Exploring Intervention Techniques to Alleviate Negative Emotions during Video Content Moderation Tasks as a Worker-centered Task Design

Dokyun Lee
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
dokyun@dgist.ac.kr

Sangeun Seo
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
seseo98@dgist.ac.kr

Chanwoo Park
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
bcw6622@dgist.ac.kr

Sunjun Kim
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
sunjun_kim@dgist.ac.kr

Buru Chang
Sogang University
Seoul, Republic of Korea
buru@sogang.ac.kr

Jean Y. Song
Electrical Engineering and Computer Science, DGIST
Daegu, Republic of Korea
jeansong@dgist.ac.kr

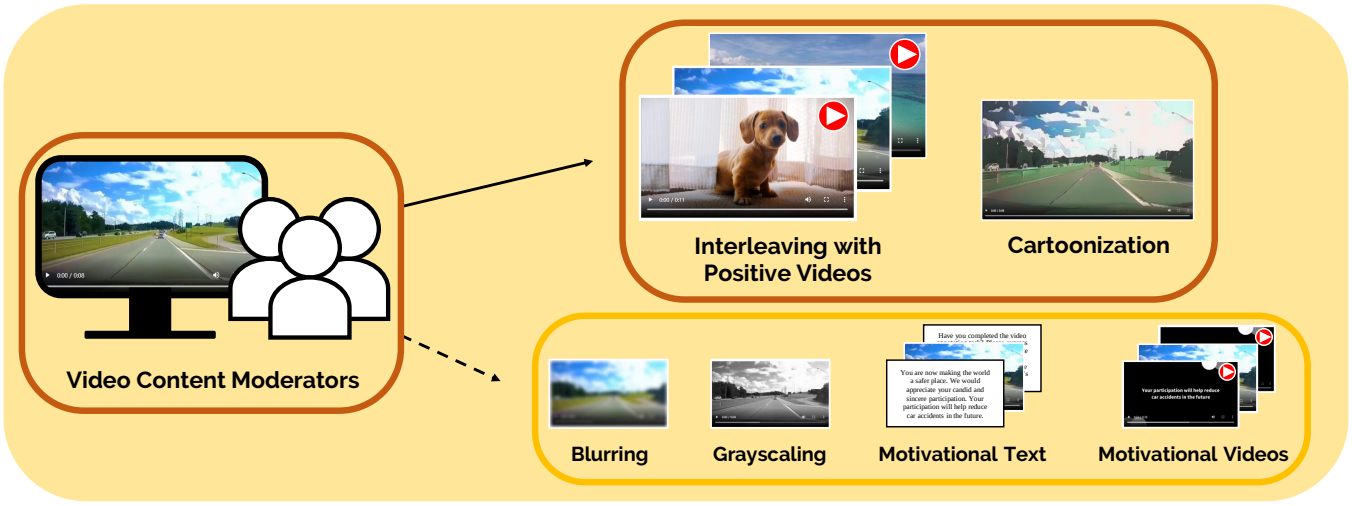

Figure 1: We examined six intervention techniques for video content moderation tasks that can be inserted and operated during the moderation tasks to alleviate negative emotions that video content moderators could experience during work. We propose that interleaving with positive videos and cartoonization are effective intervention techniques that significantly reduce negative emotions.

ABSTRACT

Videos are dynamic and multi-modal compared to other types of content, making automatic filtering difficult, which is why content moderators play a crucial role. However, video content moderators are exposed to more profound emotional labor because videos contain rich visual information, sometimes including even harmful content, such as violent or terrifying scenes. In this work, we explore the effect of six intervention techniques on alleviating negative emotions during video content moderation tasks. We conducted one online crowdsourcing experiment and two controlled user studies to find out that (i) interleaving with positive videos or (ii) cartoonization could significantly reduce negative emotions in the moderators. Participants reported that the advantages of these approaches are in helping reduce negative emotions at the time of moderation while existing approaches focus on post-task activities (e.g., relaxation, talking with others, or getting a hobby). We discuss the applicability of our findings to broader tasks, including improvement in intervention techniques.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

DIS '24, July 01–05, 2024, IT University of Copenhagen, Denmark

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0583-0/24/07
https://doi.org/10.1145/3643834.3660708

CCS CONCEPTS

• **Human-centered computing** ! **User centered design**.

KEYWORDS

video content moderation; mental health; social media; crowdsourcing; intervention techniques

ACM Reference Format:

Dokyun Lee, Sangeun Seo, Chanwoo Park, Sunjun Kim, Buru Chang, and Jean Y. Song. 2024. Exploring Intervention Techniques to Alleviate Negative Emotions during Video Content Moderation Tasks as a Worker-centered Task Design. In *Designing Interactive Systems Conference (DIS '24), July 01–05, 2024, IT University of Copenhagen, Denmark.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3643834.3660708

1 INTRODUCTION

A vast amount of user-generated video content is being poured onto the internet every hour [1, 80], with the average views per video easily surpassing a thousand times the number of videos being posted [1, 55]. Even though most content on the internet poses no significant issues, often there are posts containing violent scenes including horrific accidents, suicide, animal torture, decapitation, and pornography, which need to be filtered [14, 26, 53, 74]. That is, as the generating of various video content continues to increase, the importance of video content moderation tasks is being more emphasized than ever. However, moderating large amounts of video content through manual human review has its limitations. Fortunately, various machine learning-based prediction systems have been developed, assisting in automatically identifying problematic content [35, 58]. Nevertheless, filtering content using machine learning-based techniques is still difficult to fully automate due to high accuracy requirements, subjectivity, and safety concerns. Consequently, many people have to be involved in manual content moderation tasks, including in-house reviewers, third-party contract workers, online labor outsourcing employees, and even voluntary moderators [18, 29, 60, 61].

In general, videos are more memorable than photos or text because they engage multiple sensory areas [32] and are more effective at eliciting emotions than images [72]. Especially nowadays, since the demand for video content is more significant than compared to before, moderating video content is technically more intricate and time-consuming than moderating static images or text [30, 82]. Video content is not only dynamic and rich, but also multimodal, containing both visual and auditory elements [59, 71], requiring content moderators to thoroughly review each one by pausing, rewinding, and zooming to identify harmful scenes [25]. As a result, video content moderators may be exposed to harmful content for much longer periods. This adversely affects video content moderators' psychological well-being [16, 51, 54], potentially leading to serious conditions like Post-Traumatic Stress Disorder (PTSD) [16, 53, 54].

To overcome this, big tech companies such as Microsoft and Facebook provide video moderation tools that allow users to control blurring, black-and-white, or audio muting [69]. Unfortunately, there is little research on technical approaches to protect the emotions of content moderators beyond these intervention techniques, especially for video content moderation tasks. We believe that simple intervention techniques inserted within the moderation task can be an economical and practical solution for alleviating harmful emotional effects on video content moderators. Therefore, in this study, we introduce and investigate the effects of different intervention techniques that can be applied to any video. More specifically, we address the following research questions:

> RQ1. Which intervention techniques can reduce negative emotions for video content moderation tasks?
> RQ2. How do these intervention techniques reduce negative emotions?
> RQ3. Can these intervention techniques have lasting benefits for daily life and emotions?

We carefully chose six intervention techniques including four novel techniques by drawing from prior research and the insights we gained through iterative pilot studies on different intervention ideas. The selected techniques are blurring, grayscaling, inserting motivational text, inserting motivational videos, interleaving with positive videos, and cartoonization. We examined the changes in the participants' emotional state before and after a set of video moderation tasks using the Positive and Negative Affect Schedule (PANAS) [78]. We selected car crash videos as the main video content to use in the experiment. The reason for this was to provide stimulating videos to measure differences in participants' emotional changes based on each intervention technique. However, we excluded excessively stimulating videos that could cause mental harm to participants due to ethical concerns.

To answer RQ1 and RQ2, we conducted two independent experiments. The first experiment (n=280) recruited Amazon Mechanical Turk (AMT) workers to quantitatively measure whether six intervention techniques could emotionally benefit them. The second experiment (n=7) involved a controlled study in which qualitative data were collected through face-to-face interviews after participants experienced six intervention techniques. As a result, the first experiment showed that interleaving with positive videos and cartoonization significantly reduced negative emotions. In the second experiment, we report the characteristics underlying how each intervention technique reduces negative emotions (Table 1). Specifically, participants showed a preference for (i) interleaving with positive videos and (ii) cartoonization among the six intervention techniques. Additionally, we report various key findings and observations (Table 1), which could be utilized to improve the intervention techniques. These include preferences for the intervention techniques experienced by participants and the characteristics of each intervention technique.

Based on the results of the two experiments, we conducted the third experiment to further understand whether the positive effect of the two intervention techniques (interleaving with positive videos and applying cartoonization to the task videos) can last in daily life and emotions. This additional controlled study (n=9) was conducted for three days to answer RQ3. Participants performed daily tasks and completed self-report questionnaires for three days in a row. Then an in-depth interview was conducted after the last day's task. Later, we asked for a follow-up survey three weeks after conducting the tasks. We found out that after the three weeks, all participants had no significant changes in their daily lives and

Significant Increase in Negative Emotions for Intervention Techniques after Watching Car Crash Videos (Section 4.5)	
Significant Increase	Blurring, Grayscaling, Motivational Text, Motivational Videos.
No Significant Increase	Interleaving with Positive Videos, Cartoonization.
Preferences for Intervention Techniques (Section 5.3)	
Ranking Preference	(1) Interleaving with Positive Videos - (2) Cartoonization - (3) Grayscaling - (4) Blurring - (5) Motivational Videos - (6) Motivational Text.
Keywording the Characteristics of Each Intervention Technique (Section 5.3)	
Blurring	Low stimulation, Emotional distress reduction, Reduced task accuracy, Eye fatigue
Grayscaling	Eye fatigue improvement, Stress reduction, Low stimulation, Task interruption
Motivational Text	Motivation, Professionalism, Refresh, Annoyance, Repetitive fatigue
Motivational Videos	Sense of duty, Professionalism, Refresh, Annoyance, Repetitive fatigue
Interleaving with Positive Videos	Relaxation, Refresh, Low stress, Decreased work speed, Video preference discrepancy
Cartoonization	Low stimulation, Reduced afterimages, Stress reduction, Reduced task accuracy
The Impact of Interleaving with Positive Videos and Cartoonization in Three-Day Controlled Study (Section 6.3)	
Summary	We found out the potential of interleaving with positive videos and cartoonization to emotionally assist participants in sustaining their engagement with tasks.
Interleaving with Positive Videos	Interleaving with positive videos helps offset mood and distract concentration by allowing people to experience emotional relaxation through positive videos.
Cartoonization	Cartoonization reduces visual stimuli and fosters a sense of detachment from reality by feeling similar to cartoons.

Table 1: We present a summary of findings of the six intervention techniques based on quantitative and qualitative data analyses from three different types of experiments.

emotions. Also, participants reported the potential of interleaving with positive videos and cartoonization in emotionally assisting them and sustaining their task engagement.

Through a comprehensive analysis of the three experiments, the main contributions of our paper are the following:

- We propose four new intervention techniques in addition to the previously presented blurring and grayscaling. Also, we provide evidence on whether the six intervention techniques can be effectively applied to video content while reducing negative emotions, and the specific effects of each intervention technique.
- We present both quantitative and qualitative analyses showing that interleaving with positive videos or applying cartoonization on task videos significantly reduces negative emotions during video moderation tasks.
- We summarize the key findings obtained from the experiments (Table 1) and offer practical guidance for enhancing future intervention techniques.

