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

Keyan Guo, Ayush Utkarsh, Wenbo Ding, and Isabelle Ondracek, *University at Buffalo;* Ziming Zhao, *Northeastern University;* Guo Freeman, *Clemson University;* Nishant Vishwamitra, *The University of Texas at San Antonio;* Hongxin Hu, *University at Buffalo*

https://www.usenix.org/conference/usenixsecurity24/presentation/guo-keyan

This paper is included in the Proceedings of the 33rd USENIX Security Symposium.

August 14–16, 2024 • Philadelphia, PA, USA

978-1-939133-44-1

Open access to the Proceedings of the 33rd USENIX Security Symposium is sponsored by USENIX.

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

Keyan Guo*, Ayush Utkarsh*, Wenbo Ding*, Isabelle Ondracek*,
Ziming Zhao◇, Guo Freeman†, Nishant Vishwamitra‡, Hongxin Hu*
*University at Buffalo, ◇Northeastern University
†Clemson University, ‡The University of Texas at San Antonio,

Abstract

Online user generated content games (UGCGs) are increasingly popular among children and adolescents for social interaction and more creative online entertainment. However, they pose a heightened risk of exposure to explicit content, raising growing concerns for the online safety of children and adolescents. Despite these concerns, few studies have addressed the issue of illicit image-based promotions of unsafe UGCGs on social media, which can inadvertently attract young users. This challenge arises from the difficulty of obtaining comprehensive training data for UGCG images and the unique nature of these images, which differ from traditional unsafe content. In this work, we take the first step towards studying the threat of illicit promotions of unsafe UGCGs. We collect a real-world dataset comprising 2,924 images that display diverse sexually explicit and violent content used to promote UGCGs by their game creators. Our in-depth studies reveal a new understanding of this problem and the urgent need for automatically flagging illicit UGCG promotions. We additionally create a cutting-edge system, UGCG-GUARD, designed to aid social media platforms in effectively identifying images used for illicit UGCG promotions. This system leverages recently introduced large vision-language models (VLMs) and employs a novel conditional prompting strategy for zero-shot domain adaptation, along with chain-of-thought (CoT) reasoning for contextual identification. UGCG-GUARD achieves outstanding results, with an accuracy rate of 94% in detecting these images used for the illicit promotion of such games in real-world scenarios.

Disclaimer. This manuscript contains discussions and visual representations of sexually explicit and violent content. Reader discretion is strongly advised.

1 Introduction

In recent years, 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 [1]. Gaming platforms, such as Roblox, have established revenue-sharing models with these UGC creators [2]. 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 who are under 13 years old [3]. To attract users, the creators advertise their UGCGs leveraging online social media platforms [4–8], such as X [9] (formerly Twitter), Reddit [10], and Discord [11]. In particular, X, as a platform that brings together a large number of UGC creators and gamers, is often utilized as the first choice [4]. 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 [2, 12–16]. These games present an unprecedented safety issue to underage users who, often, are ill-prepared to confront or manage such exposures [12, 17]. The exposure to explicit content and interactions violates not only ethical norms but also poses significant challenges to their psychological, emotional, and social development [2, 15, 18].

While moderation during UGCG play is a subject of discussion [19–22], 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 like X to promote their games. As depicted in Figure 1, the creators share promotional unsafe images of UGCGs to draw a large number of young players to their harmful designs.

Currently, various existing tools, such as Google Cloud Vision API [23], Clarifai [24], and Amazon Rekognition [25], utilize artificial intelligence and machine learning (AI/ML) models for moderating harmful content [26]. However, there is a concern regarding the effectiveness of these tools in preventing the illicit promotion of unsafe UGCG images. While AI/ML-based systems, such as these detectors, have

(a) Sexually explicit UGCG (b) Violent UGCG

Figure 1: Illicit promotions of unsafe UGCGs on X.

demonstrated considerable efficacy in identifying traditional unsafe images (*i.e.*, real-world sexually explicit and violent images) [23–25, 27, 28], these systems 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 Figure 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 where there is a transition from real to artificial. The rendered avatars or personas in UGCGs embody a diverse array of forms and contexts, making their classification a complex endeavor. While such images are challenging for AI/ML, humans can easily perceive these images due to their contextual knowledge.

