"Gross!": Personalized Video Content Moderation Using Generative Image Overlays for Aversive Videos

Yejin Choi*
Chung-Ang University
Seoul, Republic of Korea
yeyeye222@cau.ac.kr

Dohwa Kim*
Chung-Ang University
Seoul, Republic of Korea
kimdohwa2@cau.ac.kr

Youngeun Jun
Chung-Ang University
Seoul, Republic of Korea
nara085@cau.ac.kr

Hyosu Kim
Chung-Ang University
Seoul, Republic of Korea
hskimhello@cau.ac.kr

Eunji Park†
Chung-Ang University
Seoul, Republic of Korea
eunjipark@cau.ac.kr

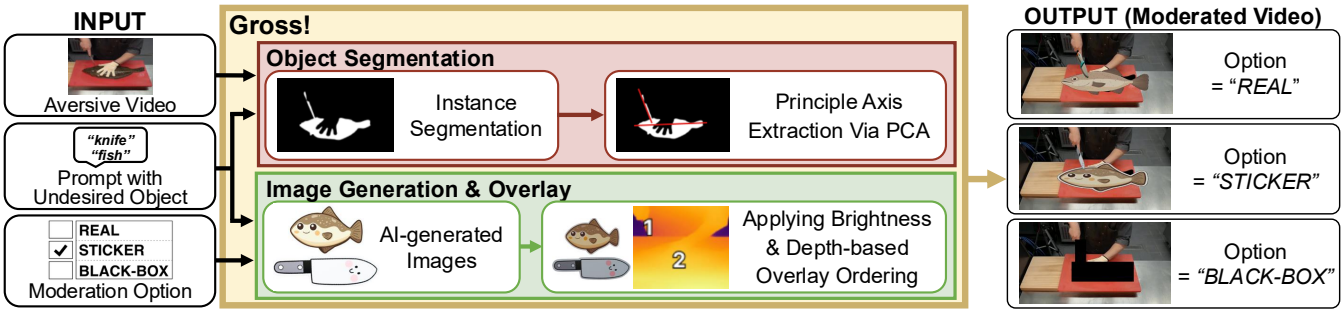

Figure 1: System overview of *Gross!* (Sample frame adapted from YouTube for illustrative purposes [3].)

Abstract

As the number of videos available online continues to increase rapidly, effective content filtering technologies have become essential. However, existing approaches lack personalization and struggle to preserve contextual understanding of the video. In this study, we propose a personalized content filtering method capable of understanding contextual information within videos. To evaluate the effectiveness of our approach, we conducted a user study with 18 participants. The results demonstrated a improvement in discomfort reduction and contextual understanding.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

content moderation, personalization, contextual preserving

ACM Reference Format:

Yejin Choi, Dohwa Kim, Youngeun Jun, Hyosu Kim, and Eunji Park. 2025. "Gross!": Personalized Video Content Moderation Using Generative Image Overlays for Aversive Videos. In *The 38th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '25), September 28–October 01, 2025, Busan, Republic of Korea.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3746058.3758458

*The first two authors contributed equally and are listed alphabetically.
†Corresponding author

UIST Adjunct '25, Busan, Republic of Korea

ACM ISBN 979-8-4007-2036-9/25/09
https://doi.org/10.1145/3746058.3758458

1 Introduction

As digital environments continue to generate more varied video content, there is a growing need for effective filtering techniques to protect viewers from harmful or sensitive materials [1, 6, 8, 9, 16]. Despite this growing need, previous research has been limited to focusing only on how to effectively obscure undesirable objects, rather than tailoring to users' individual aversions [5, 14]. Even major video platforms apply uniform filtering policies to all users, without personalization [4, 12]. This can be particularly distressing for people with specific phobias (e.g., acrophobia, clownphobia), who may experience greater discomfort [6].

In addition, conventional filtering methods can hinder users' understanding of video context, as key visual objects are obscured or removed [7]. For example, traditional filtering methods (e.g., black box [13], pixelation [15], blurring [7, 14]) make it difficult for viewers to understand the context. To address this issue, we propose *Gross!*, a personalized filtering system that preserves contextual information for users. *Gross!* moderates aversive content by receiving user-defined aversions as text prompts and masking the corresponding objects in each video frame with AI-generated images. Through a user study with 18 participants, we observed improvements in both reducing discomfort and preserving context than the conventional method.

