonstrate the feasibility of manipulating video sources, we manually crafted two videos in which a square object flashes in the video. We also manually

TABLE 2: Empirical Results of Platform Investigation

Platform Name	Video Sharing			Live Streaming	Creative Editing		Playback Settings (Mobile App)			Playback Settings (Website)	
	Video Upload	Direct Message	Tag Username		Effects Upload	Effect-Video Combination	Speed Adjustment	Video Orientation	Resolution Adjustment	Speed Adjustment	Resolution Adjustment
Dailymotion	☼●	✗	✗	✗	✗	✗	✗	Vertical	✗	○	○
Flickr	☼●	✗	☼●	✗	✗	✗	✗	Adaptive	✗	✗	✗
Youtube	☼●	✗	✗	○	✗	○	○	Adaptive	○	○	○
Twitch	N/A	✗	✗	○	✗	✗	○	Adaptive	○	○	○
Bilibili	☼●	✗	✗	○	✗	✗	○	Adaptive	✗	○	○
Rumble	☼●	✗	✗	○	✗	✗	○	Adaptive	○	○	○
Tumblr	☼●	☼●	✗	✗	✗	✗	✗	Adaptive	✗	○	✗
TikTok	☼●	☼●	☼●	○	◐	◐	○	Adaptive	✗	✗	○
Instagram	☼●	☼●	☼●	○	◐*	○	○	Vertical	✗	✗	✗
Snapchat	☼●	☼●	☼●	✗	◐	○	✗	Vertical	✗	✗	✗
Vimeo	☼●	✗	✗	○	✗	✗	○	Adaptive	○	○	○
Facebook Watch	☼●	☼●	☼●	○	✗	○	○	Adaptive	○	○	○
Josh	☼●	☼●	☼●	○	✗	○	✗	Vertical	✗	N/A	N/A
Moj	☼●	☼●	☼●	○	✗	○	✗	Vertical	✗	N/A	N/A
Pinterest	☼●	✗	✗	✗	✗	✗	✗	Adaptive	✗	✗	✗

☼● - Able to upload videos that violate the luminance and color threshold without any checks. ○ - Platform supports such features without defense in place. ◐ - Platform supports such features and has some defense in place, though not comprehensive. ✗ - Platform does not offer this functionality. *Instagram has removed all effects built by creators on January 14, 2025.



(a) (b)

Figure 4: PoC attack using video source. (a) Crafted videos with flashing objects on Dailymotion. (b) Recorded video with flashing waves and building lights on YouTube.

adjust the luminance and color profile of videos with real-world scenes as proof-of-concepts (Figure 4). To evaluate platform-side protection, we programmatically generated 40 videos that violate the guideline[3] using Python and OpenCV. The videos were uploaded to all video platforms to examine whether the platforms implement any countermeasures. We then created an account representing the victim viewer on the platform, leaving all account settings at the default. For platforms that supported direct messaging, we sent the videos from the attacker account to the viewer, without the two having any prior connection. We also tag the viewer's username in the videos' description, thus manipulating the recommendation algorithm to push the videos to the viewer.

Results. All platforms in our empirical analysis do not check for the uploads of seizure-triggering videos. There are also no warnings for videos that may contain triggers. For platforms that support user interactions, the settings are often defaulted to the most lenient options, allowing anyone to message or tag any other users. The attackers on these platforms were able to deliver harmful videos to their target users, with the users having no way to filter.

3. 5 videos violate the standard luminance threshold; 5 violate the HDR luminance threshold; 30 violate the color threshold, with various shapes in different colors. All videos exceeded the area and frequency thresholds, with the frequency set at 20 flashes per second.

5.2. Evaluating Effects from Creative Editing

Experiments. We used PRISM, described in §6, for these validation experiments to enable automation and provide a more realistic evaluation using the latest guidelines. We collected user-generated content-agnostic and content-dependent effects from platforms that support custom effects (TikTok, Instagram, and Snapchat). For *content-agnostic effects*, we overlaid them with a black-framed video and analyzed whether they contain seizure-triggering properties when tested alone. Subsequently, we applied the effects to 45 randomly sampled benign videos from TikTok to examine whether the combination could produce triggering sequences. In the case where the dimensions of the video frames and the effect frames do not match, the effect frames are scaled to match the video's dimensions. Effects are looped to match the duration of the videos. For *content-dependent effects*, we applied these on 10 face images to avoid risking human subjects.

Results. A proof of concept manipulation is shown in Figure 5. Out of the 150 content-agnostic effects collected (93 from TikTok, 40 from Instagram, and 17 from Snapchat), our assessment found 11 TikTok effects and 5 Instagram effects to exhibit photosensitive triggers when tested alone, though these were not marked as harmful by the respective video platforms. A total of 49 effects were triggering when combined with a benign video (TikTok: 30, Instagram: 14, Snapchat: 5). However, these were not triggering when tested alone. Out of the 321 content-dependent effects we collected (199 from TikTok, 50 from Instagram, and 72 from Snapchat), 75 effects were identified to be potentially dangerous (54 on TikTok, 12 on Instagram, and 9 on Snapchat), failing to pass detection when combined with images of a face. None of them has been marked by the respective video

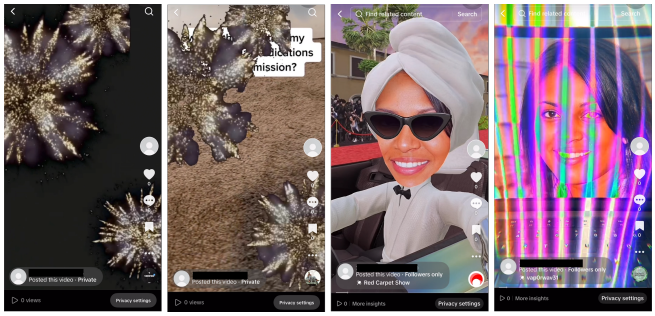


Figure 5: PoC attack using creative editing on TikTok.

platforms, and no warnings were shown when they were selected. While there exist some defense mechanisms on TikTok and Snapchat, as discussed in §5.4, most platforms with creative editing features do not offer any protection.

5.3. Evaluating Impact of Attack Constraints

Experiments. We designed seemingly benign videos whose source content is safe, but which become seizure-triggering after transcoding and playback. We uploaded these videos to each platform and evaluated the protections on both the mobile and website interfaces. To systematically study how playback environments influence the triggering rate of videos, we collected 5,447 trending TikTok videos and applied controlled transformations in resolution, playback speed, orientation, and brightness. We then simulated playback on mobile phones, tablets, and desktop displays at standard viewing distances [41], also comparing HDR and SDR renderings. These large-scale experiments were automated using PRISM, allowing us to empirically quantify how typical transformations can inadvertently alter the visual stimuli's triggering potential.

Platform Evaluation Results. As shown in Table 2, major video platforms lack mechanisms to prevent seizure-triggering content during playback. Although users can adjust playback parameters—such as resolution, speed, and orientation—or rely on adaptive behaviors that respond to network or device conditions, these features are designed for usability rather than safety. Consequently, the potential of such adjustments to induce seizures remains unchecked.

Impact of Playback Devices. Table 3 shows the results of the experiments with device variations. The change from vertical to horizontal orientation significantly decreases failed video rates. This is because short videos like the ones on TikTok are designed to be watched vertically. When played horizontally, the video is scaled and letterboxed, and the region containing dangerous flashes is effectively reduced. Meanwhile, playing vertically on mobile phones and tablets yields roughly similar assessment results, as adjusting dimensions does not greatly affect the percentage of the flashing region related to the whole. However, besides being played horizontally, the drop in total failed videos with desktop screens may be due to the larger viewing distance

TABLE 3: Various Playback Devices on Large-scale Dataset

	Original (Horizontal)	**Mobile (Vertical)**	**Mobile (Horizontal)**
Flash	162 (2.97%)	3048 (55.96%)	853 (15.66%)
Red Flash	11 (0.2%)	110 (2.02%)	45 (0.83%)
Total Failed	166 (3.05%)	3053 (56.05%)	857 (15.73%)
	Tablet (Vertical)	**Tablet (Horizontal)**	**Desktop (Horizontal)**
Flash	3254 (59.74%)	1993 (36.59%)	2256 (41.42%)
Red Flash	118 (2.17%)	74 (1.36%)	87 (1.6%)
Total Failed	3260 (59.85%)	2000 (36.72%)	2262 (41.53%)

We used *"> 25% of full video dimension"* as the flashing area threshold [17] for original videos. Mobile: 1440*2960 / 6 in. screen / 12 in. viewing distance. Tablet: 1536*2048, 8 in. screen / 16 in. viewing distance. Desktop: 1080*1920 / 24 in. screen, 37 in. viewing distance.