2 RELATED WORK AND BACKGROUND

In this section, we review prior research in three main areas: (1) the emotional toll on workers from content moderation tasks, (2) strategies to alleviate emotional burden in manual content moderation, and (3) approaches to reduce general stress in other workspaces.

2.1 Emotional Toll: The Adverse Effects of Online Content Moderation

Online social media platforms are experiencing an exponential increase in the creation of user-generated content, which may potentially include harmful content [1, 5, 80]. Harmful content includes topics such as traffic accidents, murder, child abuse, animal cruelty, suicide, and more [26]. Reports say that there are approximately 100k commercial online content moderators [60, 68] and many more voluntary online community moderators [18, 81], who participate in online content moderation to prevent internet users from being exposed to harmful content. However, this process exposes content moderation workers to a vast amount of data that can be globally unidentifiable or inappropriate [7, 14]. The harm experienced by content moderators can be severe and have long-term effects [9, 20, 60]. These long-term impacts manifest not only as mental health issues such as depression [54], PTSD [28, 53], and emotional numbness [27, 46] but also as physical health problems like alcohol addiction [47] and insomnia [26]. Despite ongoing reports about their emotional labor [11, 13, 17, 52, 65], insufficient solutions exist worldwide to protect the widely dispersed and vulnerable workforce. In particular, research on video content moderation is still insufficient. In this work, we specifically focus on exploring intervention techniques to improve the working environment of video content moderation tasks and propose effective intervention techniques to reduce their negative emotions.

2.2 Strategies to Alleviate Emotional Burden in Content Moderation

Various strategies and research have been proposed to alleviate emotional labor in content moderation and even dataset annotation tasks. One of the prominent approaches is minimizing human intervention and introducing automated content detection systems [24, 63, 64, 83]. To achieve this, various machine learning-based prediction systems have been developed, assisting in automatically identifying problematic content [4, 12, 35, 58]. However, aspects such as context understanding, user intent comprehension, and handling ambiguous content still require human judgment and interpretation [45, 49, 61, 79]. To address these challenges, Steiger [68] introduces programming and technological approaches to improve the well-being of content moderators and alleviate psychological distress. Also, intervention techniques such as emotional state monitoring using image blurring and grayscaling [37] and emotional impact mitigation through interactive blurring [21] are being explored to reduce negative emotions that may arise during visual content moderation tasks. For industry, big tech companies offer video moderation tools that enable moderators to adjust blurring and grayscaling [69]. However, research into alleviating the psychological distress experienced by *video* content moderation workers during their tasks is still underexplored. Our work focuses on video content moderation workers, investigating whether a set of intervention techniques we have implemented effectively reduces negative emotions.

2.3 Reacting to Stress in Other Workspaces

Imposing numerous technical, inappropriate, or excessive demands on individuals in the workplace can lead to stress [8, 19]. Stress can negatively impact a person's everyday life, including depression, alcohol abuse, and unexplained physical symptoms [33]. To alleviate the negative emotions of work-related stress, employees need coping mechanisms. Reducing occupational risk factors for mental health issues and fostering workers' strengths and positive capabilities can help prevent mental health problems in the workplace [42]. There exist external strategies to assist employees in coping with stress, such as providing mental health counselor support for employees in corporations [15] and applying theoretical models about worker-workspace relationships [75]. Additionally, there are self-enhancement strategies that empower individuals to cope with stress. For example, improving psychosocial well-being in the workplace through the READY (Resilience and Activity for Everyday) program [6], enhancing psychological resilience through dedication, control, and challenge [41], emphasizing the role of intrinsic motivation in stress regulation [48], and improving well-being through the MBSR (Mindfulness-Based Stress Reduction) program [10]. However, ensuring these types of support can be particularly challenging for content moderation workers who are widely dispersed globally. While content moderators do not receive specific education and support on psychological well-being, research on specific stress responses in this occupational group remains insufficient. That is, the approaches from previous studies are challenging to apply to work like content moderation tasks because they primarily focus on dealing with post-task stress. The significance of our work lies in exploring intervention techniques that can be applied during the tasks.

3 DESIGN OF SIX INTERVENTION TECHNIQUES

We considered two approaches to select intervention techniques that help with the emotions of content moderators without significantly compromising the accuracy and efficiency of moderation task: screen processing that reduces visual stimulation and psychological assistance through additional interventions (text or videos). We selected six intervention techniques based on two approaches by drawing upon previous research and gaining our own insights through iterative pilot studies on different intervention ideas. The final candidates we selected are blurring, grayscaling, inserting motivational text, inserting motivational videos, interleaving with positive videos, and cartoonization. We also summarize how we implemented each intervention in our study.

Blurring. Previous studies show that blurring task images can enable tasks to be performed while reducing exposure to harmful content [16, 21, 37]. However, Karunakaran et al. [37] also report results where viewing entirely blurred images led to a decrease in participants' positive emotions and an increase in irritability. Das et al. [21] suggested that through the design of interactive image blurring, it is possible to maintain accuracy and speed while reducing emotional impact during content moderation tasks. We expected that blurring may have a similar effect on video content moderation tasks.

To alleviate the visual exposure derived from car accident scenes, we applied an 8px Gaussian filter to blur the videos (1920px × 1080px). Through this process, we aimed to create less intense visual stimulation while maintaining the core informational elements.

Grayscaling. The interaction between emotions and colors is widely recognized [31, 67]. We specifically investigated prior studies regarding the color red, which is associated with blood. Kwallek et al. [40] report that individuals exposed to red exhibited higher levels of anxiety and stress compared to those exposed to blue. Also, red is most commonly associated with negative emotions and emotionally charged words [70]. Kuhbandner et al. [39] report emotions such as anger and failure are associated with red. Based on the prior studies, we consider the necessity of screen processing to mitigate negative emotions that can be induced by provocative colors. In fact, Karunakaran et al. [37] report that grayscaling of images significantly enhances the positive impact on human reviewers in image moderation tasks. Considering the interaction between color and emotions, we expand the idea of applying grayscaling to dynamic videos.

We applied grayscale to videos depicting car accidents to shield participants' eyes from intense color stimuli. By transforming the originally vividly colored videos into monochromatic ones, our intention was to mitigate the impact of such visual stimuli on participants while preserving the essence of the content. For this purpose, we employed the CSS property grayscale: 100% to completely remove all colors from the images and present them in monochrome.

Inserting Motivational Text or Videos. Intrinsic motivation correlates with increased sustainability, psychological well-being, and

Criteria for Selecting the Six Intervention Techniques	
To Reduce Visual Stimulation	Blurring, Grayscaling, Cartoonization
To Help Psychologically	Motivational Text, Motivational Videos, Interleaving with Positive Videos.

Table 2: Distribution of intervention techniques according to visual and psychological criteria.

enhanced performance [23]. Deci et al. [22] report that intrinsic motivation increases when positive feedback is given, and Kamar et al. [36] suggested that emphasizing the significance of personal contributions to participants prevented users from leaving volunteer-based crowdsourcing platforms and increased user engagement. Also, it is known that intrinsic motivation effectively reduces stress and influences subsequent emotions [76]. Therefore, we explore the effect of both text and video messages designed to induce intrinsic motivation in reducing negative emotions for video content moderators. More specifically, we want to compare the impact of text, which is a convenient tool for information delivery, with video, which is richer in terms of the information it can deliver because it can convey both visual and auditory elements.

The motivational text provides encouraging messages aimed at eliciting intrinsic motivation in participants by instructing that their participation in the experiment could reduce car accidents and positively impact society. A short motivational text of about 5 to 6 sentences is provided for participants to read briefly. The motivational text is presented to participants in text format a total of twice, both before and after viewing the car crash videos.

The motivational videos intervention utilizes the same content as the motivational text intervention but presents encouraging messages in video format accompanied by positive music. A video composed of text is designed as a short video lasting less than 1 minute, where participants must watch the video entirely before they can proceed to the next page. Participants are exposed to the motivational video a total of twice, before and after viewing the car crash videos. Considering emotions that can be evoked through music [34], we incorporate upbeat music in the videos to provide participants with both visual information and auditory stimuli.

Interleaving with Positive Videos. People often reduce stress by consuming positive content on the internet [3, 17]. Particularly, researchers found that viewing cute animals can contribute to decreasing stress and anxiety [50, 56]. Also, it is reported that exposure to the wonders of nature or landscapes can enhance stress relief and well-being through awe [62]. In particular, observing green landscapes can be effective in supporting relaxation and recovery after periods of high stress [73]. People being exposed to stunning natural images through images or videos has also shown effective results [38].

During the video annotation task, we inserted six videos to elicit positive emotions between car crash videos. The positive video is inserted right before watching the car crash video, and after every two-car crash videos, one positive video is inserted. After completing annotation tasks, the final positive video is inserted. These positive emotion-inducing videos featured themes based on words that can evoke positive emotions in people, sourced from the International Affective Picture System (IAPS) dataset. (Example: beautiful landscapes, baby animals) Therefore, based on the IAPS dataset, we select videos containing words associated with positive emotions [43].

Cartoonization. It has been studied that cartoons can offer short-term relief from symptoms of depression [66]. In particular, cartoon distraction has proven highly effective in alleviating anxiety in children [44]. The benefits of cartoons are not limited to children but extend to adults as well. It has been studied that adults with mental illnesses engage in self-therapy through cartoon animation [2]. Animation therapy holds the potential for psychological well-being enhancement to the extent of being utilized in art therapy [57]. Since automatic cartoonization of images and videos became possible through various machine learning algorithms, we explore the effect of cartoonization in video content moderation tasks. We expected that converting disturbing video content into a cartoon style establishes emotional distance and reduces emotional shock for participants. This transformation is offered with the intention of mitigating the emotional impact by creating a sense of detachment.

Cartoonization involves transforming original videos into a cartoon-like style. For video transformation, we employed a white-box cartoon representation [77]. We divided the video into 60 frames, converting each frame into an individual image. Then, we provided these frames as input to the model to extract the white-box cartoon representation. We obtained a cartoon version of the original video by combining the transformed individual image frames.

4 MEASURING THE EFFECT OF THE SIX INTERVENTION TECHNIQUES

In this study, we aim to evaluate the effectiveness of the six intervention techniques in reducing negative emotions while watching disturbing videos. We first conducted a between-subjects crowdsourcing study to quantitatively evaluate the emotional impact of each intervention technique. In addition to measuring participants' emotions, we compared the accuracy of the conducted tasks when using different intervention techniques. Because each intervention technique affects the original task videos differently (e.g., applying filters or inserting other content), they may affect the workers' performance differently. To understand whether there exists a trade-off between the effect of the interventions and the task performance, we analyzed not only the emotional change after conducting video moderation tasks, but also the average accuracy of the crowdsourced tasks across the six interventions.

