Poster: Moderating Illicit Online Image Promotion for Unsafe User-Generated Content Games Using Large Vision-Language Models

Keyan Guo
University at Buffalo
keyanguo@buffalo.edu

Nishant Vishwamitra
The University of Texas at San Antonio
nishant.vishwamitra@utsa.edu

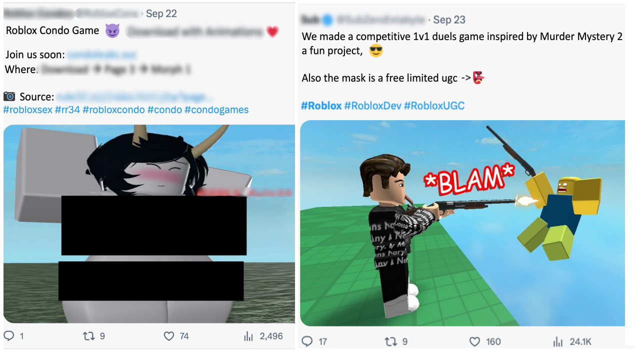
Guo Freeman
Clemson University
guof@clemson.edu

Hongxin Hu
University at Buffalo
hongxinh@buffalo.edu

I. Motivation

Online user-generated content (UGC) has steadily shifted into the limelight, captivating a widespread audience. The sphere of gaming, in particular, has experienced a transformative impact. Gaming platforms like Roblox have established revenue-sharing models with these UGC creators. This collaborative approach has attracted a multitude of UGC creators to build their UGC games (UGCGs) and, as a result, has attracted a large number of users, especially children and adolescents. Data from December 2022 illustrates that 60% of its user base is under 16 years old, with a substantial 45% comprising children under 13 years old [1]. To attract users, the creators advertise their UGCGs leveraging online social media platforms, such as X, Reddit, and Discord [2], [3]. However, the surge in user participation has also attracted individuals with malicious intentions, who have proliferated various harmful games with unsafe content, especially sexually explicit imagery and violence. These games present an unprecedented safety issue to underage users who, often, are ill-prepared to confront or manage such exposures. Exposure to explicit content and interactions violates ethical norms and poses significant challenges to their psychological, emotional, and social development. While moderation during UGCG play is a subject of discussion, alarmingly little effort has been made in moderating the *image-based illicit promotion of such UGCGs by malicious creators on social media platforms*, who are resorting to platforms to promote their games. As depicted in Figure 1, the creators share promotional unsafe images of UGCGs to draw many young players to their harmful designs.

Currently, various existing tools, such as Google Cloud Vision API [4], utilize artificial intelligence and machine learning (AI/ML) models for moderating harmful content. While such tools have demonstrated considerable efficacy in identifying traditional unsafe images (*i.e.*, real-world sexually explicit and violent images), they exhibit diminished efficiency when tasked with detecting unsafe images that are used for illicit online promotions of UGCGs. There exist two key problems in flagging such images. *First*, a paramount problem stems from the requirement of extensive training data intrinsic to traditional machine learning models. These models are adept at identifying and classifying conventional unsafe content, such as sexually explicit and violence, through bolstering by large, annotated datasets. However, the acquisition of such large-scale data becomes a formidable task in the context of UGCGs, due to the ambiguous (*i.e.*, undefined) nature of content within these virtual worlds, characterized by an eclectic mix of artificially rendered avatars and abstract geometrical representations. For example, in Fig. 1 (a), the avatar is a mix of a female-like character with animal-like horns. *Second*, unlike traditional unsafe images, the UGCG images exhibit a substantial shift in the input domain. Traditional AI/ML-based systems are adept at detecting explicit content featuring real human forms. However, UGCGs introduce a complex landscape with a transition from real to artificial. The rendered avatars or personas in UGCGs embody diverse forms and contexts, making their classification a complex endeavor.