TABLE 4: Various Playback Settings on Large-scale Dataset

	Original		**Resolution**	
	SDR	**HDR**	**720p**	**1080p**
Flash	513 (9.42%)	521 (9.56%)	435 (7.99%)	434 (7.97%)
Red Flash	37 (0.68%)	37 (0.68%)	37 (0.68%)	37 (0.68%)
Total Failed	519 (9.53%)	527 (9.68%)	441 (8.1%)	440 (8.08%)
	Playback Speed		**Brightness Adjustment**	
	0.5x	**2x**	**-50%**	**+50%**
Flash	426 (7.82%)	709 (13.02%)	132 (2.42%)	491 (9.01%)
Red Flash	35 (0.64%)	28 (0.51%)	92 (1.69%)	0 (0%)
Total Failed	431 (7.91%)	716 (13.14%)	208 (3.82%)	491 (9.01%)

Original videos are in 540p and vertical orientation. We used *"> 25% of full video dimension"* as the flashing area threshold [17].

that relaxes the threshold, which is calculated based on a 10-degree visual field, as described in §6.3.

Impact of Playback Settings. Table 4 shows the results of the experiments with playback settings. Likewise, changing only the display resolution does not greatly affect the triggering risk of the videos. However, varying the playback speed and brightness leads to significant differences in the results, with videos that are sped up being potentially more dangerous. Meanwhile, 8 more videos were identified to be triggering when played in HDR but not in SDR. While existing tools like PEAT disregarded this variation and marked the video as safe, we can see from the analysis result that the video is borderline dangerous (Figure 10). We later discuss how PRISM can identify these variations correctly.

5.4. Current Defense and Countermeasures

Platform Community Guidelines. Our survey of community guidelines and content moderation policies revealed a lack of sufficient protective measures on video platforms. Most platforms leave the decision of whether a video is harmful to the creator's or viewer's discretion, even if they include explicit clauses related to photosensitive epilepsy in their community guidelines. The community guidelines of TikTok Effect House and Snapchat claim that user-generated effects will undergo a review process before being released for public use, although the procedure is unknown and likely involves manual inspection by dedicated staff, which may lead to potential false negatives, as observed during our investigation. Only Effect House explicitly stated that potentially triggering effects would be marked.

Platform Accessibility Features. Conforming to the claim made in Effect House's community guideline, if an effect may trigger photosensitive epilepsy, a warning page will pop up when users select to apply the effect in TikTok. Videos using the flagged effects are blocked only when the *"Remove photosensitive videos"* accessibility feature, defaulted to off, is toggled on. The user is not informed of this feature at any stage. Videos that do not use a flagged effect are not checked at all, no matter if they contain triggering sequences or not.

6. Defense Design and Formal Specification

Our empirical investigation underscores the pressing need for effective detection policies that can help regulate seizure-triggering videos on video platforms. We now describe the design of PRISM, outlining the policy basis, physiological rationale, and corresponding formal specification.

Policy Basis. WCAG [18] describes a set of policies with which content on the Internet should comply to make the Web more accessible. Since major video platforms are usually available as both web pages and mobile apps, WCAG is applicable. In addition, legal cases [42] regarding website accommodation under the Americans with Disabilities Act (ADA) often reference WCAG. We further consider variations in playback conditions, drawing insights from other guidelines [17], [29] while maintaining consistencies with prior clinical psychophysiology studies [10], [11].

Physiological Foundations. While the exact mechanism of photosensitive epilepsy is still largely unknown and remains an active area of research, clinical studies consistently identify four visual properties as key seizure triggers: sudden changes in luminance, saturated chromatic transitions, large affected visual areas, and rapid flicker frequencies. These physiologically-grounded dimensions align directly with the structure of WCAG policies and therefore guide how PRISM organizes its detection logic. We adhere to the following primary principle for seizure-triggering video detection: *A video is potentially seizure-triggering if there are flashes that exceed the luminance or color threshold, and at the same time, the flashes exceed the area and frequency threshold.*

Formalization Approach. One researcher translated the guidelines into Rocq, which was then verified by two other researchers. These formal specifications were used as blueprints for system implementation. Based on this approach, PRISM acts as a policy-based detector for identifying seizure-triggering videos. Full specifications are included in Appendix B.

6.1. Policy on Luminance Sensitivity

> ***Guideline:*** A general flash is defined as a pair of opposing changes in relative luminance of 10% or more of the maximum relative luminance (1.0) where the relative luminance of the darker image is below 0.80; and where "a pair of opposing changes" is an increase followed by a decrease, or a decrease followed by an increase.

Physiological Basis. We first calculate the relative luminance of each pixel. Assuming the input video files are in three sRGB (standard RGB) channels R_s, G_s, B_s for each frame f and pixel (x, y). The relative luminance I of a pixel is the weighted sum of the gamma-expanded channels. The coefficients reflect the human eye's sensitivity to different colors derived from the luminous efficiency function, with green light perceived as the brightest:

```
Definition I (f x y : nat) : R :=
  0.2126 * gammaExpand (R_s f x y) +
  0.7152 * gammaExpand (G_s f x y) +
  0.0722 * gammaExpand (B_s f x y).
```

The human eyes perceive light intensity in a nonlinear manner, being more sensitive to changes in darker areas than brighter ones [43]. To account for this, gamma expansion helps adjust pixel values so that the displayed content aligns with human vision. Gamma-expanded pixel values are derived from sRGB values C_s, where the piecewise gamma function (γ = 2.4 per industrial standard) is:

```
Definition gammaExpand (Cs : R) : R :=
  if Rle_dec Cs 0.04045 then
    Cs / 12.92
  else
    Rpower ((Cs + 0.055) / 1.055) 2.4.
```

For HDR display, instead of luminance changes, contrast should be used to determine harmful transitions [17]. This is due to the human visual system being particularly sensitive to changes in contrast at high luminance. In addition to the standard guideline, when the relative luminance of the darker image is above 0.8, a harmful pixel transition is one where the Michelson Contrast C_M [44] is $\frac{1}{17}$ or greater.

Formal Specifications. Assuming an HDR display, where f_1 and f_2 are consecutive frames, two consecutive harmful transitions of opposing changes constitute a flashing pixel:

```
Definition harmful_lum_transition (f1 f2 x y :
 ↪  nat) : Prop :=
  let i1 := I f1 x y in
  let i2 := I f2 x y in
  (i1 > 0.8 ∧ i2 > 0.8 ∧ C_M f1 f2 x y >= (1 /
   ↪  17)) ∧ (Rabs (i2 - i1) >= 0.1).
Definition opposing_lum_changes (f1 f2 f3 x y :
 ↪  nat) : Prop :=
  let i1 := I f1 x y in
  let i2 := I f2 x y in
  let i3 := I f3 x y in
  (i2 > i1 ∧ i3 < i2) ∧ (i2 < i1 ∧ i3 > i2).
Definition is_flash (f1 f2 f3 x y : nat) : Prop
 ↪  :=
  harmful_transition f1 f2 x y ∧
   ↪  harmful_transition f2 f3 x y ∧
   ↪  opposing_changes f1 f2 f3 x y.
```

6.2. Policy on Chromatic Sensitivity

> ***Guideline:*** A red flash is defined as any pair of opposing transitions involving a saturated red, where one transition is either to or from a state with a ratio of red that is

greater than or equal to 0.8, and the difference between states is more than 0.2 (unitless) in the CIE 1976 UCS chromaticity diagram.