4.1 Research Ethics

We chose car crash videos that are oftentimes shared through social media and are also used to train autonomous vehicles through manual annotation. Car crash videos are inherently distressing, and we anticipated that exposure to these videos would increase negative emotions among the participants. Considering the sensitivity and

Positive		Negative	
Interested	Excited	Distressed	Upset
Strong	Enthusiastic	Guilty	Scared
Proud	Alert	Hostile	Irritable
Inspired	Determined	Ashamed	Nervous
Attentive	Active	Jittery	Afraid

Table 3: Positive and Negative Affect Schedule (PANAS)

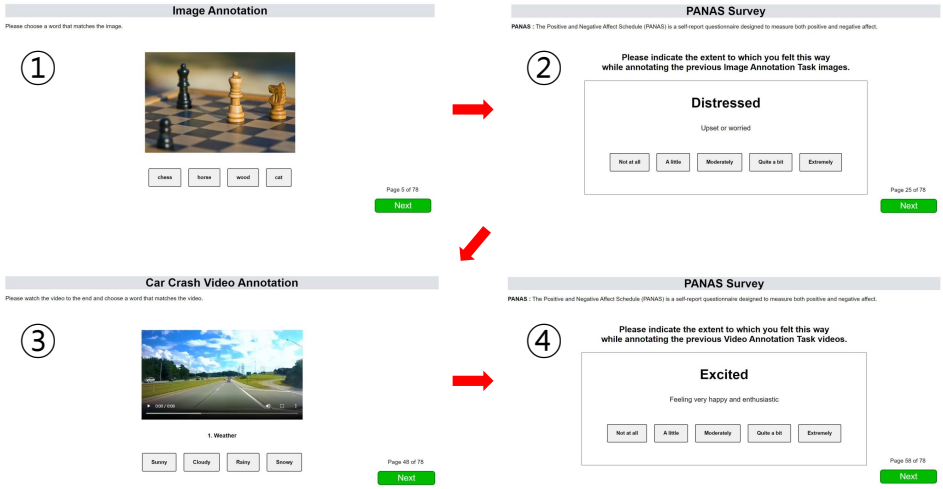

Figure 2: AMT study procedure: (1) Image Annotation: participants label the names of words that are rated with neutral emotions. (2) PANAS Survey (Before): participants answer the PANAS survey to measure emotions after image annotation but before the main moderation task of video annotation. (3) Video Annotation: participants respond to moderation tasks after watching car crash videos. (4) PANAS Survey (After): participants answer the PANAS survey to measure emotions after the moderation task.

ethical aspects of the task, we implemented the following measures to reduce the mental burden on the crowd workers involved. First, prior to the start of the experiment, we clearly informed participants that the task involved watching car crash videos that could be disgusting to some people, and participants could withdraw from the study at any time on their own decision. Second, we only showed collisions between cars, which did not include any living being injured, i.e., excluding any human or animal collisions. Third, we edited and presented the video for about 10-15 seconds, not showing any direct images after the accident, such as bloodstains or the video of an injured person after the accident. We note that we received approval from our Institutional Review Board (IRB).

4.2 Measurement

We utilized one of the most widely used emotion measurement scales, the PANAS (Positive and Negative Affect Schedule) scale [78], to measure participants' emotions before and after the moderation task. Each emotion category was measured using a 5-point Likert scale. PANAS comprises 20 emotion categories: 10 for measuring positive affect and 10 for measuring negative affect as listed in Table 3.

4.3 Participants

We recruited 280 participants from AMT to assess the effect of the six intervention techniques and a baseline condition where no intervention was applied. We recruited 40 AMT workers for each intervention and the baseline condition, excluding duplicate workers. Participants were selected based on two criteria through AMT, ensuring that their residence was limited to the United States and that they had accumulated more than 500 approved Human Intelligence Tasks (HITs) for quality control.

4.4 Procedure

Participants read and signed the informed consent form before taking part in the experiment and providing demographic data. Figure 2 visually shows the overall experimental process for the AMT study. Participants experience various emotions in their daily lives, potentially influencing the PANAS survey. Therefore, in the initial phase, participants were shown 20 neutral pictures to minimize the possibility of bias and PANAS was measured while participants were in as emotionally neutral state as possible. To make sure they pay attention to the neutral pictures, they were asked to choose the name of each picture with one of the four provided options.

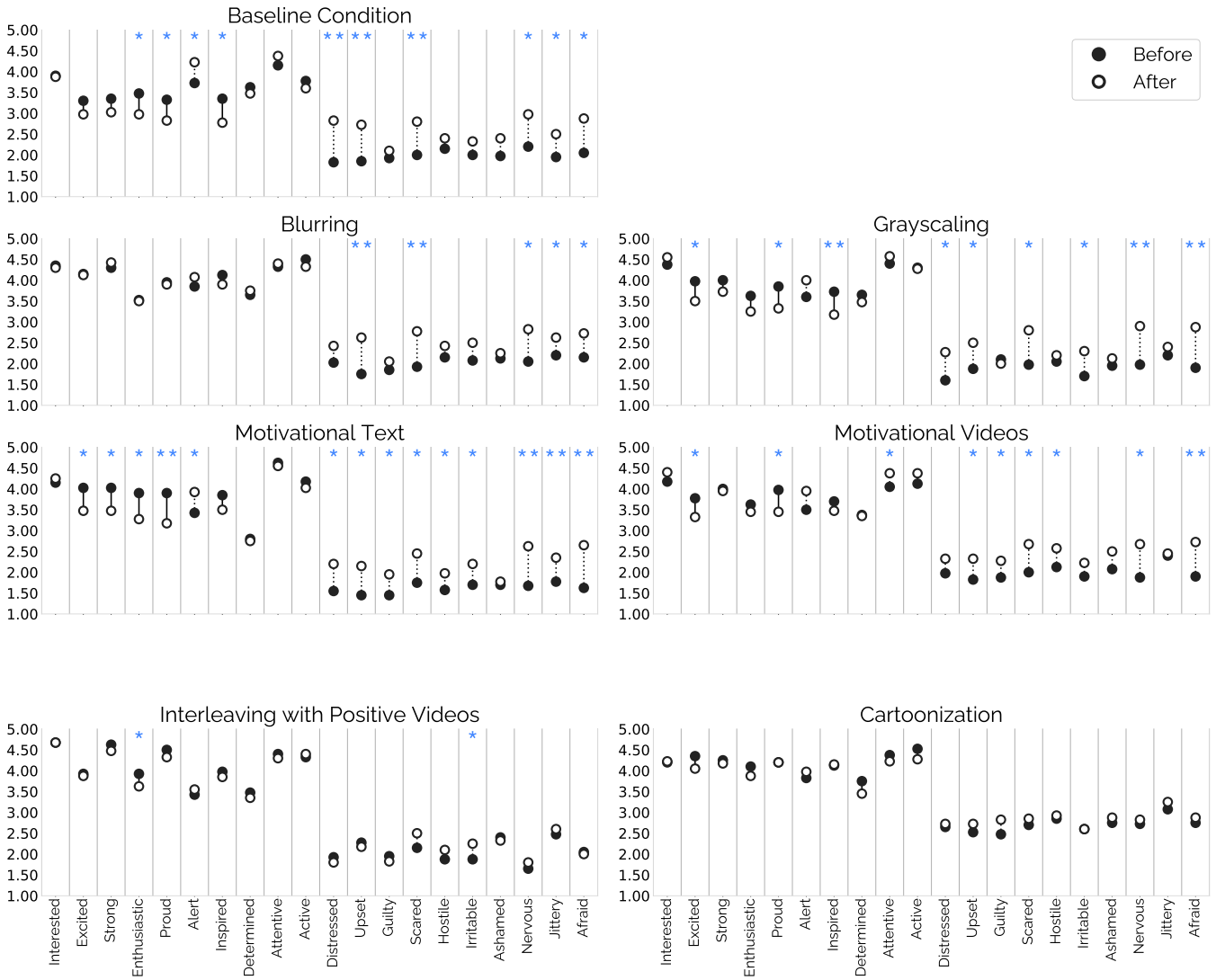

Figure 3: Visualization of participants' emotions that changed before and after the moderation task. Interleaving with positive videos and cartoonization techniques (intervention techniques in the last row) showed the least change in emotions, especially for the negative emotions. This indicates that these two interventions effectively alleviated negative emotions from the sensitive video moderation tasks. * indicates $p < 0.05$. ** indicates $p < 0.0025$, the significance level with Bonferroni correction applied.

The selection of neutral pictures was based on the dataset from the IAPS [43]. In the IAPS database, images with valence scores between 4 and 6 are considered neutral. Therefore, we selected 20 words within this range for image annotation. Following this phase, participants were asked to answer the 20 PANAS questions one by one. After answering the PANAS scale, participants were assigned to one of six intervention techniques or the baseline condition and exposed to 10 car crash videos. As a moderation task, participants were asked to report the weather of the scene, the location of the accident, and the number of cars involved in the crash for each video. After conducting tasks on all 10 car crash videos, participants were asked to repeat answering to the PANAS scale to measure their emotions.

4.5 Results

To answer RQ1, we measured PANAS score before and after the moderation tasks. We ran Shapiro-Wilk test and the responses were in ordinal levels that do not follow a normal distribution ($p < 0.05$). Therefore, we conducted statistical analyses with Wilcoxon signed-rank test, considering that PANAS was measured using an ordinal scale. Because there were twenty emotion categories, we applied the Bonferroni correction to ensure a more rigorous validation, setting the significance level at 0.0025.

4.5.1 Effectiveness in Alleviating Negative Emotions. We note that interleaving with positive videos and cartoonization demonstrated a marked reduction in the increase of negative emotions. Figure 3