Sep 22
Roblox Condo Game 😈 ❤️

Join us soon:
Where:

📷 Source:
#robloxsex #rr34 #robloxcondo #condo #condogames

[redacted]

[redacted]

1 9 74 2,496

Sep 23
We made a competitive 1v1 duels game inspired by Murder Mystery 2 a fun project, 😎

Also the mask is a free limited ugc ->👺

#Roblox #RobloxDev #RobloxUGC

17 9 160 24.1K

(a) Sexually explicit UGCG

(b) Violent UGCG

Fig. 1: Illicit promotions of unsafe UGCGs on X.

II. Dataset and Measurement

We first compile a novel, real-world dataset consisting of 2,924 UGCG images used for unsafe UGCG promotions by the actual game creators on the social media platform X. We collect these images based on keywords derived from self-reported UGCG-related stories shared by parents and children on Common Sense Media. We further measure the performance of five existing unsafe image detection systems, Clarify [5], Yahoo Open NSFW [6], Amazon Rekognition [7], Microsoft Azure [8], and Google Vision AI [4], against these illicit promotional images of UGCGs. Although these systems have effectively detected unsafe images of real-world and cartoon styles, they exhibit significant shortcomings when applied to UGCG images, with detection accuracy scores ranging from a low of 13% (Clarify) to a high of 67% (Google Vision AI). These findings underscore the urgent necessity for enhanced detection mechanisms for flagging image-based online illicit promotions of UGCGs on social media platforms.

III. Our Approach

In our work, we design UGCG-Guard, a novel system for flagging images used for the illicit promotion of unsafe UGCGs. UGCG-Guard consists of four main components: (1) Data Collection and Annotation; (2) UGCG-CoT Prompting; (3) VLM-based Detection; and (4) Content Moderation. The framework begins by compiling a dataset of illicit online promotional images for UGCGs. Following this, we develop UGCG-CoT prompts, a novel Chain-of-Thought (CoT) reasoning-based prompting strategy tailored to enable reasoning-based decision-making for the identification of images used for the illicit promotion of unsafe UGCGs, by addressing the challenges of domain shift and contextual reasoning posed by these images via conditional prompting and reasoning-based prompting. We design conditional prompting for domain adaptation to instruct UGCG-Guard to understand the specific characteristics of UGCG images in a zero-shot manner. More specifically, our designed prompts concentrate exclusively on images similar to those found in UGCGs. This strategy also helps minimize the influence of irrelevant training data on the model. This process is achieved through a two-step structure consisting of a condition and a guidance question. The condition articulated as $Condition$: *"This is an image generated from a role-playing game."* anchors the model's understanding, clarifying that the image in question is a simulation, not a real-world photograph. This foundational insight is pivotal in ensuring the model's responses are contextually anchored. Complementing this, the guidance question Q_1: *"Are there any characters or avatars in this image?"* directs the model's attention to identify human-like figures within the UGCG images. Having equipped large VLMs with the capability to interpret the distinct domain of UGCGs, we proceed to introduce reasoning-based prompts tailored to identify specific unsafe content categories, including sexually explicit material and violence. Such prompts enable the VLM to make contextual decisions and ensure that unsafe content can be identified and moderated. We pose the following question prompts to identify sexually explicit content: Q_{2A}: *"Are the characters or avatars naked?"*; Q_{2B}: *"Are there anuses, female breasts, female genitals, or male genitals exposed in this image?"*; Q_{2C}: *"Are the characters or avatars indulging in any sexual activities?"* To detect violent content, we incorporated these questions: Q_{3A}: *"Are there any knives, guns, or deadly weapons presented in this image?"*; Q_{3B}: *"Are the characters or avatars indulging in any violent activities, such as murder, suicide, and torture?"* We further exploit the analytical prowess of large VLMs with an additional question to uncover a broader spectrum of harmful activities: Q_4: *"Are the characters or avatars indulging in any other harmful activities?"* In the next stage, a large VLM is leveraged to run the prepared UGCG-CoT prompts on a post with illicit promotional images, and the output from the VLM is parsed for answers to each of the UGCG-CoT prompts. Finally, in our concluding stage, we use the parsed answer from the previous stage to determine whether the post contains illicit promotional images of unsafe UGCGs, and if yes, the post is flagged for moderation.