In this work, we take the first step towards studying the critical problem of image-based online illicit promotions of UGCGs. We first compile a real-world dataset of images collected from X. This dataset comprises a wide range of illicit online promotional images associated with UGCGs, shared by actual game creators. We collect these images based on keywords derived from a textual analysis of self-reported experiences shared by parents and children on Common Sense Media [29] about their experiences with UGCGs. We then conduct a study to explore the characteristics of these image-based illicit online promotions of UGCGs, discovering that the majority of these promotional images are screenshots taken from UGCGs. We further measure the performance of existing unsafe image detection systems against these illicit promotional images of UGCGs. The low success rate indicates that existing systems are severely limited in addressing the challenges presented by these images. Our findings underscore the urgent necessity for enhanced detection mechanisms for flagging image-based online illicit promotions of UGCGs on social media platforms.

Based on our findings, we design UGCG-GUARD [1], a novel system for flagging images used for the illicit promotion of unsafe UGCGs. UGCG-GUARD leverages recently introduced advancements of large vision-language models (VLMs) to detect these images, based on a novel conditional prompting strategy designed for zero-shot domain adaptation, which ensures that the model is attuned to the distinct and nuanced characteristics inherent in UGCG's image content, facilitating the flagging of these images without the need of a large dataset, and a chain-of-thought (CoT) reasoning mechanism [30] for contextual identification of the activities of the personas in these images, enabling UGCG-GUARD to discern and respond to the intricate patterns that define illicit promotional images of unsafe UGCGs. Our system achieves a state-of-the-art average accuracy of 94% in flagging such content.

The key contributions of this paper are as follows:

- **New dataset.** We compile a novel, comprehensive dataset consisting of 2,924 images used for unsafe UGCG promotions by the actual game creators on the social media platform X . These images were systematically gathered over the period from the beginning of 2020 to the end of 2022, serving as a valuable resource for analyzing the visual promotional strategies for UGCGs, and enhancing our understanding of how unsafe UGCGs are promoted on such platforms. Our dataset will be publicly available for verified researchers to facilitate future research in this area.

- **New understanding of unsafe UGCGs and their illicit promotions.** We conduct a study to understand the challenges presented by the illicit online image promotion of unsafe UGCGs. We find such promotions utilized inappropriate images, often screenshots taken from UGCGs, for their promotional purposes. Our measurement analysis further reveals the severe limitation of existing unsafe image detection systems in identifying unsafe UGCG images, underscoring the urgent need for new moderation approaches to counteract illicit UGCG image promotions.

- **New framework for the moderation of image-based illicit online promotion for unsafe UGCGs.** We introduce UGCG-GUARD, a state-of-the-art framework

[1] Our code and datasets are available at https://github.com/CactiLab/UGCG-Guard

to flag image-based illicit promotions of unsafe UGCGs. Rooted in a novel conditional prompting strategy and CoT reasoning approach, UGCG-GUARD effectively leverages large VLMs to achieve zero-shot adaption and contextual detection. UGCG-GUARD can efficiently distinguish content indicative of image-based illicit promotions of UGCGs, even without prior explicit training on similar content categories.

- **Extensive evaluation of UGCG-GUARD.** Our system's evaluation demonstrates its state-of-the-art average accuracy of 94%, surpassing existing baseline detectors of unsafe images by 23.7% to 77.7% in flagging image-based illicit UGCG promotions. Our experiment also shows that our prompting strategy is significantly efficient, outperforming generalized prompting with an improvement of 64.9%. In real-world scenarios, our framework successfully identifies and flags image-based illicit promotions of UGCGs, achieving an impressive average F1 score of 0.91.