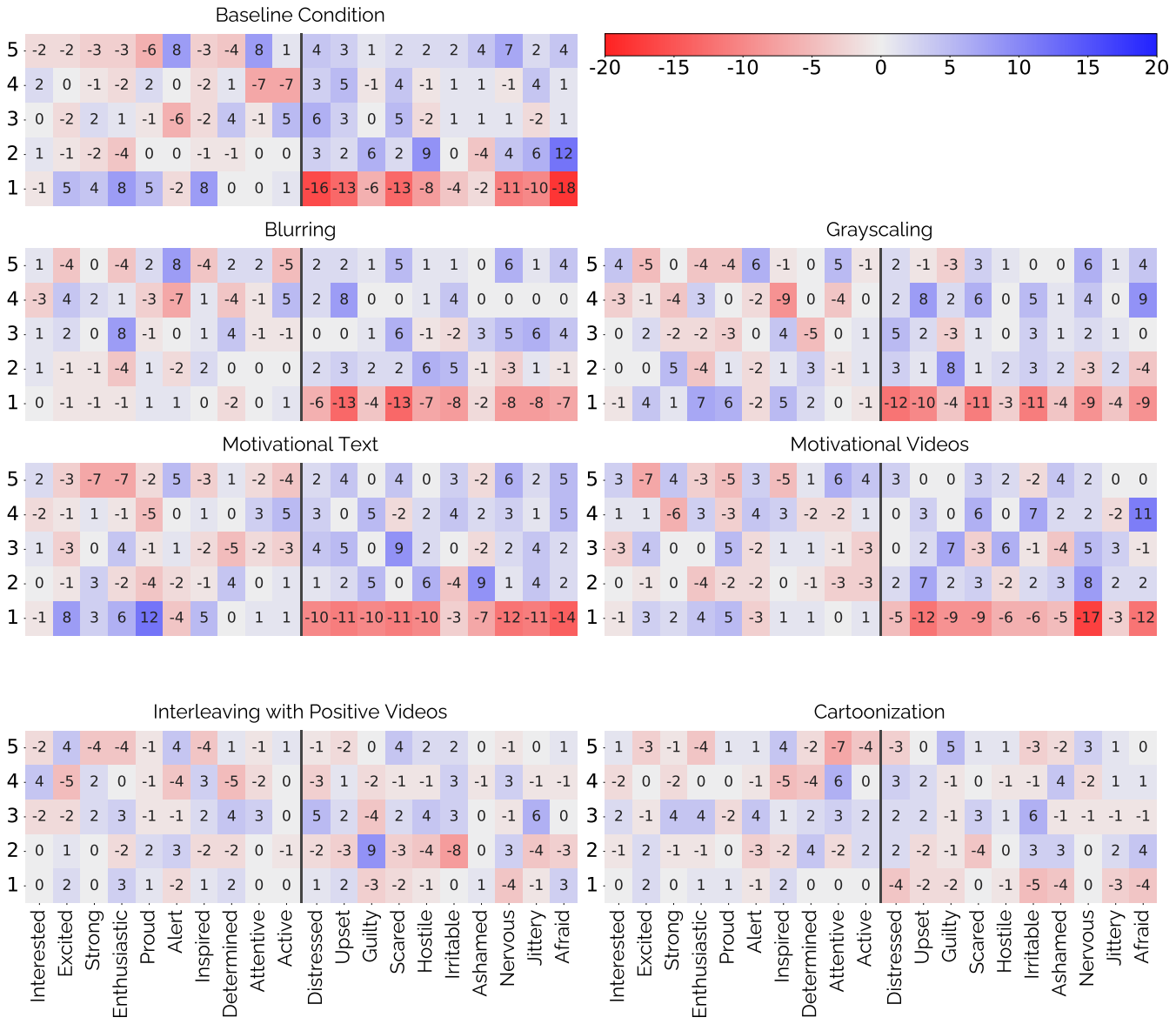

Figure 4: Color maps depicting the change in the number of participants to emotions before and after a moderation task. The red cell (with negative number) indicates participants who shifted from their initial emotion score, while the blue cell (with positive number) represents those who shifted from their final emotion score, with color intensity indicating the magnitude of change. Interleaving with positive videos and cartoonization techniques (intervention techniques in the last row) shows little change in the participants' emotions before and after the experiment compared to the rest of the conditions.

illustrates the average change in each emotion category for 40 participants before and after the experiment for each intervention technique. As shown in the figure, only one negative emotion (irritable) showed statistically significant increases after the moderation task with interleaving with positive videos ($p < 0.05$). And no statistically significant increase was observed in cartoonization. On the other hand, at least five negative emotions significantly increased in the baseline condition and other intervention techniques. We report the statistical test results for each intervention technique (Table 6).

Baseline Condition. We observed six out of ten statistically significant increases ($p < 0.05$) for the negative emotions: Distressed, Upset, Scared, Nervous, Jittery, and Afraid.

Blurring. We observed five significant increases for the negative emotions: Upset, Scared, Nervous, Jittery, and Afraid.

Grayscaling. We observed six significant increases for the negative emotions: Distressed, Upset, Scared, Irritable, Nervous, and Afraid.

	Task Accuracy						
	Baseline Condition	Blurring	Grayscaling	Motivational Text	Motivational Videos	Interleaving with Positive Videos	Cartoonization
Population	40	40	40	40	40	40	40
Mean	22.38	18.0	21.08	22.50	22.90	21.05	13.15
Std. Deviation	6.13	4.67	4.88	4.71	4.40	4.61	2.93

Table 4: The number of participants and their mean and standard deviation of the moderation task's accuracy for each intervention. Blurring and Cartoonization, which are screen processing on the original video, tend to result in lower task accuracy compared to other intervention techniques.

Inserting Motivational Text. We observed nine significant increases for the negative emotions: Distressed, Upset, Guilty, Scared, Hostile, Irritable, Nervous, Jittery, and Afraid.
Inserting Motivational Videos. We observed six significant increases for the negative emotions: Upset, Guilty, Scared, Hostile, Nervous, and Afraid.
Interleaving with Positive Videos. We observed a significant increase for the negative emotions: Irritable.
Cartoonization. We observed no significant differences in negative emotions.

Figure 4 shows a color map that depicts the absolute difference in the number of participants whose PANAS scores shifted before and after conducting the moderation task. The red (with a negative number) indicates the number of participants who selected the corresponding score for the emotion before the experiment but did not select the score after the experiment and chose a different score. The blue (with a positive number) indicates the opposite situation. That is, the more intensive the color is (or the larger the absolute value of the number is), the more people experienced a change in their emotions. In the figure, the horizontal axis represents various emotions, while the vertical axis corresponds to PANAS scores. Unlike experiments with other intervention techniques, experiments involving interleaving with positive videos or cartoonization did not show a pronounced difference for negative emotions. Through these results, we can infer that interleaving with positive videos and cartoonization experiments may protect participants from increased negative emotions induced by the task.

4.5.2 Trade-off in Accuracy. We further examined the task accuracy. In the experiment, participants watched 10 car crash videos and annotated 30 questions related to the videos. Table 4 reports participants' average accuracy and standard deviation of the annotation tasks for each intervention technique. Intervention techniques without any processing on the original videos showed a similar trend in task accuracy. In contrast, blurring and cartoonization, associated with video quality degradation, indicated lower task accuracy than other intervention techniques.

4.5.3 Summary of the Results. Our statistical data suggest that negative emotions can be aroused in the baseline condition when participants are exposed to harmful video content. Additionally, we have statistically demonstrated that interleaving positive videos and cartoonization effectively reduce negative emotions among the six intervention techniques we proposed. We observed that blurring and cartoonization tend to decrease task accuracy.

However, our quantitative analysis has limitations in explaining the underlying reasons for the results produced by these intervention techniques. Hence, we conducted an additional in-lab within-subjects study with newly recruited participants to gather qualitative data, aiming to understand the characteristics of six intervention techniques and discuss their potential correlations with quantitative data.

5 UNDERSTANDING THE EFFECT OF THE SIX INTERVENTION TECHNIQUES

To further understand why certain interventions were effective and others were not in reducing negative emotions during the video moderation tasks, we conducted an in-lab within-subjects study with seven participants who also went through in-person post-task interviews. We used the same research ethics and measurement as described in Section 4, where the recruitment and the experiments were in accordance with our institution's IRB policies.

5.1 Participants

We recruited participants through email recruitment announcements. The announcement mentioned that the process of watching a car accident video is included and that participants can drop out whenever they want and stop participating at any time they want. A group of people in their 20s (4 male, 3 female, mean age = 24.29, stdev = 1.75, max = 27, min = 21) was recruited. The study lasted for 60 minutes long, and the participants were paid 15,000 KRW (approximately 12 USD). We obtained consent forms and demographic information from the seven recruited participants.

5.2 Procedure

The study was designed as a within-subjects study, where each participant experienced all interventions sequentially. We used the Latin square design to control sequencing effects between different interventions. We explained to the participants that the purpose of our study was to find intervention techniques that would help reduce negative emotions for content moderators exposed to harmful content, and described the characteristics of each intervention technique. We used visual materials to explain how each intervention will appear on the monitor. Additionally, we explained the content moderation task, the content moderator, and the work environment

to the participants in detail, and noted that our experimental setting simulated their work environment.

Consistent with the first experiment with crowd workers on AMT, participants watched 10 car crash videos for each intervention and were asked to answer three questions per video. The videos did not overlap between different interventions. After watching one car crash video, the participants were asked to answer three questions regarding the video. After completing the simulated moderation tasks with each intervention technique applied, we asked participants through in-person interviews to obtain their opinions about the intervention techniques such as the possibility of use in workplaces, whether it helped reduce negative emotions, and their thoughts on the advantages and disadvantages of the interventions. We asked the same questions for each intervention technique, except for the baseline condition (original video without any intervention technique applied). Appendix B.1 presents the entire set of questions asked in the interview. Because no intervention technique was applied to the baseline condition, three of the questions were excluded for the baseline condition. After the participants completed all seven intervention techniques, we conducted a final interview with the participants. Appendix B.2 presents the entire set of questions asked in the final interview. We aimed to compare the seven intervention techniques experienced by the participants in the final interview. Through this comparison, we wanted to find out which interventions reduce negative emotions and why.

5.3 Results

To answer RQ2, we summarize key findings revealing why certain intervention techniques helped alleviate negative emotions of the video content moderation tasks.

5.3.1 Ranking Preference of Interventions. We asked participants the following question to evaluate the relative effectiveness of seven techniques: "If you were a content moderator handling harmful videos, what intervention techniques would you use to protect your emotions? (Please order from the most preferred to the least.)" To clearly distinguish the difference between different rankings when aggregating all seven participants' responses, we applied exponential weighting on each rank before aggregating the responses. As a result, we got the following aggregated ranking order: (1) interleaving with positive videos, (2) cartoonization, (3) grayscaling, (4) blurring, (5) motivational videos, (6) motivational text, and (7) baseline condition, from the most preferred to the least. In particular, interleaving with positive videos and cartoonization were confirmed as methods that significantly reduce negative emotions in our first experiment on AMT and also ranked high in this question. We sought to understand the reasons behind these results through in-person interviews and aimed to uncover the reasons by quoting participants' responses.

5.3.2 Emotional Protection. We asked participants the following question using a 5-point Likert scale: "If this intervention technique were implemented, would it help reduce your negative emotions when dealing with provocative or unpleasant videos?" Figure 5 visually represents participants' choices for each intervention technique in the given question. The analysis of participants' responses revealed that interleaving with positive videos and cartoonization had the highest proportions of 'strongly agree' and 'somewhat agree' responses. Following that, grayscaling and blurring had few 'neither agree nor disagree' and 'somewhat disagree' opinions and mostly consisted of 'strongly agree' and 'somewhat agree' responses. In contrast, motivational videos and motivational text had lower proportions of 'strongly agree' and 'somewhat agree' responses. An interesting observation was that cartoonization had the highest proportion of 'strongly agree' responses, surpassing interleaving with positive videos, which were suggested to be the most helpful for emotion protection.

5.3.3 Most Impressive Intervention Technique. To understand why cartoonization received more 'strongly agree' responses, we analyzed responses to the question: "What intervention technique impressed you the most? Additionally, what were the reasons for your choice?" The results are shown in Table 5. Most participants mentioned that cartoonization and interleaving with positive videos were the most impressive, with a tendency to favor cartoonization. Participants mentioned that cartoonization reduced visual stimulation more than they expected, and its forward-looking technological contribution was highlighted for its potential to alleviate negative emotions effectively. Participants noted interleaving with positive videos as an impressive technique, reporting that they experienced mood offset and reduced stimulation due to distraction through positive videos. Interestingly, some participants mentioned the potential for technological advancement and contributions related to cartoonization. P4 noted, "The others can easily be done with video editing, but cartoonization involves technology so it might get better with technological advancements."

5.3.4 Task Accuracy Order of Intervention Techniques. We provided participants with the following question to evaluate the relative effectiveness of seven intervention techniques: "If you were a content moderator handling harmful videos, what intervention techniques would you use for task accuracy and speed? (Please order from the most preferred to the least.)" The responses of participants were ranked using exponential weights for distinctiveness. The resulting aggregated order is as follows: (1) baseline condition, (2) motivational text, (3) grayscaling, (4) interleaving with positive videos, (5) motivational videos, (6) cartoonization, and (7) blurring. Our first experiment on AMT indicated that blurring and cartoonization tended to exhibit lower task accuracy than other intervention techniques, this trend was also evident in this second experiment. The underlying cause of these results can be attributed to the reduction in visual information due to the degradation in video quality, as confirmed in Table 4, which significantly impacts accuracy deterioration.