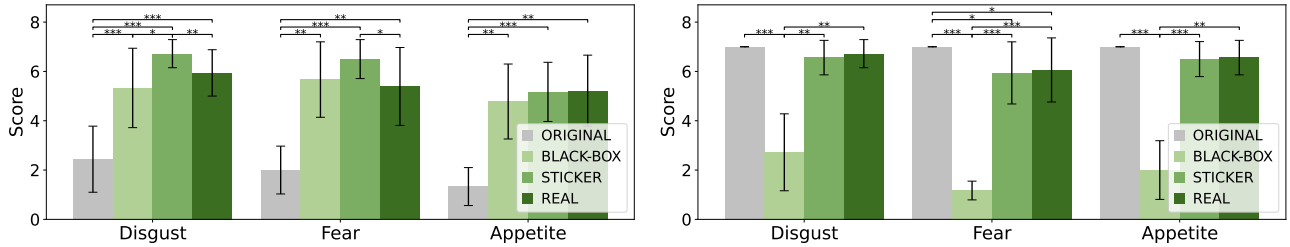

Figure 2: (Left) Content Moderation (Right) Context Preservation Effectiveness across Moderation Strategies (* = $p < .001$; ** = $p < .01$; * = $p < .05$)**

2 Implementation

2.1 Image Segmentation for Object Detection

First, we applied instance segmentation to every frame of the target video to accurately identify the location of unpleasant objects. We adopted the Grounded-SAM2 model [11], a text-prompt-based object segmentation model. Based on the user's text input specifying undesired objects (e.g., "knife", "fish"), the system generates precise segmentation masks for each instance. These masks define the exact regions to be filtered. Additionally, we apply Principal Component Analysis (PCA) to each mask to estimate object orientation, which guides the placement of generated replacement images.

2.2 Content Moderating Image Generation

We implemented one baseline (i.e., `BLACK-BOX`) and two moderation strategies (i.e., `STICKER` and `REAL`): (1) `BLACK-BOX`: a conventional black-box overlay for strict concealment, (2) `STICKER`: a playful sticker overlay designed to reduce feelings of disgust , and (3) `REAL`: a realistic image overlay that resembles the original object but repulsive features are removed. We used DALL·E 3 [2] to generate replacement images from users' text input. The orientation of the generated image was calculated using PCA to align it with the orientation of the detected object in the scene. Lastly, we adjusted the brightness of the generated image to improve visual consistency with the original scene.

2.3 Image Overlay Based on Estimated Depth

To seamlessly overlay the generated images onto the scene, it is necessary to estimate depth of the images. For example, as shown in Figure 1, when both a knife and a fish are generated, the knife should be overlaid on top of the fish. We used the MiDaS [10] model, which computes relative depth from a single image. Then, the generated images were overlaid on the original image based on the detected location of the objects, calculated orientations, and estimated depth information.

3 Pilot Study

3.1 Task Design

In the experiment, we provided the following 3 scenarios: (1) Disgust-inducing (7 seconds): A video clip showing a live fish being cut with visible flapping movements intended to elicit a strong sense of disgust, (2) Fear-inducing (8 seconds): A video clip containing a clown from a horror movie, and (3) Appetite-inducing (6 seconds): A video of a YouTuber eating fried chicken, commonly associated with stimulating appetite. Each video was processed by *Gross!* to produce three versions with different overlays (i.e., `BLACK-BOX`, `STICKER`, and `REAL`). For every processed videos and `ORIGINAL` video, we asked participants to evaluate (1) Context Moderation Effectiveness: the effectiveness of the content moderation (e.g., reduction in inducing disgust, fear, or appetite), and (2) Context Preservation Effectiveness: the degree of contextual understanding on a 7-point Likert scale after viewing each video. A higher score indicated a greater reduction in disgust, fear, or appetite, and a higher level of contextual understanding.

	Content Moderation Effectiveness			Context Preservation Effectiveness		
Version	Disgust	Fear	Appetite	Disgust	Fear	Appetite
ORIGINAL	2.44 (1.34)	2.00 (0.97)	1.33 (0.77)	7.00 (0.00)	7.00 (0.00)	7.00 (0.00)
BLACK-BOX	5.33 (1.61)	5.67 (1.53)	4.78 (1.52)	2.72 (1.56)	1.17 (0.38)	2.00 (1.19)
STICKER	6.72 (0.57)*	6.50 (0.79)	5.17 (1.20)	6.56 (0.70)**	5.94 (1.26)***	6.50 (0.71)***
REAL	5.94 (0.94)	5.39 (1.58)	5.22 (1.44)	6.72 (0.57)**	6.06 (1.30)***	6.56 (0.70)**

Table 1: Content Moderation and Context Preservation Effectiveness scores of proposed moderation strategies (`STICKER`, `REAL`) compared to `BLACK-BOX` (* = $p < .001$; ** = $p < .01$; * = $p < .05$)**

3.2 Procedure

We recruited a total of 18 participants (7 females and 11 males) from the local university. The average age was 28 years (SD = 13.67). In the experiment, we repeated the process of showing each video and conducting questionnaires evaluating the effectiveness of content moderation and context preservation. For each scenario, we provided `BLACK-BOX` option first to prevent exposure to the covered object by the black box. After that, the remaining versions (i.e., `STICKER` and `REAL`) were presented in a counter-balanced order to mitigate order effects. After the experiment, we conducted brief interviews with the participants.