Physiological Basis. We first calculate the red ratio of each pixel. Recalling the piecewise gamma expansion function, sRGB values are converted into linear RGB values $(R_{lin}, G_{lin}, B_{lin})$. Then the ratio of red are defined as:

```
Definition red_ratio (f x y : nat) : R :=
  let r := R_lin f x y in
  let g := G_lin f x y in
  let b := B_lin f x y in
  let sum := r + g + b in
  if Rle_dec sum 0 then 0 else r / sum.
```

To calculate the chromatic difference between transitioning pixels, pixels are first converted from linear RGB to the CIE XYZ colorspace, which models an idealization of the color vision of a normal human. The chromatic difference is the Euclidean distance between the chromaticity coordinates.

```
Definition denom (f x y : nat) : R :=
  X f x y + 15 * (Y f x y) + 3 * (Z f x y).
Definition u_prime (f x y : nat) : R :=
  let d := denom f x y in
  if Rle_dec d 0 then 0 else (4 * X f x y) / d.
Definition v_prime (f x y : nat) : R :=
  let d := denom f x y in
  if Rle_dec d 0 then 0 else (9 * Y f x y) / d.
Definition color_diff (f1 f2 x y : nat) : R :=
  let u1 := u_prime f1 x y in
  let v1 := v_prime f1 x y in
  let u2 := u_prime f2 x y in
  let v2 := v_prime f2 x y in
  sqrt ((u1 - u2) ^ 2 + (v1 - v2) ^ 2).
```

Formal Specifications. According to the guideline, where f_1 and f_2 are consecutive frames, two consecutive harmful transitions of opposing changes constitute a flashing pixel.

```
Definition harmful_color_transition (f1 f2 x y :
↪  nat) : Prop :=
  (red_ratio f1 x y >= 0.8 ∧ red_ratio f2 x y >=
  ↪  0.8) ∧ (color_diff f1 f2 x y > 0.2).
Definition opposing_color_changes (f1 f2 f3 x y :
↪  nat) : Prop :=
  let rr1 := red_ratio f1 x y in
  let rr2 := red_ratio f2 x y in
  let rr3 := red_ratio f3 x y in
  (rr2 > rr1 ∧ rr3 < rr2) ∧ (rr2 < rr1 ∧ rr3 >
  ↪  rr2).
Definition is_red_flash (f1 f2 f3 x y : nat) :
↪  Prop :=
  harmful_color_transition f1 f2 x y ∧
  ↪  harmful_color_transition f2 f3 x y ∧
  ↪  opposing_color_changes f1 f2 f3 x y.
```

6.3. Policy on Spatial Extent

Guideline: Combined area of flashes occurring concurrently occupies no more than a total of .006 steradians within any 10-degree visual field on the screen at typical viewing distance.

Physiological Basis. For web content, WCAG assumes a 15 to 17-inch screen of 1024 x 768 resolution as a reference [39]. Therefore, a 10-degree visual field at a typical viewing distance (22-26 inches) captures about 341 x 256 pixels. Since 25% of any 10-degree visual field on the screen approximates 0.006 steradians, a harmful frame transition is found if the total number of flashing pixels exceeds $341 * 256 * 0.25$. However, this made strong assumptions about the device resolution, screen size, and the viewing distance, which vary depending on the actual usage conditions. Therefore, we consider how these factors may affect the calculation of the flash area threshold.

In estimation, the central 10-degree of the human visual field is 10 degrees wide and 7.5 degrees high [18]. Suppose that the screen diagonal is S inches, the screen resolution is w by h pixels, and the screen is being viewed at a distance d. We calculate the device resolution in pixels per inch (PPI):

```
Definition theta_h_deg : R := 10.
Definition theta_v_deg : R := 7.5.
Definition deg_to_rad (x : R) : R := x * PI /
↪  180.
Definition flash_area_threshold (d w h : R) : R
↪  :=
  let PPI := sqrt (w^2 + h^2) / S in
  let theta_h := deg_to_rad theta_h_deg in
  let theta_v := deg_to_rad theta_v_deg in
  let area_inch := (d * theta_h) * (d * theta_v)
  ↪  in
  let area_px := area_inch * (PPI ^ 2) in
  0.25 * area_px.
```

Formal Specifications. According to the guideline, the total number of flashing pixels A between f_1 and f_2 should not exceed this threshold.

```
Definition no_harmful_flash :
  forall (f1 f2 : Frame) (d w h : R),
  0 <= d -> 0 < w -> 0 < h ->
  A f1 f2 <= flash_area_threshold d w h.
```

6.4. Policy on Temporal Frequency

Guideline: There are no more than three general flashes and/or no more than three red flashes within any one-second period.

Formal Specifications. Let $F_{gen}(t)$ and $F_{red}(t)$ be the number of general flashes and red flashes in the interval $[t, t+1)$. Let T be the total duration of the video in seconds. We say that for every time t in $[0, T-1)$, there are at most 3 general flashes AND at most 3 red flashes in the interval $[t, t+1)$.

```
Definition respects_flash_rate : Prop :=
 forall t : R,
   0 <= t <= T - 1 ->
   (F_gen t <= 3)%nat ∧ (F_red t <= 3)%nat.
```

Equivalently, a harmful video is one where there exists some time t in $[0, T-1)$ that has more than 4 general flashes or red flashes within that 1-second interval.

```
Definition harmful_video : Prop :=
  exists t : R,
    0 <= t <= T - 1 ∧
    (F_gen t >= 4)%nat ∧ (F_red t >= 4)%nat.
```

The algorithm is designed to analyze incoming frames in a sliding window fashion. Each incoming frame is processed and added to a queue, where one second's worth of frames is kept. We use the video's framerate to comply with the "one-second period" specification. For instance, for a video played at 30 FPS, the queue will be filled with 30 processed frames for seizure-trigger identification.

7. Evaluation

In this section, we formally verified PRISM's implementation to be strictly following the presented formal specification. With a self-generated edge-case benchmark dataset, we examined the correctness and effectiveness of PRISM and compared its performance with prior implementations. Finally, we evaluate the performance overhead of PRISM.

7.1. Formal Verification

Verification Implementation. PRISM was implemented using C, which has complex semantics and language features that are not supported by many verification tools. To address this challenge, we employ the Spoq verification framework [45]. We leverage the Clang compiler front end to parse C code into LLVM's language-independent intermediate representation (IR), then translate the resulting control flow graphs into Rocq representation using program-style functions with if-then-else and loop statements using Spoq. This makes verification more amenable and enables the verification for richer C semantics, including complex libraries such as OpenCV[4], compared to alternatives like CompCert's ClightGen [46].

Verification Approaches. To formally verify PRISM's functional correctness, we conducted verification on two main categories: *structural correctness* and *pattern correctness*. Then, we leverage contextual refinement to demonstrate that each component correctly implements its specification within the context of the larger system, ensuring that verified properties are preserved across component boundaries.

(1) Structural correctness ensures that our frame processing functions preserve data integrity throughout the analysis:

```
Theorem frame_processing_safety :
  forall rows : list (list Pixel),
  let (lum_harmful, col_harmful) :=
   ↪  process_frame_rows rows in
  length lum_harmful = length rows ∧
  forall i, i < length rows ->
    length (nth i lum_harmful []) = length (nth i
     ↪  rows []).
```

4. Verification of OpenCV's internal implementation is beyond the scope of this work; we focus our verification efforts on the system logic that utilizes these library functions.

The theorem shows that the frame processing functions preserve input data structures. This guarantees that PRISM operates correctly on videos of arbitrary size and content.

(2) Pattern correctness ensures that our detection algorithms accurately identify harmful flashing patterns according to established guidelines. We verify several key properties to ensure flash detection accuracy across three critical components: pattern identification, numerical stability, and color space transformations.

First, we demonstrate that our flash detection function correctly identifies opposing luminance changes. Verification proves that when `is_flash` returns true for consecutive frames f_1, f_2, f_3 at position (x, y), one of two conditions holds: either the luminance values l_1, l_2, l_3 satisfy both harmfulness conditions (the l_1 to l_2 and l_2 to l_3 transitions exceed thresholds) while exhibiting opposing changes (forming a peak or valley pattern), or the corresponding red ratio values exhibit similar harmful opposing patterns.

To ensure these pattern detections remain reliable, we prove mathematical bounds for our Michelson Contrast calculations, which underpin the luminance threshold checks. For luminance values $l_1, l_2 \in [0, 1]$, the contrast formula $C_M = \frac{|l_2 - l_1|}{l_1 + l_2}$ satisfies $0 \leq C_M \leq 1$. This bound verification prevents overflow conditions and ensures that contrast measurements remain within physically meaningful ranges. Thereby guaranteeing the stability of our threshold comparisons.

Similarly, for the color-based detection pathway, our color space conversion verification establishes that transformations preserve the essential properties needed for detection. When the color detection function identifies harmful patterns based on red ratios r_1, r_2, and chromaticity differences Δx, Δy, we prove there exists a positive threshold τ such that $(\Delta x)^2 + (\Delta y)^2 \geq \tau$. This confirms that detected color changes correspond to perceptually significant differences, ensuring that luminance-based and color-based detection maintain accuracy throughout the processing pipeline.