5.3.5 Summary of the Results. We comprehensively investigated the characteristics of six intervention techniques through an in-lab within-subjects study. We asked participants to evaluate the advantages and disadvantages of each intervention technique, and the results are summarized in Table 7. We report suggested improvements to the intervention techniques based on their characteristics in discussion section 7.2.

We note that the quantitative results from the AMT study and the qualitative results from the in-lab within-subjects study are consistent. Our study results reveal that interleaving with positive

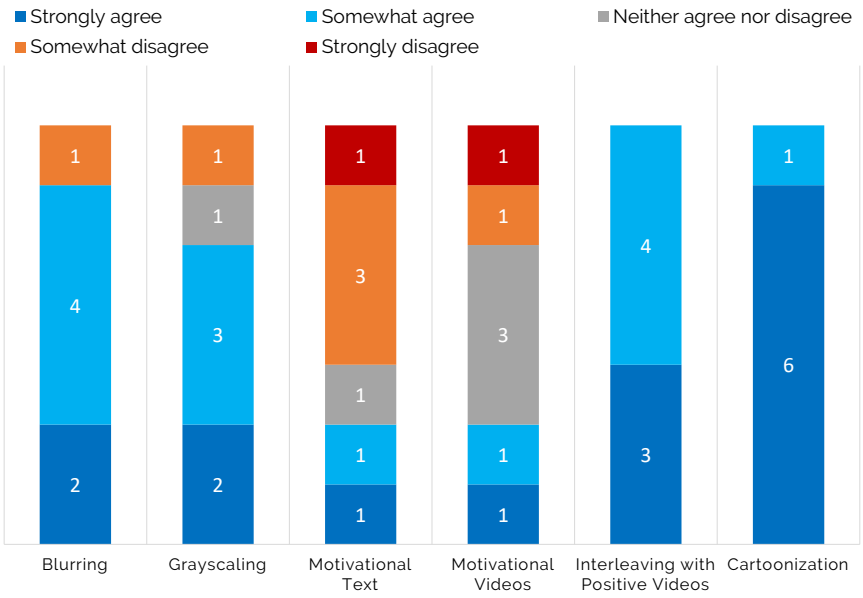

Figure 5: A visual representation of participants' choices about whether each intervention technique would help reduce negative emotions. The response rates of 'strongly agree' and 'somewhat agree' were the highest in cartoonization and interleaving with positive videos.

Participants	Intervention Technique	Reason
P1	Cartoonization	Clear cartoon-like transformation
P2	Cartoonization	Sharper deformation than expected
P3	Cartoonization	Most impressive with anticipated future applications
P4	Cartoonization	Expectation of effectiveness based on technological advancement
P5	Interleaving with Positive Videos	Potentially helpful way to counterbalance the negative with something positive
P6	Interleaving with Positive Videos	Significantly better than expected results, dispersal effect of concentration
P7	Blurring	Videos looked better than I expected and stimulation reduction effect

Table 5: Summary of all participants' impressions of intervention techniques that they were the most impressed with and their reasons. Most participants mentioned cartoonization and interleaving with positive videos was the most impressive.

videos helps offset mood and distract concentration by allowing people to experience emotional relaxation through positive videos. On the other hand, cartoonization reduces visual stimuli and fosters a sense of detachment from reality by feeling similar to cartoons or games. We further aimed to better understand the effectiveness of the interventions in daily life, especially regarding the two most effective interventions: interleaving with positive videos and cartoonization. Therefore, we conducted a three-day between-subjects study as our last investigation.

6 UNDERSTANDING THE IMPACT OF THE TWO EFFECTIVE INTERVENTIONS IN DAILY LIFE

The main goal of this study is to find whether continuous exposure to harmful content affects participants' daily lives and emotions and whether the intervention techniques found effective would still be effective. We conducted an in-depth investigation of baseline condition, interleaving with positive videos, and cartoonization through a three-day between-subjects study. Three participants were recruited per each condition, totaling nine participants.

6.1 Participants

We conducted a three-day user study with a group of participants (4 male, 5 female, mean age=23, stdev=2.58, max=26, min=19) recruited from our institution. This study was 60 minutes long per day, and participants were paid 50,000 KRW (approximately 39 USD) in total. Initially, we recruited participants through e-mail announcements. The participants did not overlap with any previous studies. Again, we used the same research ethics and measurement as described in Section 4, following our institution's IRB policies.

6.2 Procedure

A between-subjects study was conducted with nine participants divided into three condition groups (baseline condition, interleaving with positive videos, cartoonization), where each group consisted of three participants. We asked the participants to conduct the task from the perspective of content moderators who have to watch sensitive videos. They watched 20 car crash videos every day and conducted an annotation task for each video, the same as the annotation tasks conducted in the previous two studies. After each day of the experiment, participants filled out their answers to a brief self-report questionnaire. The following experiments were conducted for a total of three days. Appendix C.1 presents the self-report questionnaire on the first day. Appendix C.2 shows the self-report questionnaire on the second and third days of the experiment. The questionnaires for the first day and the rest of the days are different because the latter days had questions to find out if the previous day's experiment had affected the next day's routine and emotions.

After the three-day user study, we conducted face-to-face interviews with the participants. We asked the same questions for each intervention technique except the baseline condition. Appendix C.3 presents the questions asked in face-to-face interviews. Because the intervention technique was not applied to the baseline condition, questions asking for information about the intervention technique were excluded from interviews with participants who experimented with the baseline condition. We conducted a final survey three weeks after the experiment to see what changes had occurred in their emotions and daily lives. Appendix C.4 presents the questions asked in the final survey.

6.3 Results

Participants were free to conduct experiments at any time of the day, from anywhere for three days in a row. After completing the experiment, participants immediately responded to self-report questionnaires on a Google form we provided. To answer RQ3, our questions were designed to assess participants' emotional state at the time of finishing the simulated moderation task, and to check whether the previous day's experiment had an impact on the participants' emotional state and daily life. Many questions overlapped with the questions from the previous in-lab within-subjects study, and their answers also exhibit similar patterns. Here, we only report newly discovered responses in the three-day experiment.

6.3.1 Impact of Intervention Techniques in Reducing Negative Emotion in daily life. Participants were asked to respond to the daily survey question: "Did you experience any changes in your mood after performing this experiment today?" Most participants reported experiencing negative emotions due to the car crash videos and provided explanations. Some participants also described the effects of intervention techniques. This content is summarized in keyword format on Table 8.

Additionally, in face-to-face interviews conducted after the completion of the three-day user study, participants' responses to the question, "If this intervention technique were implemented, would it help reduce your negative emotions when dealing with provocative or unpleasant videos?", and similar responses from the in-lab within-subjects study was observed. Participants reported that cartoonization and interleaving with positive videos were effective in reducing negative emotions. This suggests that the intervention techniques that were found to be more effective in the short-term between-subjects study may be also effective in longer-term usage.

We note that, in the follow-up survey after three weeks, all participants reported that they did not have impact from the user study anymore, which indicates that the three-day study was long enough to impact the participants' mood, but did not have a longer-term negative impact to cause more serious issues.

6.3.2 Feasibility of the Intervention Techniques to Work on Content Moderation Tasks. When we asked participants if they could perform content moderation as a job, all participants who had experienced baseline condition answered that they would not. However, some of the participants who had experienced interleaving with positive videos or cartoonization intervention techniques responded that they might be able to do it if those intervention techniques were applied.

> "After encountering various stimulating videos, it seems I can handle the job if I have my own regular break time with positive videos." - P6

> "If cartoonization intervention technique were applied, I think continuous content moderation work would be possible. On the contrary, if not, lingering images of stimulating videos would make daily life uncomfortable, and I might not be able to continue working." – P8

When asked whether the intervention technique they experienced would help them when exposed to harmful videos for a long time as part of their job, all those who experienced interleaving with positive videos responded that it would be helpful.

> "If I watch another video that is completely unrelated to the harmful video during my break, I think it will help a lot in fading the memory of the harmful video, and I have experienced it myself." – P4

Additionally, all participants who experienced cartoonization said that it would be helpful.

> "Scenes with the cartoon filter seem to reduce visual stimulation compared to real videos, which could alleviate the burden on the workers." – P9

6.3.3 Task Accuracy. Participants watched 20 car crash videos daily and responded to 60 questions in total as their moderation tasks. We measured task accuracy for each intervention technique over three days, and the average task accuracy for participants appeared as follows: baseline condition (mean=44.67 stdev=4.29), interleaving with positive videos (mean=44.33 stdev=3.83), and cartoonization (mean=45.67 stdev=3.13). This result indicates that, contrary to the AMT study's findings, cartoonization tended towards similar accuracy to the baseline condition. In the AMT study, anonymous participants may have been inclined to rush through tasks for efficiency, whereas in the three-day between-subjects study, participants worked under the supervision of experimenters, potentially allowing them to focus more on the task. The three-day between-subjects study was designed to create an environment

similar to that in which content moderators work. Content moderators would strive for precise moderation, and it is our expectation that cartoonization's accuracy could yield task accuracy similar to the baseline condition, as observed in the results of the three-day between-subjects study.

6.3.4 Summary of the Results. Participants reported experiencing negative psychological impact when viewing car crash videos, and from the perspective of content moderators, they indicated that continuing such work could lead to severe psychological harm. However, some participants who watched videos with intervention techniques applied mentioned that they could continue content moderation work if these intervention techniques were applied. For the reason, they mentioned that interleaving with positive videos helps improve mood and reduce distraction by enabling individuals to experience emotional relaxation through such content, while cartoonization diminishes visual stimuli and cultivates a sense of detachment from reality, resembling the experience of cartoons or games. Additionally, we report the effects of the intervention techniques mentioned by the participants, and we conclude that the results of both the AMT study and the in-lab within-subjects study indicate consistent effects in this study.

7 DISCUSSION

In this section, we discuss the emotional well-being improvement and potential impacts of interleaving with positive videos and cartoonization based on the results obtained from three studies and the possibility for improvement in other intervention techniques. Additionally, we discuss efforts to overcome limitations in recruiting participants and whether our findings can be applied to content moderators in the wild.

7.1 The Immediate Effects of Intervention Techniques during Tasks

Some participants in a three-day controlled study expressed a preference for managing stress by separating work and daily life activities after engaging in moderation tasks. They mentioned examples such as exercising, reading books, or taking breaks. Additionally, there were also opinions about seeking psychological assessments or treatment to prevent trauma. P9 noted, "I will make an effort to separate daily life and work, and if necessary, consult with experts to ensure that trauma does not develop, and I can receive medical treatment." These responses demonstrate that participants need to spend additional time and cost after leaving their work to relieve stress. Additionally, continuous exposure to harmful content during work has the risk of causing psychological damage to content moderators.

Participants mentioned that the use of intervention techniques has the advantage of preventing stressful situations in advance rather than finding solutions after the stress has occurred. They also noted that interleaving with positive videos and cartoonization could provide mental support for the emotional labor of long-term content moderation tasks. These intervention techniques not only indicate important benefits for individuals performing content moderation tasks in terms of time and cost spent on stress relief, but can also help individuals maintain their work.