Defenses	**Accuracy**	**Precision**	**Recall**	**F1**
Google Vision AI	0.68	0.79	0.68	0.65
NSFW-CNN	0.5	0.63	0.57	0.47
InstructBLIP	0.6	0.88	0.42	0.57
UGCG-Guard	0.94	0.98	0.91	0.94

TABLE I: Comparing UGCG-Guard against the baselines.

IV. Evaluation

We evaluated the efficacy of UGCG-Guard in identifying image-based illicit promotion of unsafe UGCGs, comparing its performance against existing baseline detectors for unsafe images. We employed Google Vision AI [4] as a representative of cutting-edge commercial systems for unsafe image detection. Additionally, we considered the NSFW-CNN [9], which is fine-tuned on an extensive dataset of unsafe images to enhance its detection capabilities. Also, to understand the effectiveness of UGCG-CoT, we contrasted its performance with the same large VLM, InstructBLIP [10], which was probed using a general zero-shot prompting strategy, and our general prompt is crafted by aligning with the widely accepted definition of unsafe content: "*Is this image unsafe? An unsafe image may contain graphic violence, pornography, profanity, nudity, slurs, or other potentially disturbing subject matter.*" Our results in TABLE I show that UGCG-Guard achieves a state-of-the-art accuracy of 94%, surpassing other baseline detectors in identifying image-based illicit UGCG promotions. Additionally, our UGCG-CoT prompting strategy significantly outperforms the generalized prompt with an improvement of 64.9%. Moreover, we experimented on 322 *unlabeled* samples collected from two other social media platforms, Reddit and Discord, to simulate an "in-the-wild" running scenario that leverages our approach to control the real-world image-based illicit promotion of UGCGs. In this experiment, our framework successfully identifies and flags image-based illicit promotions of UGCGs, achieving an impressive average F1 score of 0.91.

References

[1] Statista. Distribution of Roblox audiences worldwide as of December 2022, by age group, 2023.

[2] Roblox Developer. How to market your game?, 2022.

[3] Youtube. 3 FREE WAYS to PROMOTE your Roblox GAME and make it POPULAR, 2023.

[4] Google. Google Vision AI. https://cloud.google.com/vision/.

[5] Clarifai. Clarifai. https://www.clarifai.com/.

[6] Yahoo. Yahoo Open NSFW. https://github.com/yahoo/open_nsfw, 2016.

[7] Amazon. Amazon Rekognition. https://aws.amazon.com/rekognition/.

[8] Microsoft. Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/.

[9] Gant Laborde. Deep NN for NSFW Detection.