2 Background and Related Work

2.1 Online UGCGs and Unsafe Content

The advent of the internet and social media spurred an interactive cultural evolution characterized by UGC posts online, including images, videos, text, and audio. UGC has significantly influenced gaming, promoting creativity, prolonging game lifespans, and fortifying communities [1]. UGCGs like "Minecraft" and "Roblox" epitomize this shift, transforming players into creators and expanding gameplay possibilities exponentially. Now, such UGCGs are wildly presented in today's game platforms for social interaction and gameplay. They allow players to create their own UGC and share it with others [1, 31, 32].

However, as the volume of UGCGs proliferates online, there is a corresponding surge in malicious content and activities within these games. Online communities in UGCGs are thriving spaces where participants interact through their avatars. While avatars can be vehicles for positive engagement, there is a dark side to this liberty. Some participants exploit this platform to introduce and engage in malicious activities, turning otherwise positive virtual interactions into avenues for undesirable behaviors such as fighting, having sex, and killing others. [14–16]. In addition, a recent study by Kou et al. [2] revealed another safety concern, the UGC creators misuse the creative latitude offered by platforms like Roblox to introduce harmful designs. Examples include games promoting Nazi roleplay and embedding gambling-like mechanisms within these user generated environments. The risks are manifold, from the direct introduction of inappropriate content to the more subtle integration of problematic incentive structures within the UGCGs, each posing significant ethical and safety concerns.

2.2 Content Moderation

Content moderation has been extensively studied as an effective modern mechanism to address issues of toxicity and harassment in online spaces. At a high level, content moderation can be broadly defined as *"the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse"* [33], and can often be characterized as a series of trade-offs between actions, styles, philosophies, and values based on the context and facilitators of moderation [34]. Current moderation methods for UGCGs include filtering inappropriate or harmful words, filtering personal information, providing a reporting feature, and posting a public Code of Conduct the users of the games in question should obey [19, 21, 22]. Such content moderation strategies have been widely used in various online contexts but have also demonstrated several limitations. For example, TikTok users, who engage with a video-centric social media platform, have devised a strategy to circumvent the platform's algorithmic content moderation by deliberately misspelling prohibited words (e.g., using "seggs" instead of "sex") [35]. Cho et al. expand on this topic, discussing a wider array of text-based content moderation evasion tactics in their study [36]. These moderation strategies, however, are primarily text-based. Although there are effective methods for moderating textual content, there are no comprehensive solutions for moderating illicit image sharing for images used in UGCGs.

There exist AI/ML techniques aimed at identifying unsafe visual content, including inappropriate images and videos. Platzer et al. developed a machine learning approach utilizing Support Vector Machines (SVM) for recognizing pornographic images by analyzing distinct image features [37]. In another study, Yuan et al. explored the real-world issue of illicit online promotions involving pornographic images [38]. They introduced a novel deep learning approach that prioritizes regions in an image where sexual content is minimally obscured, aiming to overcome the challenges posed by adversarial sexually explicit images in the real world. Concerning video content, Tahir et al. highlighted the importance of detecting inappropriate videos within child-centric content-sharing platforms, such as YouTube Kids [39]. They collected a dataset of such videos and developed a multimodal model for detection purposes. Additionally, another study [40] provided a thorough analysis of inappropriate or disturbing videos aimed at toddlers, assembling a large-scale, labeled dataset for this purpose. The researchers trained a deep learning classifier, which yielded a promising detection rate. However, these techniques often rely on evaluating identified features, such as the extent of skin exposure in images and videos, or assessing the "humanness" of the subjects depicted [41]. Such criteria can be limiting and may not ef-

fectively address the diverse and complex nature of visual content in UGCGs. A significant gap persists in effectively moderating image content within UGCGs, as the existing AI/ML techniques are not tailored to discern the nuances of unsafe images specific to this domain [42]. Our study underscores this gap and emphasizes the imperative need for specialized moderation methods adept at identifying and mitigating the sharing of unsafe images in UGCGs, ensuring a safer and more inclusive gaming environment.