7.2 Potential for Improvement in Intervention Techniques

Participants shared insights on the potential development of intervention techniques and alternative intervention techniques through in-lab within-subjects study. Each intervention technique has diverse characteristics and needs to address shortcomings. We summarize improvements in intervention techniques based on their feedback, providing keywords.

Blurring Sharpness, Interactive Intensity Control. Participants mentioned that allowing users to directly control the sharpness and position of the blur when applying blurring on the screen would improve accuracy. "It would be better if the blurring looked slightly sharper [P1]," "While blurring overall is good, it would be even better if I could blur selectively [P7]."

Grayscaling Switching to Original. Participants expressed concern that while grayscaling could emotionally help by removing stimulating colors like blood, it might decrease the accuracy of moderation tasks related to color. Therefore, they mentioned that a switching function between the original video and the processed video would likely be frequently used. "It didn't feel as real and the visual stimulation was lessened. However, in the case of videos that require color identification, I will switch back to the original videos frequently [P6]."

Sequence and Frequency of Motivational Text, Providing Audio Support. Participants noted that repetitive phrases might lose meaning when encountered daily. They showed a preference for having motivational text read aloud by a person rather than reading it themselves. Additionally, to maximize the benefits of motivational text, it was suggested that providing it before starting work would be beneficial. "It would be great if the meaning is conveyed through someone speaking. Also, it would be best if it's placed at the beginning before starting work [P4]," "Having the text appear at the beginning is good. If there's auditory support, like someone verbally explaining, it would be even more effective [P7]."

Maximizing Visual Materials, Providing Audio Support in Motivational Videos. Participants mentioned that it would be more effective if someone explained the content in the video or if the visual materials were maximized. "I think conveying the content through someone speaking in the video rather than text would be more effective. It would be great to maximize the visual aspect [P4]."

Simultaneous Provision of Positive Videos, User-Specific Frequency Control, Preferred Video. Some participants expressed concerns that unrelated videos appearing during moderation tasks could disrupt the workflow. Instead, they mentioned that showing positive videos constantly alongside content could increase the efficiency of tasks. Additionally, some other participants suggested that it would be more effective if workers could decide the frequency and content of positive videos they want. "To prevent the workflow from being disrupted, it would be better to have the videos displayed next to the task rather than provided separately [P1]," "It would be better to be able to adjust the frequency of occurrences of positive videos according to individual preferences. It would be more effective if I could see videos about things I like [P7]."

Diversity of Cartoonization Filters, Adjustment of Cartoonization Range. Participants praised the effect of cartoonization while mentioning that it would be better to provide various

cartoonization filters by AI techniques. Also, they mentioned that if the cartoonization could be applied locally to areas of the video that pose a risk rather than applied to the entire video, it could increase efficiency and accuracy. "It would be better to have a feature where we can choose our preferred filter through a variety of cartoonization filters [P3]," "I think it would be more effective to cartoonize only parts, not the entire cartoonization. If there is an act of stabbing someone with a knife, changing the act itself into a cartoon would make it less visually stimulating and improve work performance [P6]."

Other Intervention Techniques: Utilization of Auditory Materials, Sound Filtering. Participants mentioned the need for intervention techniques that would be helpful auditorily in addition to those we proposed. "It would be better to hear something auditory like a radio that would be helpful. It would be better to hear interesting words rather than words of comfort. Baby cries were shocking. [P1]," "Even though the screen changed, the sound remained. So, I felt scared. It would be better to filter sounds like baby cries or people's screams [P7]."

7.3 Bridging the Gap Between Experiment Participants and Content Moderators in the Wild

We recruited young users with a high understanding of social media platforms as participants in our experiment. Although they are not professional content moderators, the experimental results can adequately represent the experiences of content moderators because the emotional mechanisms are similar between individuals regardless of their occupations. The primary focus of our research is not on improving or investigating the accuracy or performance of content moderation, but rather on examining how intervention techniques can alleviate negative *emotions* when exposed to harmful content.

Nevertheless, to overcome the limitations of participant recruitment, we made efforts to ensure that participants could understand and approach the experiment from the perspective of content moderators. We provided detailed explanations to those unfamiliar with content moderation and provided a controlled experimental environment similar to the environment in which content moderators work. This approach could also help in alleviating negative emotions for content moderators.

8 LIMITATIONS AND FUTURE WORK

We have recognized several limitations in our research from user studies and their results, which future research can explore more deeply.

First, we selected videos that would not cause significant mental harm to participants through IRB review, considering the ethical issues that could arise from experiments. Therefore, we used car crash videos where humans do not appear directly and have no interaction with any animals. Many participants in two controlled user studies mentioned that car crash videos are frequently encountered in everyday life and are insufficient to evoke negative emotions. In future work, we will explore the impact of intervention techniques on reducing negative emotions using a variety of video content.

Second, while most participants considered the use of intervention techniques to be sufficiently useful, some mentioned that the use of techniques such as blurring, grayscaling, and cartoonization to transform original videos could reduce accuracy. Based on the results of the AMT study, we need to consider the accuracy to maximize the utilization of intervention techniques. Although the differences in accuracy between baseline condition, interleaving with positive videos, and cartoonization were not reported in a three-day user study, there is a need for future work to improve the accuracy of intervention techniques to maintain accuracy and efficiency in content moderation tasks.

Third, a three-day experiment is a short-term investigation that may be insufficient to evaluate the long-term effects of intervention techniques. Future work may conduct a longitudinal study (3-6 months) evaluating the impact of interleaving with positive videos and cartoonization on the well-being of video content moderators, compared to previously suggested intervention techniques.

Finally, through the potential for improvement in intervention techniques mentioned in the discussion section 7.2, we anticipate that various intervention techniques can be newly proposed and developed, utilizing the characteristics of proposed intervention techniques well.

9 CONCLUSION

In this paper, we introduce six intervention techniques that can be applied to video content moderation. We found that interleaving with positive videos and cartoonization are the optimal intervention techniques and the potential for these intervention techniques to assist content moderators in better coping with their tasks. Also, we comprehensively summarize user opinions on the six intervention techniques and provide guidelines for improvement. Finally, content moderators work behind the scenes to safeguard the mental well-being of internet users. We want to make the cyber environment safer and happier for everyone by exploring new intervention techniques to protect content moderators.

ACKNOWLEDGMENTS

We are grateful to all the participants for lending their time and sharing their experiences. We also thank the reviewers for their valuable insights and constructive feedback. This work was supported in part by the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (NRF-2022R1C1C1003123, NRF-RS202300211872) and the DGIST Start-up Fund Program (2021070007).

REFERENCES

[1] Muninder Adavelli. 2023. How many videos are uploaded on Youtube every day? https://techjury.net/blog/how-many-videos-are-uploaded-to-youtube-a-day/
[2] Anu Aggarwal. 2023. Do cartoons positively affect your mental health. https://anuaggarwalfoundation.org/do-cartoons-positively-affect-your-mental-health/
[3] Joe Bagliere. 2020. Science shows watching cute animals is good for your health. *CNN. https://www. cnn. com/2020/09/27/us/watching-cute-animals-studyscn-trnd/index. html* (2020).
[4] Monika Bickert. 2019. Publishing our internal enforcement guidelines and expanding our appeals process. https://about.fb.com/news/2018/04/comprehensive-community-standards/
[5] Branka. 2023. Facebook statistics 2023. https://truelist.co/blog/facebook-statistics/
[6] Nicola W Burton, Ken I Pakenham, and Wendy J Brown. 2010. Feasibility and effectiveness of psychosocial resilience training: a pilot study of the READY