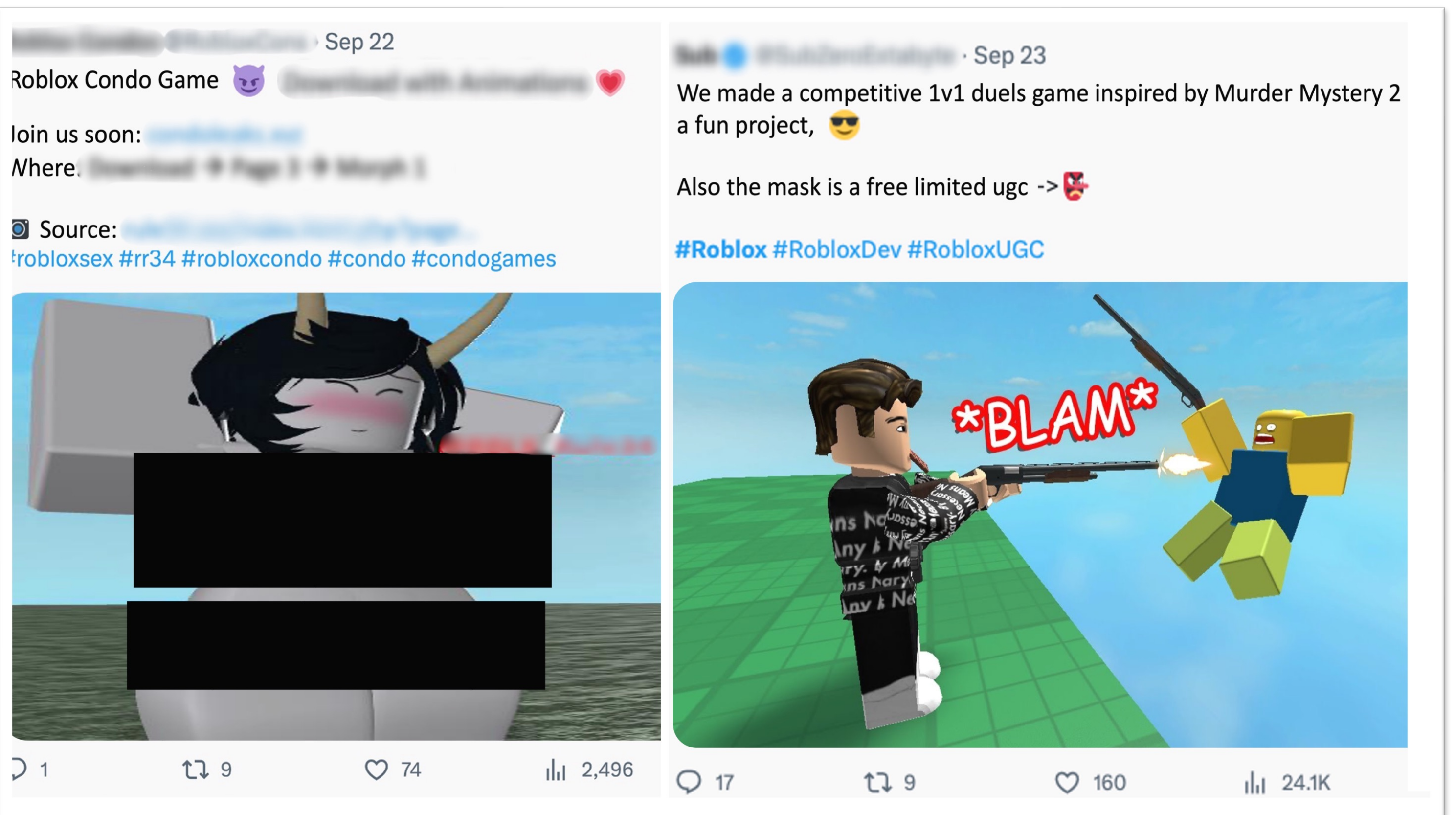
[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning, 2023.

Moderating Illicit Online Image Promotion for Unsafe User-Generated Content Games Using Large Vision-Language Models

Keyan Guo[1], Nishant Vishwamitra[2], Guo Freeman[3], Hongxin Hu[1]
[1]University at Buffalo, [2]The University of Texas at San Antonio, [3]Clemson University

Introduction

Online user-generated content games (UGCGs) are increasingly popular among children and adolescents. However, they pose a heightened risk of exposure to explicit content, raising growing concerns for the online safety of children and adolescents. In our study, we identified an emerging threat of illegally promoting unsafe UGCG based on images on social media, which may inadvertently attract younger users.



[redacted] · Sep 22
Roblox Condo Game 😈 [redacted] ❤️

Join us soon: [redacted]
Where: [redacted]

Source: [redacted]
#robloxsex #rr34 #robloxcondo #condo #condogames

1 9 74 2,496

[redacted] · Sep 23
We made a competitive 1v1 duels game inspired by Murder Mystery 2 a fun project, 😎