2.3 Large Vision-Language Models and Chain-of-thought Reasoning

The landscape of multimodal learning, especially in the domain of vision-language multimodal learning, has experienced significant advancements. Pioneering the contemporary discourse, models like CLIP [43] and BLIP [44] have marked noteworthy milestones. CLIP, for instance, has revolutionized the field by learning visual concepts from natural language descriptions, forging a symbiotic relationship between vision and language components to enhance performance across a multitude of tasks. BLIP further accentuates this integration, exemplifying robust performance and versatility in applications ranging from zero-shot to few-shot learning scenarios. As the discourse evolves, the emergence of large vision-language models heralds a new epoch in multimodal learning. These models, characterized by their incorporation of large language models (LLMs) to navigate vision-language tasks, are driving unprecedented progress. LLaVA [45], another significant development, underscores the synergy between visual and linguistic elements, optimizing performance in complex, dynamic environments. InstructBLIP [46], building upon the foundational principles of BLIP, integrates instructional learning paradigms to enhance model interpretability and task-specific adaptability. The GPT-4Vision [47] stands as a testament to the ongoing evolution, amalgamating extensive language modeling capacities with intricate visual comprehension, paving the way for a future where the confluence of vision and language is not just integrated but inherently synergistic. These cutting-edge large VLMs are characterized by their interpretability, adaptability, and enhanced contextual capability, offering many opportunities to tackle intricate and emerging challenges within the vision-language domain. In our study, we exploit the capabilities of large VLMs to address a specific issue: the illicit online promotion of unsafe images in UGCGs. These advanced models provide the necessary tools to identify and analyze subtle and complex patterns of unsafe content dissemination effectively.

Chain-of-thought (CoT) reasoning refers to the process where AI models, particularly LLMs, follow a sequence of logical steps to arrive at a conclusion or answer [30]. It involves connecting various information and ideas coherently and logically, akin to a human's natural thought process. This methodology enhances the model's ability to handle complex queries and problems, offering more contextually relevant and nuanced responses. The large VLMs with such enhanced reasoning capabilities herald a new phase in AI decision-making [48–50]. Although CoT has been instrumental in elucidating AI decisions [48, 51], its application and potential challenges in the specific realm of unsafe content moderation remain to be thoroughly explored and understood.

3 Threat Model

In our work, we examine the threat posed by adversaries who exploit social media platforms to promote unsafe UGCGs with inappropriate images, typically screenshots taken from UGCGs. We consider both adversaries and victims as users of these platforms and do not consider promotions outside the social media platforms. We do not attribute advanced capabilities to these adversaries nor assume the application of intricate adversarial tactics to evade content moderation. Our focus is on the straightforward yet effective strategies these actors employ to pervade digital platforms, exposing vulnerable online users to unsafe UGCGs.

In addition, we do not consider in-game content moderation of harmful content in our paper. Our system specifically targets identifying the illicit promotional images for unsafe UGCGs on social media platforms following a "soft moderation" strategy [52], *i.e.*, providing warnings to users about the potential harmfulness of the images, since some platforms may not consider such images illegal. X and Reddit allow users to flag inappropriate images as sensitive or "NSFW", which can prevent such content from appearing automatically in timelines, offering a form of user controlled moderation. Platforms such as X, Reddit, and Discord maintain their community guidelines and systems for reporting and eliminating inappropriate content. However, the enforcement of these guidelines can greatly differ across various community spaces, largely relying on the vigilance of community moderators or the effectiveness of automated detection tools. Despite these mechanisms, inconsistencies in enforcement create loopholes that enable the continued illicit image-based promotion of unsafe UGCGs, highlighting a significant safety challenge. The persistent threat of illicit UGCG promotions underscores a universal challenge: no social media platform should permit content that risks the safety of its users, particularly minors. Therefore, enhancing content moderation frameworks to address these gaps is crucial in our model, ensuring the digital environment remains safe and conducive to positive interactions.