program. *Psychology, health & medicine* 15, 3 (2010), 266–277.
[7] K Canegallo. 2019. Meet the teams keeping our corner of the internet safer. *The Keyword* (2019).
[8] Robert D Caplan. 1983. Person-environment fit: Past, present and future. *Stress research* (1983).
[9] Elinor Carmi. 2019. The hidden listeners: regulating the line from telephone operators to content moderators. *International Journal of Communication* 13 (2019), 440–458.
[10] James Carmody and Ruth A Baer. 2008. Relationships between mindfulness practice and levels of mindfulness, medical and psychological symptoms and well-being in a mindfulness-based stress reduction program. *Journal of behavioral medicine* 31 (2008), 23–33.
[11] Eshwar Chandrasekharan, Chaitrali Gandhi, Matthew Wortley Mustelier, and Eric Gilbert. 2019. Crossmod: A cross-community learning-based system to assist reddit moderators. *Proceedings of the ACM on human-computer interaction* 3, CSCW (2019), 1–30.
[12] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. 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*. 3175–3187.
[13] A Chen. 2012. Facebook releases new content guidelines, now allows bodily fluids. *Gawker* (2012).
[14] Adrian Chen. 2014. The laborers who keep dick pics and beheadings out of your Facebook feed. *Wired* 23 (2014), 14.
[15] Jill Collins, Alison Gibson, Sarah Parkin, Rosemary Parkinson, Diana Shave, and Colin Dyer. 2012. Counselling in the workplace: How time-limited counselling can effect change in well-being. *Counselling and Psychotherapy Research* 12, 2 (2012), 84–92.
[16] Cambridge Consultants. 2019. Use of AI in Online Content Moderation. (2019).
[17] Christine L Cook, Jie Cai, and Donghee Yvette Wohn. 2022. Awe Versus Aww: The Effectiveness of Two Kinds of Positive Emotional Stimulation on Stress Reduction for Online Content Moderators. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–19.
[18] Christine L Cook, Aashka Patel, and Donghee Yvette Wohn. 2021. Commercial versus volunteer: Comparing user perceptions of toxicity and transparency in content moderation across social media platforms. *Frontiers in Human Dynamics* 3 (2021), 626409.
[19] M Csikszentmihalyi. 1990. Flow. The Psychology of Optimal Experience. New York (HarperPerennial) 1990. (1990).
[20] Brandon Dang, Martin J Riedl, and Matthew Lease. 2018. But who protects the moderators? the case of crowdsourced image moderation. *arXiv preprint arXiv:1804.10999* (2018).
[21] Anubrata Das, Brandon Dang, and Matthew Lease. 2020. Fast, accurate, and healthier: Interactive blurring helps moderators reduce exposure to harmful content. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 33–42.
[22] Edward L Deci. 1971. Effects of externally mediated rewards on intrinsic motivation. *Journal of personality and Social Psychology* 18, 1 (1971), 105.
[23] Edward L Deci and Richard M Ryan. 2008. Facilitating optimal motivation and psychological well-being across life's domains. *Canadian psychology/Psychologie canadienne* 49, 1 (2008), 14.
[24] Oscar Deniz, Ismael Serrano, Gloria Bueno, and Tae-Kyun Kim. 2014. Fast violence detection in video. In *2014 international conference on computer vision theory and applications (VISAPP)*, Vol. 2. IEEE, 478–485.
[25] Elizabeth Dwoskin. 2019. Internet gatekeepers pay a psychic toll. *Post Reports, the daily podcast of The Washington Post* (2019).
[26] Jennifer Elias. 2020. Former YouTube content moderator describes horrors of the job in new lawsuit. https://www.cnbc.com/2020/09/22/former-youtube-content-moderator-describes-horrors-of-the-job-in-lawsuit.html
[27] Paul A Frewen and Ruth A Lanius. 2006. Toward a psychobiology of posttraumatic self-dysregulation: Reexperiencing, hyperarousal, dissociation, and emotional numbing. *Annals of the New York Academy of Sciences* 1071, 1 (2006), 110–124.
[28] Abhimanyu Ghoshal. 2017. Microsoft sued by employees who developed ptsd after reviewing disturbing content. *The next web* (2017).
[29] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press.
[30] Joanne E Gray and Nicolas P Suzor. 2020. Playing with machines: Using machine learning to understand automated copyright enforcement at scale. *Big Data & Society* 7, 1 (2020), 2053951720919963.
[31] Michael Hemphill. 1996. A note on adults' color–emotion associations. *The Journal of genetic psychology* 157, 3 (1996), 275–280.
[32] Hope Horner. 2019. The Psychology of Video: Why Video makes people more likely to buy. https://sproutvideo.com/blog/psychology-why-video-makes-people-more-likely-buy.html
[33] Matthew Hotopf and Simon Wessely. 1997. Stress in the workplace: unfinished business. *Journal of Psychosomatic Research* 43, 1 (1997), 1–6.
[34] Patrick G Hunter, E Glenn Schellenberg, and Ulrich Schimmack. 2010. Feelings and perceptions of happiness and sadness induced by music: Similarities, differences, and mixed emotions. *Psychology of Aesthetics, Creativity, and the Arts* 4, 1 (2010), 47.
[35] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.
[36] Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. 2016. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. (2016).
[37] Sowmya Karunakaran and Rashmi Ramakrishan. 2019. Testing stylistic interventions to reduce emotional impact of content moderation workers. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 50–58.
[38] Alethea HQ Koh, Eddie MW Tong, and Alexander YL Yuen. 2019. The buffering effect of awe on negative affect towards lost possessions. *The Journal of Positive Psychology* 14, 2 (2019), 156–165.
[39] Christof Kuhbandner and Reinhard Pekrun. 2013. Joint effects of emotion and color on memory. *Emotion* 13, 3 (2013), 375.
[40] Nancy Kwallek, Carol M Lewis, and Ann S Robbins. 1988. Effects of office interior color on workers' mood and productivity. *Perceptual and Motor Skills* 66, 1 (1988), 123–128.
[41] Vickie A Lambert, Clinton E Lambert, and Hiroaki Yamase. 2003. Psychological hardiness, workplace stress and related stress reduction strategies. *Nursing & Health Sciences* 5, 2 (2003), 181–184.
[42] Anthony D LaMontagne, Angela Martin, Kathryn M Page, Nicola J Reavley, Andrew J Noblet, Allison J Milner, Tessa Keegel, and Peter M Smith. 2014. Workplace mental health: developing an integrated intervention approach. *BMC psychiatry* 14, 1 (2014), 1–11.
[43] Peter J Lang, Margaret M Bradley, Bruce N Cuthbert, et al. 2005. *International affective picture system (IAPS): Affective ratings of pictures and instruction manual*. NIMH, Center for the Study of Emotion & Attention Gainesville, FL.
[44] Jeongwoo Lee, Jihye Lee, Hyungsun Lim, Ji-Seon Son, Jun-Rae Lee, Dong-Chan Kim, and Seonghoon Ko. 2012. Cartoon distraction alleviates anxiety in children during induction of anesthesia. *Anesthesia & Analgesia* 115, 5 (2012), 1168–1173.
[45] Kalev Leetaru. 2018. Why we still need human moderators in an AI-Powered World. https://www.forbes.com/sites/kalevleetaru/2018/09/08/why-we-still-need-human-moderators-in-an-ai-powered-world/?sh=4d05c50f1412
[46] Sara Lindberg. 2022. What is emotional numbing? https://www.verywellmind.com/emotional-numbing-symptoms-2797372
[47] Clare J Mackie, Natalie Castellanos-Ryan, and Patricia J Conrod. 2011. Personality moderates the longitudinal relationship between psychological symptoms and alcohol use in adolescents. *Alcoholism: Clinical and Experimental Research* 35, 4 (2011), 703–716.
[48] Silvia Meyer, Alexander Grob, and Markus Gerber. 2021. No fun, no gain: The stress-buffering effect of physical activity on life satisfaction depends on adolescents' intrinsic motivation. *Psychology of Sport and Exercise* 56 (2021), 102004.
[49] Online Moderation. 2019. https://www.onlinemoderation.com/human-moderators-versus-technology/
[50] Jessica Gall Myrick. 2015. Emotion regulation, procrastination, and watching cat videos online: Who watches Internet cats, why, and to what effect? *Computers in human behavior* 52 (2015), 168–176.
[51] Casey Newton. 2019. The emotional toll of content moderation. https://www.npr.org/2019/12/21/790492548/the-emotional-toll-of-content-moderation
[52] Casey Newton. 2019. The trauma floor: The secret lives of Facebook moderators in America. *The Verge* 25, 02 (2019).
[53] Casey Newton. 2020. Facebook will pay $52 million in settlement with moderators who developed PTSD on the job. https://www.theverge.com/2020/5/12/21255870.
[54] Casey Newton. 2020. YouTube moderators are being forced to sign a statement acknowledging the job can give them PTSD. https://www.theverge.com/2020/1/24/21075830/youtube-moderators-ptsd-accenture-statement-lawsuits-mental-health
[55] Annabelle Nyst. 2023. 134 social media statistics you need to know for 2023. https://www.searchenginejournal.com/social-media-statistics/480507/
[56] Faculty of Biological Sciences. 2020. What are the health benefits of watching cute animals? https://biologicalsciences.leeds.ac.uk/school-biomedical-sciences/news/article/273/what-are-the-health-benefits-of-watching-cute-animals
[57] Se-Hyung Park, So-Young Kim, Jinny Hye-Jin Choo, Won-jae Lee, and Jun-sang Kang. 2009. Using new media to create integrating art therapy: animation therapy. In *ACM SIGGRAPH ASIA 2009 Educators Program*. 1–5.
[58] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep learning for user comment moderation. *arXiv preprint arXiv:1705.09993* (2017).
[59] Emily R. 2023. The Ultimate Guide to Content moderation. https://getstream.io/blog/content-moderation/
[60] Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. (2016).
[61] Sarah T. Roberts. 2017. *Content Moderation*. Springer International Publishing, Cham, 1–4. https://doi.org/10.1007/978-3-319-32001-4_44-1

[62] Melanie Rudd, Kathleen D Vohs, and Jennifer Aaker. 2012. Awe expands people's perception of time, alters decision making, and enhances well-being. *Psychological science* 23, 10 (2012), 1130–1136.

[63] Napa Sae-Bae, Xiaoxi Sun, Husrev T Sencar, and Nasir D Memon. 2014. Towards automatic detection of child pornography. In *2014 IEEE International Conference on Image Processing (ICIP)*. IEEE, 5332–5336.

[64] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.

[65] Joseph Seering, Robert Kraut, and Laura Dabbish. 2017. Shaping pro and anti-social behavior on twitch through moderation and example-setting. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*. 111–125.

[66] Elisabeth Sherman. 2020. Therapists Explain How Cartoons Affect Your Mental Health.

[67] Satyendra Singh. 2006. Impact of color on marketing. *Management decision* 44, 6 (2006), 783–789.

[68] Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[69] Mark Sullivan. 2019. Facebook is expanding its tools to make content moderation less toxic.

[70] Tina M Sutton and Jeanette Altarriba. 2016. Color associations to emotion and emotion-laden words: A collection of norms for stimulus construction and selection. *Behavior research methods* 48 (2016), 686–728.

[71] Tan Tang, Yanhong Wu, Yingcai Wu, Lingyun Yu, and Yuhong Li. 2021. Videomoderator: A risk-aware framework for multimodal video moderation in e-commerce. *IEEE Transactions on Visualization and Computer Graphics* 28, 1 (2021), 846–856.

[72] Meike K Uhrig, Nadine Trautmann, Ulf Baumgärtner, Rolf-Detlef Treede, Florian Henrich, Wolfgang Hiller, and Susanne Marschall. 2016. Emotion elicitation: A comparison of pictures and films. *Frontiers in psychology* 7 (2016), 180.

[73] Magdalena MHE Van den Berg, Jolanda Maas, Rianne Muller, Anoek Braun, Wendy Kaandorp, René Van Lien, Mireille NM Van Poppel, Willem Van Mechelen, and Agnes E Van den Berg. 2015. Autonomic nervous system responses to viewing green and built settings: differentiating between sympathetic and parasympathetic activity. *International journal of environmental research and public health* 12, 12 (2015), 15860–15874.

[74] Bertie Vidgen, Helen Margetts, and Alex Harris. 2019. How much online abuse is there. *Alan Turing Institute* 11 (2019).

[75] Jacqueline C Vischer. 2007. The effects of the physical environment on job performance: towards a theoretical model of workspace stress. *Stress and health: Journal of the International Society for the Investigation of Stress* 23, 3 (2007), 175–184.

[76] Jiaxin Wang, Xiaoxiao Fu, and Youcheng Wang. 2021. Can "bad" stressors spark "good" behaviors in frontline employees? Incorporating motivation and emotion. *International journal of contemporary hospitality management* 33, 1 (2021), 101–124.

[77] Xinrui Wang and Jinze Yu. 2020. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8090–8099.

[78] David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology* 54, 6 (1988), 1063.

[79] Webhelp. 2022. Why are human moderators still essential? https://webhelp.com/news/why-are-human-moderators-still-essential/

[80] Jason Wise. 2023. How many Instagram posts are there in 2023? https://earthweb.com/how-many-instagram-posts-are-there/

[81] Donghee Yvette Wohn. 2019. Volunteer moderators in twitch micro communities: How they get involved, the roles they play, and the emotional labor they experience. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[82] Jing Zeng and D Bondy Valdovinos Kaye. 2022. From content moderation to visibility moderation: A case study of platform governance on TikTok. *Policy & Internet* 14, 1 (2022), 79–95.

[83] Haoti Zhong, Hao Li, Anna Cinzia Squicciarini, Sarah Michele Rajtmajer, Christopher Griffin, David J Miller, and Cornelia Caragea. 2016. Content-Driven Detection of Cyberbullying on the Instagram Social Network. In *IJCAI*, Vol. 16. 3952–3958.