Also the mask is a free limited ugc ->

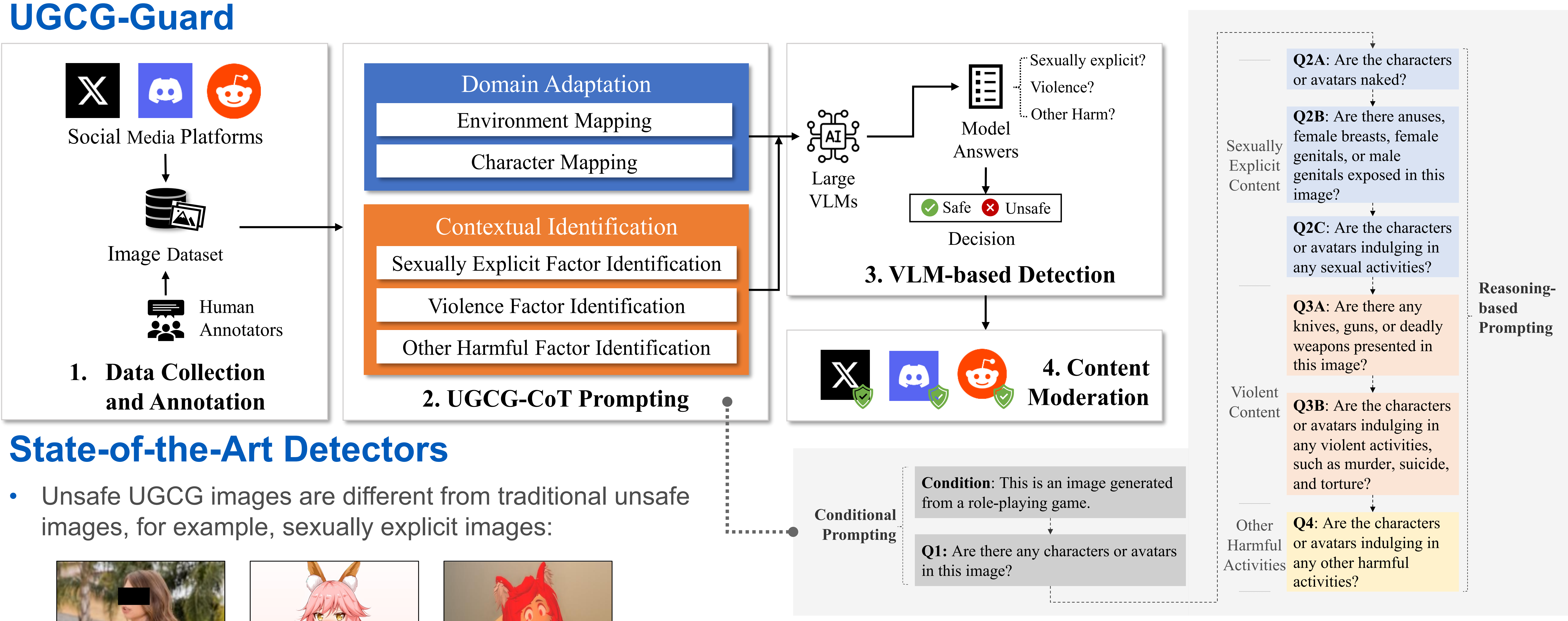
#Roblox #RobloxDev #RobloxUGC

17 9 160 24.1K

(a) Sexually explicit UGCG (b) Violent UGCG

Our studies show a new understanding of this problem and the need for automatically flagging illicit UGCG promotions. We create a novel system, UGCG-Guard to aid social media platforms in effectively addressing this new challenge.