(3) Compositional Verification via Contextual Refinement. To prove that the composition of structural correctness and pattern correctness maintains overall system correctness, we employ contextual refinement [47]–[50]. Refinement techniques are not always modular in the structure of a program (*i.e.,* they may require whole-program reasoning). Contextual refinement is a compositional verification technique that proves component integration correctness by establishing that each component correctly implements its specification within the operational context of the larger system. This approach allows us to compositionally verify that *structural correctness* and *pattern correctness* work together without one undermining the guarantees provided by the other. We first establish refinement relations between abstract specifications and concrete implementations. The abstract specification defines the essential outputs for pixel processing as a tuple $(L, R, (x, y))$ representing luminance, red ratio, and chromaticity coordinates. The concrete implementation achieves this through the composition of gamma expansion, RGB-to-XYZ conversion, and chromaticity calculation. Our

TABLE 5: General Flash Detection Result on Edge-Case Dataset (Pass, Fail)

Frequency	3 Flashes								5 Flashes							
Area (pixels)	100x100				200x200				100x100				200x200			
ΔI	> 0.1		< 0.1		> 0.1		< 0.1		> 0.1		< 0.1		> 0.1		< 0.1	
I **of Darker Frame**	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8
Ground Truth	P	P	P	P	P	P	P	P	P	P	P	P	F	F	P	P
PEAT	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
IRIS	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
PRISM	P	P	P	P	P	P	P	P	P	P	P	P	F	F	P	P

TABLE 6: Red Flash Detection Result on Edge-Case Dataset (Pass, Fail)

Frequency	3 Red Flashes								5 Red Flashes							
Area (pixels)	100x100				200x200				100x100				200x200			
Distance	> 0.2		< 0.2		> 0.2		< 0.2		> 0.2		< 0.2		> 0.2		< 0.2	
Rr	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8	< 0.8	> 0.8
Ground Truth	P	P	P	P	P	P	P	P	P	P	P	P	P	F	P	P
PEAT	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	F
IRIS	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
PRISM	P	P	P	P	P	P	P	P	P	P	P	P	P	F	P	P

refinement theorem proves that $abstract\_pixel\_spec(p) = concrete\_pixel\_impl(p)$ for all valid pixels p, establishing that the mathematical specification and the actual implementation are equivalent. The refinement extends to frame-level processing, where we define processing contexts that capture the computational state at each verification step. A processing context inductively builds from individual pixel contexts, each containing a pixel and its derived luminance, red ratio, and chromaticity values. Frame contexts aggregate pixel contexts while maintaining structural invariants about dimensions and data integrity.

The key challenge in compositional verification is proving that pattern correctness properties remain valid when operating on data that has undergone structural transformations. Our composition theorem proves structural correctness provides the necessary invariants for pattern correctness. Specifically, the frame processing operations preserve pixel relationships and spatial locality that pattern detection algorithms depend upon. When *process_frame_rows* maintains dimensional integrity (length preservation and row-column correspondence), the downstream pattern detection functions receive data in the exact format their verification assumes. The refinement establishes that for any sequence of frames $F = [f_1, f_2, f_3]$, if structural processing produces $F' = [f_1', f_2', f_3']$ where each f_i' maintains the structural invariants of f_i, then pattern detection results on F' are equivalent to pattern detection on F. That is, if *structural correctness* preserves detection context and *pattern correctness* is correct on context-preserving data, then *structural correctness* $\circ$ *pattern correctness* is correct. The verification provides mathematical assurance of PRISM's safety preservation through composition.

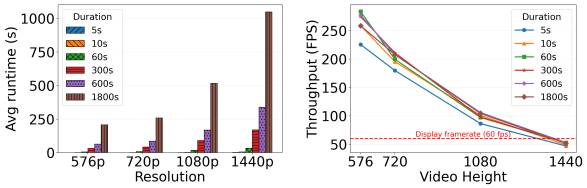
7.2. Accuracy on Edge-Case Dataset

An adaptive attacker can craft harmful videos that exhibit borderline trigger characteristics in an attempt to confuse the detector. In this section, we evaluate PRISM and prior implementations' sensitivity to the defined thresholds.

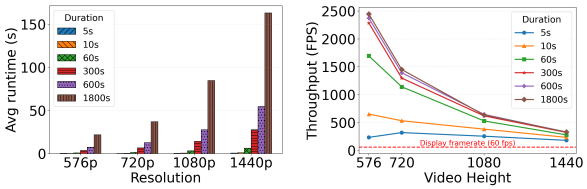
Method. We generated an edge-case dataset using the specifications and varying the luminance, color, area, and frequency of flashes above and below the thresholds specified in the guideline. We compared the results of PRISM with PEAT [36] and IRIS [51], which were the only freely available video detection software at the time of the experiment. PRISM and IRIS analyze videos in their original MP4 format. Since PEAT can only analyze files in AVI, videos were converted using FFmpeg. We manually checked the disagreement between the systems and discussed the results.

Result. Table 5 and Table 6 show the detection results. To establish a benchmark, we assumed the WCAG reference of 15 in. screen size with 1024 x 768 resolution at 23 in. viewing distance. PRISM detected all videos correctly and achieved an accuracy of 100% when content is displayed in SDR, since its implementation is derived directly from the formal specifications. It also correctly identifies dangerous cases when the content is assumed to be displayed in HDR. On the other hand, while PEAT can correctly identify most videos, false-positive and false-negative results exist. For instance, PEAT reported a video that contains both general flashes and red flashes as safe without identifying any flashes (false-negative). It does not consider the variations between SDR and HDR displays. It also flagged a video as containing red flash, even though the Euclidean distance between flashing pixels is less than 0.2 (false-positive). IRIS passes every video without identifying any as triggering.

Discrepancy between PRISM, PEAT, and IRIS. The discrepancy between the detection results may be partially due to PEAT and IRIS implementing an older version of WCAG, where the calculation of flash luminance and color threshold had been different: 1) the threshold for calculating gamma-expanded RGB values was 0.03928, and 2) instead of calculating the euclidean distance in the XYZ colorspace, a condition of $(R - G - B) * 320 > 20$ was used. While changes in the gamma-expansion threshold would not affect the result much, the change to use the XYZ colorspace may have impacted the detection results. We also found that the implementation of PEAT may have deviated even from the older version of WCAG. As a case study, we generated a



(a) Average runtime - CPU. (b) Throughput - CPU.



(c) Average runtime - GPU. (d) Throughput - GPU.

Figure 6: Impact of video duration and resolution on runtime performance.

video with 4 flashes in one second by inserting nonconsecutive white frames in a stream of black frames. By definition, the video is harmful since it satisfies all requirements of harmful flashes. However, the video was flagged as safe. We refer to §11 for the responsible disclosure process.

7.3. Runtime Performance

We evaluate PRISM on a commodity CPU and a single discrete GPU. We report end-to-end validation time per video and effective processing throughput in frames per second (FPS). The CPU baseline demonstrates feasibility on general-purpose hosts such as offline audit or regulatory review, while the GPU implementation targets high-throughput content moderation.

Design for Throughput. The GPU implementation is designed such that nearly all per-frame work remains on the device with minimal transfers. Each pixel is handled by a dedicated CUDA thread that performs luminance and chromaticity analysis with coalesced memory access, and only the moderation results are returned to the host. Host input buffers are page-locked, and decoding of frame $N+1$ on the CPU is overlapped with GPU processing of frame N using asynchronous direct memory access and persistent device allocations. This organization reduces host-to-device traffic, amortizes fixed costs, and yields stable scaling with video resolution and length.

Method. All experiments were conducted on an x86_64 server equipped with an Intel(R) Core(TM) i9-12900K CPU (24 cores), an NVIDIA GeForce RTX 3080, and running Ubuntu 24.04. CUDA 12.5 and OpenCV 4.6.0 were used for GPU acceleration and frame acquisition. We sampled videos of different durations (from 5 seconds to 3 minutes) and formatted them in different resolutions common on video platforms (576p, 720p, 1080p, 1440p with 9:16 aspect ratio). All of the videos have a framerate of 30 frames per second. To ensure consistency, each experiment was repeated five times, with the average runtime reported.

Results. Figure 6 shows that the GPU implementation delivers substantial reductions in end-to-end runtime for longer and higher-resolution content while preserving the exact detection logic of the CPU baseline. For the most common format on short-video platforms (5 to 10-second videos of 576p resolution), PRISM was able to bound the runtime to 1 second even without hardware acceleration. Meanwhile, optimization on the GPU leads to a 6-8x speed up across video samples (for instance, 32.7s on CPU versus 3.9s on GPU with a 576p, 5-minute clip). It is worth noting that throughput comfortably exceeds display frame rates on the GPU across all tested settings (for example, 54002 frames in 163.3 s at 1440×2560, which is ≈330 FPS). This shows that PRISM has the potential to perform content detection in real time when deployed on the user end, and the disruption to usability would be minimal, as processing the frames would not result in dropped frames. We will leave this exploration as future work.