A EXPERIMENT RESULT TABLE

A.1 Statistical analysis of the emotional changes in AMT study

	Baseline Condition		Blurring		Grayscaling		Motivational Text		Motivational Videos		Interleaving with Positive Videos		Cartooni-zation	
	Z	p	Z	p	Z	p	Z	p	Z	p	Z	p	Z	p
Interested	-0.254	0.800	-0.229	0.819	-1.100	0.271	-0.977	0.329	-1.299	0.194	0.000	1.000	-0.182	0.856
Excited	-1.931	0.054	-0.108	0.914	-2.747	0.006*	-2.279	0.023*	-2.337	0.019*	-0.248	0.804	-1.469	0.142
Strong	-1.522	0.128	-0.587	0.557	-1.676	0.094	-2.658	0.008*	-0.319	0.750	-1.732	0.083	-0.468	0.640
Enthusiastic	-2.140	0.032*	-0.532	0.595	-1.628	0.104	-2.479	0.013*	-0.912	0.362	-2.294	0.022*	-1.062	0.288
Proud	-2.414	0.016*	-0.261	0.794	-2.582	0.010*	-3.130	0.002**	-2.118	0.034*	-1.585	0.113	-0.233	0.816
Alert	-2.996	0.003*	-0.778	0.437	-1.289	0.197	-2.802	0.005*	-1.850	0.064	-0.592	0.554	-0.735	0.462
Inspired	-2.516	0.012*	-0.787	0.431	-3.069	0.002**	-1.747	0.081	-1.177	0.239	-0.645	0.519	-0.171	0.864
Determined	-0.567	0.571	-0.198	0.843	-0.637	0.524	-0.168	0.867	-0.099	0.921	-1.127	0.260	-1.691	0.091
Attentive	-1.459	0.145	-0.436	0.663	-1.262	0.207	-0.131	0.896	-2.101	0.036*	-0.791	0.429	-0.873	0.383
Active	-1.137	0.256	-0.907	0.365	-0.593	0.553	-0.807	0.420	-1.779	0.075	-0.690	0.490	-1.189	0.234
Distressed	-3.530	0.000**	-1.653	0.098	-2.723	0.006*	-2.808	0.005*	-1.355	0.176	-0.741	0.458	-0.401	0.688
Upset	-3.358	0.001**	-3.050	0.002**	-2.234	0.026*	-2.723	0.006*	-2.237	0.025*	-0.760	0.447	-1.153	0.249
Guilty	-1.106	0.269	-0.805	0.421	-0.452	0.651	-2.560	0.010*	-2.381	0.017*	-0.811	0.417	-1.494	0.135
Scared	-3.128	0.002**	-3.198	0.001**	-2.862	0.004*	-2.636	0.008*	-2.890	0.004*	-1.700	0.089	-0.597	0.550
Hostile	-0.956	0.339	-1.073	0.283	-0.875	0.382	-2.639	0.008*	-2.097	0.036*	-1.639	0.101	-0.317	0.751
Irritable	-1.564	0.118	-1.634	0.102	-2.828	0.005*	-2.314	0.021*	-1.423	0.155	-2.147	0.032*	-0.075	0.940
Ashamed	-1.921	0.055	-0.493	0.622	-1.328	0.184	-0.453	0.651	-1.793	0.073	-0.832	0.405	-0.746	0.456
Nervous	-2.654	0.008*	-2.843	0.004*	-3.210	0.001**	-3.294	0.001**	-2.653	0.008*	-0.836	0.403	-0.289	0.772
Jittery	-2.869	0.004*	-1.993	0.046*	-1.402	0.161	-3.084	0.002**	-0.324	0.746	-0.673	0.501	-1.380	0.167
Afraid	-2.891	0.004*	-1.998	0.046*	-3.389	0.001**	-3.676	0.000**	-3.173	0.002**	-0.225	0.822	-0.947	0.343

Table 6: Statistical analysis of the emotional changes before and after the moderation task is reported for each intervention technique. Interleaving with positive videos and cartoonization significantly reduced negative emotions compared to other intervention techniques, indicating that it can be emotionally helpful. * indicates $p < 0.05$ and ** indicates $p < 0.0025$, the significance level with Bonferroni correction applied.

A.2 Summary of advantages and disadvantages of each intervention technique

	Advantages	Disadvantages
Baseline Condition	• Ability to distinguish content in tasks. • Accessibility to original material. • Combined advantages of accuracy and speed.	• Possibility of exposure without contrast. • Emotional burden from graphic content. • Stress from stimulating content.
Blurring	• Direct avoidance of shocking videos. • Reduction of shocking experiences and stress. • Psychological assistance for workers. • Reduced perception of bad situations.	• Decreased accuracy. • Challenges in censorship and difficulty in perception. • Difficulty in detecting small movements. • Eye discomfort and increased fatigue.
Grayscaling	• Preventing negative emotions related to blood. • Maintaining information dissemination. • Stress reduction through non-stimulating content. • Reduced eye strain and alleviation of brutality.	• Constraints in tasks related to color. • Emotional strain due to maintaining sharpness. • Increased fatigue due to accuracy reduction. • Discomfort and eye fatigue.
Motivational Text	• Motivation through a sense of sacrifice. • Daily mental preparation. • Achievement through a positive impact. • Opportunities for refreshment.	• Time consumption in reading text. • Boredom and fatigue due to text fixation. • Doubts about the contribution of repetitive text.
Motivational Videos	• Instilling a sense of purpose and responsibility. • Providing mental preparedness. • Positive outlook through visual media. • Opportunities for refreshment.	• Inconvenience due to time consumption. • Tediousness in repetitive content.
Interleaving with Positive Videos	• Opportunities for motivation and relaxation. • Fatigue and stress management through refreshment. • Stress relief and utilization of rest time. • Psychological help for workers.	• Annoyance due to reduced work speed. • Reduced effectiveness for disliked positive videos. • Reduced efficiency due to diminished concentration. • Dissatisfaction due to time issues during work.
Cartoonization	• Emotional support through cartoon processing. • Reduction of negative emotions. • Comfortable viewing through cartoon processing. • Reduced impact of stimulating videos.	• Possible decrease in accuracy. • Moderation burden due to lack of visual information.

Table 7: Summary of advantages and disadvantages of each intervention technique reported by the in-lab within-subjects study participants.

A.3 Changes in participants' mood

		Day 1	Day 2	Day 3
Baseline Condition	P1	Subtle anxiety, guilt, concern for victims during accident video viewing.	Discomfort similar to yesterday and increased sensitivity.	More desensitized today but remained sensitive to sound and felt discomfort.
	P2	Post-accident video exposure elevated heart rate, sense of burden.	Post-experiment elevated heart rate and felt tension.	Increased post-experiment heart rate and slight tension.
	P3	Anxiety, worry while watching accidents, followed by feeling vacant afterward.	Intense tension after viewing large accident video, but not much emotional impact.	Large accident videos still evoke worry and distress, while others felt familiar with no mood change.
Interleaving with Positive Videos	P4	Surprise, racing heart during accident video viewing, discomfort when thinking about victims.	Pre-experiment anxiety, startle responses, and increased heart rate during the experiment.	Felt heaviness, slight sadness, and vigilance. Less startle and emotional numbness on the last day.
	P5	Initial tension, relief due to lack of stimulating content.	Occasional startle from accident sounds, but no significant emotional change.	Only slight tension before the experiment started.
	P6	Diminished mood due to accident videos, improved by cute animal videos.	Initial tension while watching videos, relief from peaceful intermissions.	Felt some anger while watching accident videos but no change after the experiment.
Cartoonization	P7	Stressed more by vehicle passengers' sounds, more like watching animation than real accidents.	Visual stimuli led to fatigue, heightened sensitivity to collision sounds.	Auditory stimuli led to fatigue, but no significant emotional change.
	P8	Overall, experienced a slight increase in fatigue and fear.	Feeling uneasy, thoughts of loved ones.	Watching accident videos left feelings of unease and guilt.
	P9	Increased fatigue and fear from exposure to sound and videos, but no negative emotional changes.	Continued sensitivity and increased emotional impact from stimulating accident videos.	Post-experiment, emotions became calmer with a sense of fear.

Table 8: Changes in participants' mood by day after doing the moderation task of annotating car crash videos. Most participants reported negative physical and psychological effects after watching car crash videos.

B IN-LAB WITHIN-SUBJECTS STUDY

B.1 Questions to be asked after completing each intervention technique

(1) If this intervention technique were available, how likely would you use it?
   Extremely likely
   Highly likely
   Neutral
   Somewhat likely
   Not at all likely
(2) In our experiments, we did not offer the option to switch to the original video. However, if this option were available while using this intervention technique, how frequently would you choose to switch to the original video? (Especially when the video is hard to recognize or the process is complicated)
   Always
   Often
   Sometimes
   Rarely
   Never
(3) If this intervention technique were actually implemented, would it help reduce your negative emotions when dealing with provocative or unpleasant videos?
   Strongly agree
   Somewhat agree
   Neither agree nor disagree
   Somewhat disagree
   Strongly disagree
(4) What advantages can workers experience when using this intervention technique?
(5) What disadvantages can workers experience when using this intervention technique?
(6) Please share your experiences after watching videos processed with this intervention technique.

B.2 Questions to be asked after completing all intervention techniques

(1) If you were a content moderator handling harmful videos, what intervention techniques would you use to protect your emotions? (Please order from the most preferred to the least.)
(2) If you were a content moderator handling harmful videos, what intervention techniques would you use for task accuracy and speed? (Please order from the most preferred to the least.)
(3) What intervention technique impressed you the most? Additionally, what were the reasons for your choice?
(4) What intervention technique did you find least effective? Additionally, what were the reasons for your choice?
(5) When you actually performing tasks, which intervention techniques would you use?
(6) Are there any other intervention techniques that could be helpful?

C THREE-DAY CONTROLLED STUDY

C.1 Day 1

(1) Please write your name.
(2) What is today's date?
(3) Where did you conduct today's experiment?
(4) Before the experiment, please describe your mood in detail.
(5) Did you experience any changes in your mood after performing this experiment today? If so, please explain the reasons.

C.2 Day 2 & 3

(1) Please write your name.
(2) What is today's date?
(3) Where did you conduct today's experiment?
(4) Before the experiment, please describe your mood in detail.
(5) Did yesterday's experiment experience affect your mood today?
(6) Did yesterday's experiment experience affect your daily life today?
(7) Did you experience any changes in your mood after performing this experiment today? If so, please explain the reasons.

C.3 Face-to-face interview questions

(1) Have you ever been unintentionally exposed to harmful content on social media platforms? If so, please describe your experiences.
(2) If this intervention technique were available, how likely would you use it?
   Extremely likely
   Highly likely
   Neutral
   Somewhat likely
   Not at all likely
(3) In our experiments, we did not offer the option to switch to the original video. However, if this option were available while using this intervention technique, how frequently would you choose to switch to the original video? (Especially when the video is hard to recognize or the process is complicated)
   Always
   Often
   Sometimes
   Rarely
   Never
(4) If this intervention technique were actually implemented, would it help reduce your negative emotions when dealing with provocative or unpleasant videos?
   Strongly agree
   Somewhat agree
   Neither agree nor disagree
   Somewhat disagree
   Strongly disagree
(5) What advantages can workers experience when using this intervention technique?
(6) What disadvantages can workers experience when using this intervention technique?

(7) If you were a content moderator, would using this intervention technique help protect your emotions when dealing with provocative or unpleasant videos?
(8) What were the most prominent emotions you felt while watching car crash videos?
(9) Did watching these videos over three days affect your daily life?
(10) Did watching these videos over three days progressively affect your emotions?
(11) If you were exposed to these videos for a longer period than three days, how does it impact your daily life?
(12) How could this intervention technique be improved?
(13) Please share your experiences after watching videos processed with this intervention technique.
(14) Are there any other intervention techniques that could be helpful?

C.4 Final interview questions

(1) Please describe your mood today in detail.
(2) It has been three weeks since this experiment, and you were exposed to continuous car accident videos for three days. Did this experience affect your daily life?
(3) It has been three weeks since this experiment, and you were exposed to continuous car accident videos for three days. Did this experience affect your emotions?
(4) If you were a content moderator, how would you manage stress?
(5) If you were a content moderator, would prolonged exposure to provocative videos affect your daily life?
(6) If you were a content moderator, would prolonged exposure to provocative videos affect your emotions?
(7) Could you actually do this work like content moderators do? If so, why?
(8) If you were a content moderator, would using this intervention technique be helpful for you when exposed to provocative or unpleasant videos? If so, why?