3.3 Results

We conducted a Friedman test and pairwise Wilcoxon signed-rank tests to compare the effectiveness of context moderation and preservation between the proposed strategies (See Table 1). Among the three strategies, `STICKER` show the highest effectiveness for the disgust- and fear-inducing videos, while `REAL` outperformed others in the appetite-inducing video (See Figure 2). In terms of contextual understanding, both `STICKER` and `REAL` showed significantly higher comprehension scores than the `BLACK-BOX`. Notably, the two

strategies that received high ratings for moderation effectiveness were also found to preserve context relatively well.

During interviews, P5 mentioned that they preferred not to see aversive objects but still wished to understand the movie's content clearly. Interestingly, some participants pointed out that sound played a significant role in eliciting feelings of disgust, fear, or appetite. Moreover, P2, P6, and P8 noted that `BLACK-BOX` was perceived as even more disgusting or frightening than `ORIGINAL`, as it stimulated imagination. However, P2 pointed out that the high realism of `REAL` amplified the fear response, rather than alleviating it, under fear-inducing conditions. This finding indicates that too highly realistic sticker-style visual elements can inadvertently elicit severe aversive responses.

4 Conclusion

We propose *Gross!*, automated user-specific content moderation system that allows fine-grained filtering and preserves contextual information for the user. Our system outperformed conventional filtering methods in both content moderation and context awareness. *Gross!* demonstrates strong potential as a fully personalized filtering algorithm for video platforms.

References

[1] Ava Bartolome and Shuo Niu. 2023. A literature review of video-sharing platform research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2, 3 (2023), 8.

[3] Bocheol Jeong. 2020. [Cooking Tutorial] How to Fillet Flatfish (Paralichthys) — A Detailed, Beginner-Friendly Guide. https://www.youtube.com/watch?v=dADWqQWDpfA&t=651s.

[4] Victoria ME Bridgland, Benjamin W Bellet, and Melanie KT Takarangi. 2023. Curiosity disturbed the cat: Instagram's sensitive-content screens do not deter vulnerable users from viewing distressing content. *Clinical psychological science* 11, 2 (2023), 290–307.

[5] Ryuhaerang Choi, Subin Park, Sujin Han, and Sung-Ju Lee. 2024. FoodCensor: Promoting Mindful Digital Food Content Consumption for People with Eating Disorders. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[6] Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–33.

[7] 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.

[8] Haonan Nan, Zixiao Wang, Yu Zhang, Xuantao Zhang, and Saixing Zeng. 2025. ResponSight: Explainable and Collaborative Moderation Approach for Responsible Video Content in UGC Platform. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–12.

[9] Miran Park, Kyuri Park, Hyewon Cho, Hwan Choi, and Hajin Lim. 2024. Exploring Design Approaches for Reducing Viewers' Discomfort with Distressing Short-form Videos. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[10] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.

[11] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).

[12] Becca Ricks and Jesse McCrosky. 2022. Does This Button Work? Investigating YouTube's Ineffective User Controls. *Mozilla Foundation. Online: https://foundation. mozilla. org/en/research/library/user-controls/report* (2022).

[13] Shagan Sah, Ameya Shringi, Raymond Ptucha, Aaron Burry, and Robert Loce. 2017. Video redaction: a survey and comparison of enabling technologies. *Journal of Electronic Imaging* 26, 5 (2017), 051406–051406.

[14] Ioannis Sarridis, Jochen Spangenberg, Olga Papadopoulou, and Symeon Papadopoulos. 2025. Mitigating viewer impact from disturbing imagery using ai filters: A user-study. *International Journal of Human–Computer Interaction* 41, 2 (2025), 1234–1245.

[15] Kohske Takahshi and Katsumi Watanabe. 2015. Effects of image blur on visual perception and affective response. In *2015 7th International Conference on Knowledge and Smart Technology (KST)*. IEEE, 169–172.

[16] Wenxin Zhao, Fangyu Yu, Peng Zhang, Hansu Gu, Lin Wang, Siyuan Qiao, Tun Lu, and Ning Gu. 2025. YouthCare: Building a Personalized Collaborative Video Censorship Tool to Support Parent-Child Joint Media Engagement. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–20.