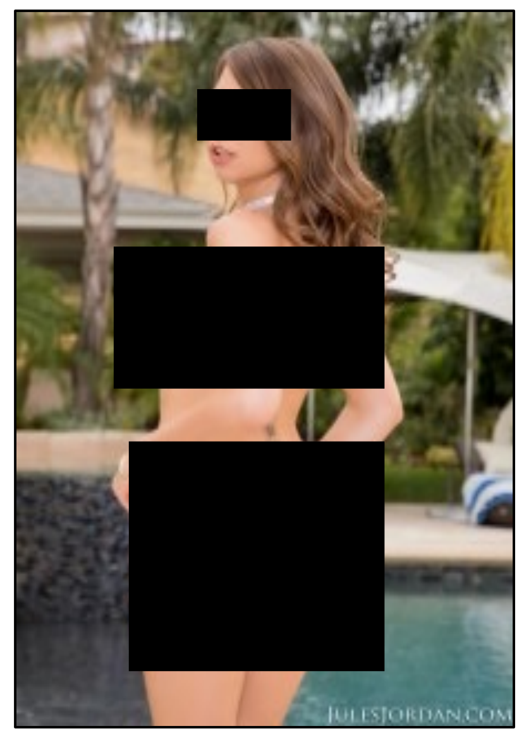
UGCG-Guard



Social Media Platforms
Image Dataset
Human Annotators
1. Data Collection and Annotation
Domain Adaptation
Environment Mapping
Character Mapping
Contextual Identification
Sexually Explicit Factor Identification
Violence Factor Identification
Other Harmful Factor Identification
2. UGCG-CoT Prompting
Large VLMs
Model Answers
Sexually explicit?
Violence?
Other Harm?
Safe
Unsafe
Decision
3. VLM-based Detection
4. Content Moderation
Conditional Prompting
Condition: This is an image generated from a role-playing game.
Q1: Are there any characters or avatars in this image?
Sexually Explicit Content
Q2A: Are the characters or avatars naked?
Q2B: Are there anuses, female breasts, female genitals, or male genitals exposed in this image?
Q2C: Are the characters or avatars indulging in any sexual activities?
Violent Content
Q3A: Are there any knives, guns, or deadly weapons presented in this image?
Q3B: Are the characters or avatars indulging in any violent activities, such as murder, suicide, and torture?
Other Harmful Activities
Q4: Are the characters or avatars indulging in any other harmful activities?
Reasoning-based Prompting

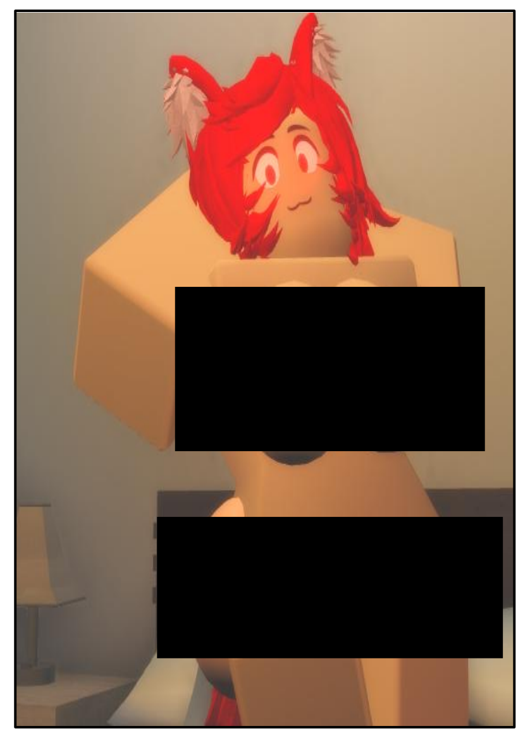
State-of-the-Art Detectors

- Unsafe UGCG images are different from traditional unsafe images, for example, sexually explicit images:



Sexually-explicit-human

Sexually-explicit-anime



Sexually-explicit-UGCG

- We evaluate five SOTA detectors that are widely used to detect unsafe images for different sexually explicit images.

Image Type	State-of-the-Art Unsafe Image Detectors				
	Clarify	Yahoo Open NSFW	Amazon Re-kognition	Microsoft Azure	Google Vision AI
Human	88%	92%	98%	92%	98%
Anime	89%	81%	91%	90%	99%
UGCG	13%	13%	17%	15%	67%

Evaluation

Against Baselines

Defenses	Accuracy	Precision	Recall	F1
Google Vision AI	0.68	0.79	0.68	0.65
NSFW-CNN	0.5	0.63	0.57	0.47
InstructBLIP	0.6	0.88	0.42	0.57
UGCG-Guard	0.94	0.98	0.91	0.94

Detection Rate of Prompt Ablations

- UGCG-Guard without Conditional Prompts: 74%
- Conditional Prompts +
  - Q2A: 86.5% | Q2B: 56.6 | Q2C: 40%
  - Q2A&Q2B: 74.6% | Q2A+Q2C: 93.3% | Q2B+Q2C: 73.3%
  - Q2A & Q2B & Q2C: **98.2%**

An extended version of this poster has been accepted to USENIX Security 2024