Furthermore, we focus on illicit promotions of unsafe UGCGs based on images and do not consider the textual data in such posts, since images are the major components that drive promotions due to their visual appeal. We also posit that image-based promotions, by virtue of their visual appeal, pose a significant risk, particularly to children and adolescents.

4 Motivation and Observation

In this section, we present studies on understanding the nature of unsafe UGCGs in Roblox and their illicit promotions particularly among children, and the challenges in their detection. Our study begins with the analysis of real-world, self-reported stories related to unsafe UGCGs. Utilizing the identified keywords from our analysis, we then collect potential UGCG promotional images from social media platforms and engage human annotators to systematically annotate the images. We analyze the challenges of illicit online image promotions for unsafe UGCGs and assess the necessity for moderation techniques to flag illicit image promotions associated with unsafe UGCGs. This involves a comprehensive evaluation of the current tools designated for the detection of unsafe images online.

4.1 Data Collection and Annotation

Hashtag Identification. To find effective hashtags for collecting valid UGCG images, we first gathered online discussions with self-reported stories related to unsafe UGCGs from a public Internet resource, Common Sense Media [29]. This platform was chosen for two main reasons: it offers a comprehensive repository of real-world narratives related to UGCG, and it features insights specifically from children and parents, providing valuable perspectives on the effects of unsafe UGCG on younger groups. By October 2, 2023, we had compiled all Roblox-related discussions from this platform using an automated crawling technique, resulting in 7,081 stories collected. These stories span from July 9, 2009, to October 1, 2023. Then, a cleaning process ensued, during which we filtered out duplicated stories, those not written in English, and stories comprising fewer than five words to ensure the quality and relevance of the data. We employed KeyBERT [53] to automatically extract keywords from the collected stories. Then, we collected illicit promotional images of UGCGs by identifying hashtags associated with unsafe UGCGs. By analyzing self-reported stories, we first compiled keywords related to unsafe UGCGs. Utilizing these keywords, we constructed an initial hashtag list, including terms such as *#RobloxUGC*, *#RobloxCondo*, *#RobloxSex*, *#RobloxR34*, *#RobloxKiller*, and *#RobloxMurderMystery*. During the data collection phase, we systematically incorporated any new hashtag associated with a tweet containing one of the predetermined keywords into our evolving list. This list was continuously updated with emerging hashtags from new tweets until a saturation point was reached where no additional unique hashtags were discovered. The comprehensive list of these hashtags is detailed in Appendix A.

Image Collection. We used the Official X Streaming API[2] to collect public tweets during the period from January 1, 2020, to December 31, 2022 (*i.e.*, 2 years) based on the hashtags. More specifically, we used the tool's flag 'image' to make sure all the posts we collected were along with images. At the end of our data collection process, we have retrieved 38,182 image-based tweets and extracted 29,858 images immediately between January 2020 and December 2022.

Image Filtering and Annotation. We selected a random sample of 4,000 images to serve as the subjects of our study. Each image in the dataset was initially processed to enhance its clarity and quality. Inspired by the methodology outlined by Phan et al. [54], we employed a three-step filtering process to further refine our dataset. In the first step, we utilized the Python library called Pillow [55] to automatically eliminate images with either a width or height of fewer than 300 pixels. In the second step, we manually inspected the remaining images, filtering out those not associated with UGCGs. In the third step, we rigorously reviewed the images to ensure each was sufficiently clear to be easily perceived by humans. Finally, the images were independently annotated by three authors. We utilized Natural Language Processing (NLP) techniques, particularly the agglomerative hierarchical clustering (AHC) method [56], to categorize the stories into unsafe groups, resulting in the identification of four primary categories of unsafe UGCGs: "sexually explicit", "violent", "bullying", and "scam". By reviewing each group and their representative stories, we built our codebook for UGCG image annotation (see Appendix B). We further employed Fleiss' Kappa score [57], a statistical measure used to assess reliability among multiple annotators. The scores revealed a progressive improvement in inter-annotator agreement throughout the three coding rounds, beginning with a *substantial* agreement level of 74%, then increasing to an *almost perfect* agreement level of 92%, and finally reaching an *almost perfect* agreement level of 100%. Ultimately, this process resulted in the identification of 2,924 valid UGCG images, including 1,621 images classified as "sexually explicit", 202 as "violent", and 1,101 as "safe". However, we did not find any images that could be categorized as "bullying" or "scam".