8. Discussions

In this study, we showed that photosensitive threats can be observed in the video platform ecosystem. The prevalence of sensitive content calls for urgent measures and mitigation to help photosensitive users avoid such content. To this end, we discuss the following and suggest future work directions.

Guideline Adaptation and Enforcement. Video platforms should act as the first line of defense against seizure-triggering materials. Unlike those for violence or other explicit content, current moderation of seizure-triggering content is neither comprehensive nor enforced. Platforms should take the initiative to provide more stringent protection for users with PSE. Developing concrete legal enforcement for moderation can help facilitate the establishment and adaptation of relevant guidelines for video platforms.

Individual Differences Regarding Triggering Conditions. In addition to different viewing conditions, physiological differences among patients may be another key factor affecting the likelihood of videos triggering seizures. Though rare, a patient can find, for instance, a frequency of below 3 flashes triggering or uncomfortable. Other patients may feel that the guidelines are too stringent and wish for more flexibility. According to Binnie et al. [11], the percentage of patients who remain at risk with the guidelines in place would be 0.06% of the total patient population. While PRISM was able to protect the majority of the patients, which we consider important as we bring about the first step towards exploring the threats and defenses in video platforms, it should be noted that symptoms of PSE are particular to individuals, and the set of guidelines is yet to be comprehensive. Exploring other factors and features that are tailored to individual needs may be the next step to further increase platform accessibility.

Creator Labeling for Risk Disclosure. Video platforms can promote or incentivize creators to disclose hidden photosensitive risks. While malicious actors are likely to ignore such measures, this is helpful for video filtering and also raises

awareness among content creators, preventing the onset of accidental attacks. Together with platform-side detectors, users can be more informed, and their exposure to sensitive videos can be minimized.

8.1. Future Work

Platform-Side ML-Based Detector. The possibility of using ML-based classifiers as a defense for seizure-triggering content on video platforms has yet to be explored. Given that video platforms can readily access large amounts of video data, the dataset can be of great use in the training and testing of ML models. At the same time, policy-based detectors can help provide ground-truth labels for training data. ML-based classifiers can serve as agents of a second opinion in identifying seizure-triggering videos due to their capability to capture underlying patterns from large amounts of data. In addition, such classifiers may help identify new triggering characteristics previously unknown, facilitating revisions of outdated policies regarding PSE protections.

Epileptic Triggers from AI-Generated Videos. Advancements in generative AI have allowed the emergence of new service platforms that deploy artificial intelligence to automatically synthesize, transform, and edit videos using text prompts or other input videos [52]. Similar to how large language models have been exploited to create harmful texts [53], the technology has opened up new threat surfaces as models can accidentally or intentionally be used to generate provocative visualizations that are dangerous to photosensitive patients. Attacks and defenses related to seizure-triggering content produced by generative AI should be further explored, with facilitation and restrictions in place to prevent the dissemination of harmful sequences.

Epileptic Triggers in Emerging Devices. Besides video platforms, seizure-triggering content can also be found on emerging devices, such as in AR/VR [54]. The unique immersive viewing condition of headsets suggests moderation strategies that could be vastly different from traditional flat-screen displays, requiring consideration of the interaction between virtual objects, real-world scenes, and user posture.

Usable Trigger Warning Interface. Developing a usable warning interface to convey the result of detection is essential for users to be informed of the elements that could lead to triggering content. Video platforms can tag content with a risk score, along with a warning that describes high-risk viewing conditions. A usable interface can also allow patients to adjust detection thresholds to their needs, enabling tailored modifications to individual conditions.

9. Related Work

Visual Content Monitoring. As video platforms become more popular, not only can text-based content be intrusive [57]–[60], but video-based content can also harm users [2], [61]. Past work studying video platforms has discussed different forms of toxic content targeting various at-risk populations, including exploitative monetization [62], user radicalization [63], and abortion misinformation [64]. Meanwhile, content moderation strategies against harmful videos have been systematized and proposed [65]–[67], such as using ML to detect inappropriate content on video platforms for children [68], [69].

TABLE 7: Existing Tools for PSE Content Detection

Name	Technique	Moderation Scope	Cyber-Physical Components			
			Digital File	Display Hardware	Playback Setting	Viewing Condition
PRISM	Policy	Video platform	✓	✓	✓	✓
FPA [55]	Policy	TV broadcast	✓			
PEAT [36]	Policy	Offline video files	✓			
IRIS [51]	Policy	Offline video files	✓			
South et al. [37]	Policy	Browser, GIFs	✓			
Barbu et al. [56]	ML	Offline video files	✓			

Display Hardware: screen size, screen resolution. Playback Setting: screen brightness, SDR/HDR, playback speed, video resolution. Viewing Condition: display orientation, viewing distance

Detecting Epilepsy-Triggering Content. The establishment of guidelines has helped foster policy-based detection systems that help video creators identify triggering sequences before publishing their videos. For instance, the Harding Flash and Pattern Analyser (FPA) [55] was developed as a paid commercial software for TV broadcasting content. Meanwhile, Photosensitive Epilepsy Analysis Tool (PEAT) [36] provides a graphical user interface for creators and developers to visualize the occurrence of triggering sequences in their products. The video game company Electronic Arts has also developed IRIS [51] to allow developers to identify photosensitivity issues early in the development pipeline. Besides those built for creators, tools were also developed to empower photosensitive users with the ability to protect themselves. South et al. [37] presented PhotosensitivityPal, a browser extension capable of analyzing GIFs used in social media, showing GIFs to the users only when it is safe. Barbu et al. [56] proposed a prototype ML-based solution and used stacked LSTM U-Net to accomplish video-to-video transformation, identifying and removing seizure-inducing flashes from video inputs. Table 7 shows a systematic comparison. It should be noted that all prior work focused solely on digital media, but neglected to account for the variations introduced in video processing and playback.

10. Conclusion

We presented an empirical investigation of 15 video platforms, studying how these platforms moderate harmful content for users with PSE. We demonstrated several proof-of-concept attack videos that leverage different components in the content delivery pipeline to produce and propagate seizure-triggering videos, and found the current content moderation strategies lacking. Our findings motivated us to develop PRISM, a formally verified content moderation framework that securely and effectively identifies seizure-triggering content for at-risk users. We finally discuss directions for future research and accessibility improvements for photosensitive users on video platforms.

11. Ethics Considerations

Human Subjects and Platform Users. Our research does not involve any human subjects, as flashing stimuli may trigger seizures or other adverse neurophysiological responses. Instead, we validated PRISM against medically informed thresholds derived from WCAG and clinical literature. This approach ensures that our results remain ethically sound while consistent with established methodologies in accessibility and safety research.

We acknowledge that we did not coordinate with the video platforms for the proof-of-concept experiments, as the practical challenges in identifying and contacting the appropriate individuals within large organizations who are authorized to provide consent made it infeasible to obtain such consent. Additionally, accounts used for the investigations were set to private whenever possible and were not accessible by anyone outside of the research group. These accounts were not used to communicate with real users. All the videos uploaded are set as private/unlisted, so they cannot be viewed by other general users.

Statistics show that video platforms such as YouTube receive 3.7 million video uploads daily, amounting to more than 518,000 hours of new content every day [1]. This makes the impact of our video uploads to the platforms themselves negligible. In light of these statistics and our overall experimental setup, we conclude that the video platforms and their users are not negatively affected by the experiments. Instead, our findings can help motivate the development of related regulations and content moderation tools that strengthen video platforms against attacks that are already exploited by malicious actors in the wild. Hopefully, these measures can be widely adopted by video platforms to benefit users with PSE.

Vulnerability Disclosure. Upon discovery of the vulnerability, we followed a responsible disclosure protocol, where the vulnerabilities we identified were reported to the platforms through their respective issue reporting channels. At the time of submission, most platforms have acknowledged the issue. While some do not recognize it as having a security implication, others have put forward plans and updates on their community guideline to address the issue. We are actively communicating with the respective platforms to explore potential mitigation options. We have also contacted the developers of PEAT and IRIS about their respective performance on the edge-case dataset. We have reported the issues to MITRE, yet these were deemed as "detection results" and were not assigned CVEs, even though the software failed to identify a specific pattern that it explicitly claims to be able to detect [70].

References

[1] M. Adavelli, "How many videos are uploaded to youtube every day in 2025?," Jul 2023.

[2] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. G. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini, "Sok: Hate, harassment, and the changing landscape of online abuse," in *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 247–267, 2021.

[3] E. Mariconti, G. Suarez-Tangil, J. Blackburn, E. De Cristofaro, N. Kourtellis, I. Leontiadis, J. L. Serrano, and G. Stringhini, ""you know what to do": Proactive detection of youtube videos targeted by coordinated hate attacks," *Proc. ACM Hum.-Comput. Interact.*, no. CSCW, 2019.

[4] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, and M. Bailey, "Designing toxic content classification for a diversity of perspectives," in *Proceedings of the Seventeenth USENIX Conference on Usable Privacy and Security*, SOUPS'21, (USA), USENIX Association, 2021.

[5] K. Mahar, A. X. Zhang, and D. Karger, "Squadbox: A tool to combat email harassment using friendsourced moderation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, p. 1–13, ACM, 2018.

[6] E. Chandrasekharan, M. Samory, A. Srinivasan, and E. Gilbert, "The bag of communities: Identifying abusive behavior online with preexisting internet data," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, p. 3175–3187, ACM, 2017.

[7] R. S. Fisher, G. Harding, G. Erba, G. L. Barkley, and A. Wilkins, "Photic-and pattern-induced seizures: a review for the epilepsy foundation of america working group," *Epilepsia*, vol. 46, no. 9, 2005.

[8] B. J. Dlouhy, B. K. Gehlbach, and G. B. Richerson, "Sudden unexpected death in epilepsy: basic mechanisms and clinical implications for prevention," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 87, no. 4, pp. 402–413, 2016.

[9] S. Shorvon and T. Tomson, "Sudden unexpected death in epilepsy," *The Lancet*, vol. 378, no. 9808, pp. 2028–2038, 2011.

[10] R. S. Fisher, J. N. Acharya, F. M. Baumer, J. A. French, P. Parisi, J. H. Solodar, J. P. Szaflarski, L. L. Thio, B. Tolchin, A. J. Wilkins, *et al.*, "Visually sensitive seizures: An updated review by the epilepsy foundation," *Epilepsia*, vol. 63, no. 4, pp. 739–768, 2022.

[11] C. Binnie, J. Emmett, P. Gardiner, G. Harding, D. Harrison, and A. Wilkins, "Characterizing the flashing television images that precipitate seizures," *SMPTE journal*, vol. 111, no. 6-7, 2002.

[12] W. Yates, "Trolls try to trigger seizures - is it assault?," Jan 2017.

[13] "Eichenwald v. rivello," *United States District Court for The District of Maryland*, May 2018.

[14] K. Poulsen, "Hackers assault epilepsy patients via computer." https://www.wired.com/2008/03/hackers-assault-epilepsy-patients-via-computer/, Mar 2008.

[15] B. Ertl, "Hooligans attack epilepsy patients during epilepsy awareness month." https://www.pr.com/press-release/60959, 2007.

[16] M. uNiTeD, "Banned pokemon seizure scene." https://www.youtube.com/watch?v=4LQLvgXLguk, Aug 2015.

[17] ITU-R, "Guidance for the reduction of photosensitive epileptic seizures caused by television," Oct 2019.

[18] C. Adams, A. Campbell, R. Montgomery, M. Cooper, and A. Kirkpatrick., "Web content accessibility guidelines 2.2," 2023.

[19] W. H. Organization, "Epilepsy," Feb 2022.

[20] A. M. Da Silva and B. Leal, "Photosensitivity and epilepsy: current concepts and perspectives—a narrative review," *Seizure*, 2017.

[21] J. B. Jordan and G. C. Vanderheiden, "International guidelines for photosensitive epilepsy: Gap analysis and recommendations," *ACM Transactions on Accessible Computing*, vol. 17, no. 3, pp. 1–35, 2024.

[22] T. Publishing, "New study reveals surprising content viewing habits for 2021, forecasting how social media will evolve this year," *PR Newswire*, Jan 2022.

[23] D. Winter, "15 essential tiktok statistics for marketers in 2025," Jul 2024.

[24] G. Conti, M. Ahamad, and J. Stasko, "Attacking information visualization system usability overloading and deceiving the human," in *Proceedings of the 2005 Symposium on Usable Privacy and Security*, SOUPS '05, p. 89–100, ACM, 2005.

[25] A. M. McNutt, L. Huang, and K. Koenig, "Visualization for villainy," 2021.

[26] A. Wilkins, J. Emmett, and G. Harding, "Characterizing the patterned images that precipitate seizures and optimizing guidelines to prevent them," *Epilepsia*, vol. 46, no. 8, pp. 1212–1218, 2005.

[27] G. Harding and P. Harding, "Photosensitive epilepsy and image safety," *Applied ergonomics*, vol. 41, no. 4, pp. 504–508, 2010.

[28] M. Funatsuka, M. Fujita, S. Shirakawa, H. Oguni, and M. Osawa, "Study on photo-pattern sensitivity in patients with electronic screen game–induced seizures (esgs): Effects of spatial resolution, brightness, and pattern movement," *Epilepsia*, vol. 42, no. 9, 2001.

[29] Ofcom, "Guidance notes - section 2: Harm and offense," Jul 2017.

[30] Consumer and G. A. Bureau, "The FCC and speech," *Federal Communications Commission*, Aug 2022.

[31] B. News, "Japanese cartoon triggers fits in children," Dec 1997.

[32] G. News and Media, "Citroën ad banned after causing epileptic seizure," Jan 2012.

[33] C. News, "London olympic animation withdrawn over epilepsy fears," Jun 2007.

[34] hindleyite, "Banned pot noodle advert - ace of spades." https://www.youtube.com/watch?v=Bs0WCUZqoJg, Mar 2008.

[35] M. Gajanan, "Hackers posted seizure-inducing images on epilepsy foundation," Dec 2019.

[36] T. R. . D. Center, "Photosensitive epilepsy analysis tool (PEAT)," Feb 2021.

[37] L. South, D. Saffo, and M. A. Borkin, "Detecting and defending against seizure-inducing gifs in social media," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–17, 2021.

[38] Cloudfare, "What is adaptive bitrate streaming?," 2025.

[39] "G176: Keeping the flashing area small enough," *WCAG2.2 Techniques*, 2024.

[40] N. A. R. Center, "Luminance contrast in color graphics," *NASA*, 2024.

[41] statcounter, "Mobile screen resolution stats united states of america," Jul 2023.

[42] "Settlement agreement under the americans with disabilities act between the united states of america," *Ada.gov*, Dec 2021.

[43] "Understanding gamma correction," *Cambridge in Color*.

[44] D. G. Pelli and P. Bex, "Measuring contrast sensitivity," *Vision research*, vol. 90, pp. 10–14, 2013.

[45] X. Li, X. Li, W. Qiang, R. Gu, and J. Nieh, "Spoq: Scaling {Machine-Checkable} systems verification in coq," in *17th USENIX Symposium on Operating Systems Design and Implementation (OSDI 23)*, pp. 851–869, 2023.

[46] X. Leroy, *The CompCert C verified compiler: Documentation and user's manual*. PhD thesis, Inria, 2024.

[47] Y. Song, M. Cho, D. Lee, C.-K. Hur, M. Sammler, and D. Dreyer, "Conditional contextual refinement," *Proceedings of the ACM on Programming Languages*, vol. 7, no. POPL, pp. 1121–1151, 2023.

[48] A. Turon, D. Dreyer, and L. Birkedal, "Unifying refinement and hoare-style reasoning in a logic for higher-order concurrency," in *Proceedings of the 18th ACM SIGPLAN international conference on Functional programming*, pp. 377–390, 2013.

[49] D. Frumin, R. Krebbers, and L. Birkedal, "Reloc: A mechanised relational logic for fine-grained concurrency," in *Proceedings of the 33rd Annual ACM/IEEE Symposium on Logic in Computer Science*, pp. 442–451, 2018.

[50] L. Gäher, M. Sammler, S. Spies, R. Jung, H.-H. Dang, R. Krebbers, J. Kang, and D. Dreyer, "Simuliris: a separation logic framework for verifying concurrent program optimizations," *Proceedings of the ACM on Programming Languages*, vol. 6, no. POPL, pp. 1–31, 2022.

[51] electronicarts, "Iris." https://github.com/electronicarts/IRIS.

[52] P. Esser, J. Chiu, P. Atighehchian, J. Granskog, and A. Germanidis, "Structure and content-guided video synthesis with diffusion models," *arXiv preprint arXiv:2302.03011*, 2023.

[53] A. Wei, N. Haghtalab, and J. Steinhardt, "Jailbroken: How does llm safety training fail?," *arXiv preprint arXiv:2307.02483*, 2023.

[54] S. Baldassi, T. Kohno, F. Roesner, and M. Tian, "Challenges and new directions in augmented reality, computer security, and neuroscience – part 1: Risks to sensation and perception," 2018.

[55] C. R. S. Ltd, "Harding Flash and Pattern Analyzer," null.

[56] A. Barbu, D. Banda, and B. Katz, "Deep video-to-video transformations for accessibility with an application to photosensitivity," *Pattern Recognition Letters*, vol. 137, pp. 99–107, 2020.

[57] S. Zannettou, J. Finkelstein, B. Bradlyn, and J. Blackburn, "A quantitative approach to understanding online antisemitism," in *Proceedings of the International AAAI conference on Web and Social Media*, vol. 14, pp. 786–797, 2020.

[58] A. Arunasalam, H. Farrukh, E. Tekcan, and Z. B. Celik, "Understanding the security and privacy implications of online toxic content on refugees," in *USENIX Security Symposium*, 2024.

[59] S. Kumar, J. Cheng, and J. Leskovec, "Antisocial behavior on the web: Characterization and detection," in *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, (Republic and Canton of Geneva, CHE), International World Wide Web Conferences Steering Committee, 2017.

[60] N. Sambasivan, A. Batool, N. Ahmed, T. Matthews, K. Thomas, L. S. Gaytán-Lugo, D. Nemer, E. Bursztein, E. Churchill, and S. Consolvo, ""they don't leave us alone anywhere we go": Gender and digital abuse in south asia," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, p. 1–14, ACM, 2019.

[61] S. Niu, Z. Lu, A. X. Zhang, J. Cai, C. F. Griggio, and H. Heuer, "Building credibility, trust, and safety on video-sharing platforms," in *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, ACM, 2023.

[62] A. Chu, A. Arunasalam, M. O. Ozmen, and Z. B. Celik, "Behind the tube: Exploitative monetization of content on YouTube," in *31st USENIX Security Symposium (USENIX Security 22)*, 2022.

[63] M. H. Ribeiro, R. Ottoni, R. West, V. A. Almeida, and W. Meira Jr, "Auditing radicalization pathways on youtube," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 131–141, 2020.

[64] F. Sharevski, J. Vander Loop, P. Jachim, A. Devine, and E. Pieroni, "Talking abortion (mis) information with chatgpt on tiktok," in *European Symposium on Security and Privacy Workshops*, IEEE, 2023.

[65] M. Singhal, C. Ling, P. Paudel, P. Thota, N. Kumarswamy, G. Stringhini, and S. Nilizadeh, "Sok: Content moderation in social media, from guidelines to enforcement, and research to practice," in *2023 IEEE 8th European Symposium on Security and Privacy*, IEEE, 2023.

[66] V. Schmitt, V. Solopova, V. Woloszyn, and J. d. J. de Pinho Pinhal, "Implications of the new regulation proposed by the european commission on automatic content moderation," in *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, 2021.

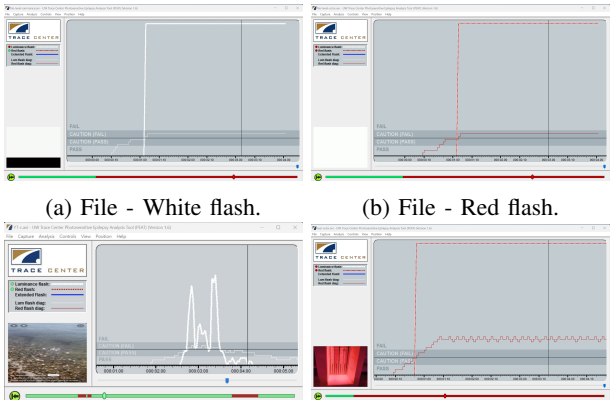
[67] N. Aggarwal, S. Agrawal, and A. Sureka, "Mining youtube metadata for detecting privacy invading harassment and misdemeanor videos," in *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, pp. 84–93, IEEE, 2014.

[68] K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, and M. Sirivianos, "Disturbed youtube for kids: Characterizing and detecting inappropriate videos targeting young children," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, pp. 522–533, 2020.

[69] R. Tahir, F. Ahmed, H. Saeed, S. Ali, F. Zaffar, and C. Wilson, "Bringing the kid back into youtube kids: Detecting inappropriate content on video streaming platforms," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 464–469, 2019.

[70] T. M. Corporation, "Cv numbering authority (cna) operational rules," *CVE.org*, 2024.

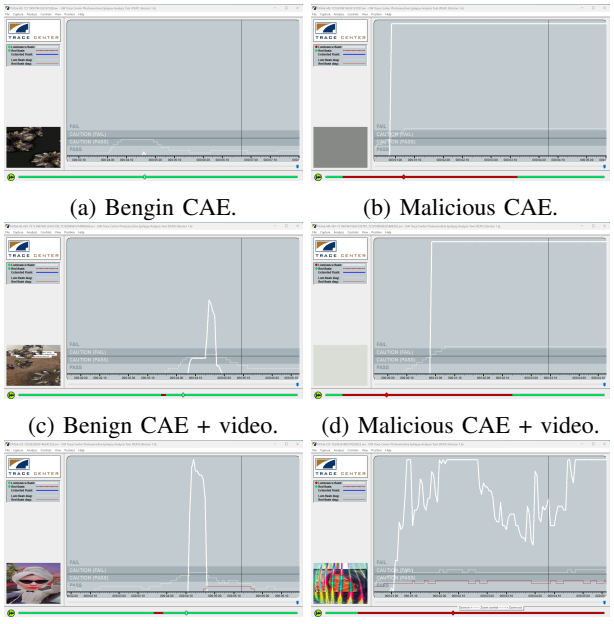
Appendix A.
Validation of PoC Samples



(a) File - White flash. (b) File - Red flash.

(c) Real-world - White flash. (d) Real-world - Red flash.

Figure 7: Crafting seizure-triggering videos by directly manipulating file content and filming real-world flashes.

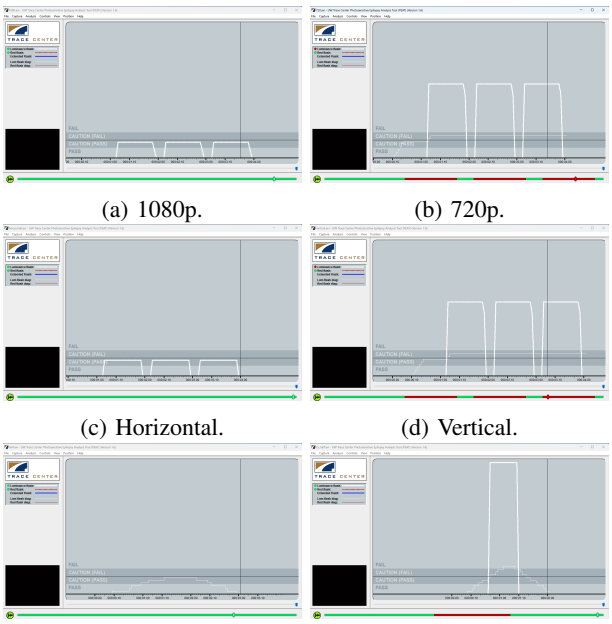


(a) Bengin CAE. (b) Malicious CAE.

(c) Benign CAE + video. (d) Malicious CAE + video.

(e) CDE #1. (f) CDE #2.

Figure 8: Crafting seizure-triggering videos by using effects alone or leveraging effect-video combinations. (CAE: Content-agnostic effects; CDE: Content-dependent effect)



(a) 1080p. (b) 720p.

(c) Horizontal. (d) Vertical.

(e) Normal speed. (f) 2x speed.

Figure 9: Variations in video playback *resolution, orientation*, and *speed* cause videos to become seizure-triggering.

Figure 10: Video designed to be seizure-triggering when played in HDR but not in SDR. PEAT analysis shows it exceeded the flash threshold but did not raise a warning.

Appendix B.
Rocq Formal Specifications

B.1. Flash Luminance Threshold

```
Section FlashLuminanceGuideline.

(* Gamma Expansion from sRGB to Linear *)
(* We assume sRGB values are in 0,1. The
↪  piecewise gamma function is per the sRGB
↪  standard. *)
Definition gammaExpand (Cs : R) : R :=
  if Rle_dec Cs 0.04045 then
    Cs / 12.92
```

```
  else
    Rpower ((Cs + 0.055) / 1.055) 2.4.

(* Relative Luminance I(f,x,y) *)
(* We assume your program provides three sRGB
↪  channels R_s, G_s, B_s for each frame f and
↪  pixel (x,y). *)
Parameter R_s : nat -> nat -> nat -> R.
Parameter G_s : nat -> nat -> nat -> R.
Parameter B_s : nat -> nat -> nat -> R.

(* The relative luminance is computed as a
↪  weighted sum of the gamma-expanded channels.
↪  *)
Definition I (f x y : nat) : R :=
  0.2126 * gammaExpand (R_s f x y) +
  0.7152 * gammaExpand (G_s f x y) +
  0.0722 * gammaExpand (B_s f x y).

(* Harmful Pixel Transition *)
(* Michelson Contrast *)
C_M = |I2 - I1| / (I1 + I2), for I1+I2 > 0.
Definition michelson_contrast (f1 f2 x y : nat) :
↪  R :=
  let i1 := I f1 x y in
  let i2 := I f2 x y in
  Rabs (i2 - i1) / (i1 + i2).

(* Harmful transition *)
(* If I(f1,x,y) and I(f2,x,y) both exceed 0.8,
↪  then we require Michelson contrast >= 1/17;
↪  otherwise, we require |delta I| >= 0.1. *)
Definition harmful_transition (f1 f2 x y : nat) :
↪  Prop :=
  let i1 := I f1 x y in
  let i2 := I f2 x y in
  (i1 > 0.8 /\ i2 > 0.8 /\ michelson_contrast f1
  ↪  f2 x y >= (1 / 17)) \/
  (Rabs (i2 - i1) >= 0.1).

(* Opposing Changes *)
(* A pair of opposing changes is an increase
↪  followed by a decrease, or a decrease
↪  followed by an increase. *)
Definition opposing_changes (f1 f2 f3 x y : nat)
↪  : Prop :=
  let i1 := I f1 x y in
  let i2 := I f2 x y in
  let i3 := I f3 x y in
  (i2 > i1 /\ i3 < i2) \/ (i2 < i1 /\ i3 > i2).

(* Two consecutive harmful pixel transitions of
↪  opposing changes constitute a flash. *)
Definition is_flash (f1 f2 f3 x y : nat) : Prop
↪  :=
  harmful_transition f1 f2 x y /\
  harmful_transition f2 f3 x y /\
  opposing_changes f1 f2 f3 x y.

End FlashLuminanceGuideline.
```

B.2. Flash Color Threshold

```
Section RedFlashGuideline.

(* Gamma expansion and sRGB -> XYZ *)
(* For simplicity, we reuse the piecewise gamma
↪  expansion from sRGB. *)
Definition gammaExpand (C : R) : R :=
  if Rle_dec C 0.04045 then
    C / 12.92
  else
    Rpower ((C + 0.055) / 1.055) 2.4.

(* We assume sRGB values in [0,1]. *)
Parameter R_s : nat -> nat -> nat -> R.
Parameter G_s : nat -> nat -> nat -> R.
Parameter B_s : nat -> nat -> nat -> R.

(* Convert sRGB to linear RGB via gamma
↪  expansion. *)
Definition R_lin (f x y : nat) : R := gammaExpand
↪  (R_s f x y).
Definition G_lin (f x y : nat) : R := gammaExpand
↪  (G_s f x y).
Definition B_lin (f x y : nat) : R := gammaExpand
↪  (B_s f x y).

(* Convert linear RGB to XYZ (D65 reference
↪  white). This is a common matrix for sRGB ->
↪  XYZ. *)
Definition X (f x y : nat) : R :=
  (0.4124 * R_lin f x y)
  + (0.3576 * G_lin f x y)
  + (0.1805 * B_lin f x y).

Definition Y (f x y : nat) : R :=
  (0.2126 * R_lin f x y)
  + (0.7152 * G_lin f x y)
  + (0.0722 * B_lin f x y).

Definition Z (f x y : nat) : R :=
  (0.0193 * R_lin f x y)
  + (0.1192 * G_lin f x y)
  + (0.9505 * B_lin f x y).

(* CIE 1976 UCS Chromaticity Coordinates (u', v')
↪  *)
Definition denom (f x y : nat) : R :=
  X f x y + 15 * (Y f x y) + 3 * (Z f x y).

Definition u_prime (f x y : nat) : R :=
  let d := denom f x y in
  if Rle_dec d 0 then 0 else (4 * X f x y) / d.

Definition v_prime (f x y : nat) : R :=
  let d := denom f x y in
  if Rle_dec d 0 then 0 else (9 * Y f x y) / d.

(* Euclidean Difference in (u', v') *)
Definition color_diff (f1 f2 x y : nat) : R :=
  let u1 := u_prime f1 x y in
  let v1 := v_prime f1 x y in
  let u2 := u_prime f2 x y in
  let v2 := v_prime f2 x y in
  sqrt ((u1 - u2) ^ 2 + (v1 - v2) ^ 2).

(* Red Ratio *)
(* The ratio of red in the linear RGB domain. If
↪  R+G+B = 0, define ratio = 0 to avoid division
↪  by zero. *)
Definition red_ratio (f x y : nat) : R :=
  let r := R_lin f x y in
  let g := G_lin f x y in
  let b := B_lin f x y in
  let sum := r + g + b in
  if Rle_dec sum 0 then 0 else r / sum.

(* Harmful Red Transition *)
Definition harmful_red_transition (f1 f2 x y :
↪  nat) : Prop :=
```

```
  (red_ratio f1 x y >= 0.8 \/ red_ratio f2 x y >=
  ↪  0.8)
  /\ color_diff f1 f2 x y > 0.2.

(* "Opposing changes" means: an increase followed
↪  by a decrease, or a decrease followed by an
↪  increase, in the red ratio. *)
Definition opposing_changes (f1 f2 f3 x y : nat)
↪  : Prop :=
  let rr1 := red_ratio f1 x y in
  let rr2 := red_ratio f2 x y in
  let rr3 := red_ratio f3 x y in
  (rr2 > rr1 /\ rr3 < rr2) \/ (rr2 < rr1 /\ rr3 >
  ↪  rr2).

(* A red flash is defined as any pair of opposing
↪  transitions involving a saturated red (>=0.8)
↪  with a color difference > 0.2. *)
(* Concretely, we say: Two consecutive harmful
↪  red transitions (f1->f2 and f2->f3) of
↪  opposing changes in the red ratio. *)
Definition is_red_flash (f1 f2 f3 x y : nat) :
↪  Prop :=
  harmful_red_transition f1 f2 x y
  /\ harmful_red_transition f2 f3 x y
  /\ opposing_changes f1 f2 f3 x y.

End RedFlashGuideline.
```

B.3. Flash Area Threshold

```
Section FlashAreaThreshold.

(* Screen Geometry Parameters *)
Parameter Frame : Set.
Parameter A : Frame -> Frame -> R.
Parameter S : R.
Axiom S_pos : 0 < S.

Definition theta_h_deg : R := 10.
Definition theta_v_deg : R := 7.5.
Definition deg_to_rad (x : R) : R := x * PI /
↪  180.

Definition flash_area_threshold (d w h : R) : R
↪  :=
  let ppi := sqrt (w^2 + h^2) / S in
  let theta_h := deg_to_rad theta_h_deg in
  let theta_v := deg_to_rad theta_v_deg in
  let area_inch := (d * theta_h) * (d * theta_v)
  ↪  in
  let area_px := area_inch * (ppi ^ 2) in
  0.25 * area_px.

(* Specifications for Flash Area *)
Definition no_harmful_flash :
  forall (f1 f2 : Frame) (d w h : R),
  0 <= d -> 0 < w -> 0 < h ->
  A f1 f2 <= flash_area_threshold d w h.

End FlashAreaThreshold.
```

B.4. Flash Frequency Threshold

```
Section FlashRateGuideline.

(* Parameters: Video duration and flash-counting
↪  functions *)
(* T is the total video duration in seconds. *)
Parameter T : R.
(* F_gen t = number of general flashes in the
↪  interval [t, t+1). *)
Parameter F_gen : R -> nat.
(* F_red t = number of red flashes in the
↪  interval [t, t+1). *)
Parameter F_red : R -> nat.

(* Specification *)
(* We say the video respects the flash rate if
↪  for every time t in [0, T-1], there are at
↪  most 3 general flashes AND at most 3 red
↪  flashes in the interval [t, t+1). *)
Definition respects_flash_rate : Prop :=
  forall t : R,
    0 <= t <= T - 1 ->
    (F_gen t <= 3)%nat /\ (F_red t <= 3)%nat.

(* Equivalently, a harmful video is one where
↪  there exists some time t in [0, T-1] that has
↪  4 or more (>=4) general flashes or 4 or more
↪  red flashes within that 1-second interval. *)
Definition harmful_video : Prop :=
  exists t : R,
    0 <= t <= T - 1 /\
    ((F_gen t >= 4)%nat \/ (F_red t >= 4)%nat).

End FlashRateGuideline.
```