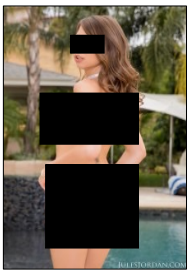
4.2 Detection Challenges of Illicit Promotional Images of UGCGs

4.2.1 Nature of UGCG Promotional Images

In the course of our data collection and annotation process, we noted a clear distinction in the nature of promotional images for UGCG advertisements compared to conventional game promotions, which typically use renditions crafted by professional entities. Specifically, we noted that the images for UGCG advertisements are predominantly generated by individual users who tend to utilize direct screenshots from the games to showcase their content, as exemplified in Figure 1.

To validate this observation and support our hypothesis regarding the nature of UGCG promotion, we conducted an analysis on a randomly selected subset of 500 images from our dataset. The findings were significant: 97.8% of

[2] https://developer.twitter.com/en/docs/twitter-api


Sexually-explicit-human


Sexually-explicit-anime

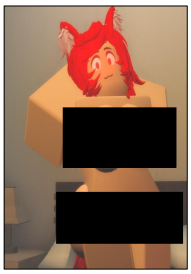

Sexually-explicit-UGCG

Figure 2: Samples of sexually explicit images. The Google Vision API's prediction for the image as *sexually-explicit-human*: "VERY_LIKELY" for "Adult" and "Racy"; for the image as *sexually-explicit-anime*: "VERY_LIKELY" for "Adult", "POSSIBLE" for "Racy"; for the image as *sexually-explicit-UGCG*: "UNLIKELY" for "Adult" and "Racy".

these images were indeed screenshots directly taken from user generated games. This prevalence underscores the subtlety with which these advertisements integrate into the platform, highlighting the need for enhanced moderation tools, which are essential to identify and mitigate the discreet yet pervasive spread of unsafe UGCG advertisements, safeguarding users from potentially unsafe content.

4.2.2 Evading State-of-the-Art Unsafe Image Detectors

Following our previous studies, We wanted to investigate the feasibility of using existing detectors for moderating the image-based illicit online promotion of UGCGs. To understand the effectiveness of existing unsafe image detectors, we conducted an experiment regarding state-of-the-art (SOTA), commercially available detectors by measuring their capability to detect promotional UGCG images from our dataset. In our work, we selected five SOTA detectors that are widely used and have the capability to detect unsafe images, which are Clarifai [24], Yahoo Open Not Safe For Work (NSFW) [28], Amazon Rekognition [25], Microsoft Azure [27], and Google Vision AI [23]. Due to the ubiquity and effectiveness of these detectors, they can be considered representative of the technology used to defend against unsafe content in existing online platforms. To study the capability of these detectors about new and traditional types of sexually explicit images, our experiments incorporated an existing dataset [58] that encompasses traditional sexually explicit content, both from real-world scenarios and animated sources. From this dataset, we randomly selected 1,000 images each from the categories labeled "porn" and "hentai", representing real-world (*i.e.*, sexually-explicit-human) and animated explicit content (*i.e.*, sexually-explicit-anime), respectively. Concurrently, another 1,000 images were randomly selected from our annotated dataset, all labeled as "sexually-explicit" (*i.e.*, sexually-explicit-UGCG). This selection served to evaluate the detectors' proficiency in identifying sexually explicit content in UGCGs, offering a comparative insight into their capability to detect traditional and emerging forms of explicit imagery. Figure 2 is an example demonstrating the different categories of images used in this study.

The responses from existing detectors exhibited a diversity in their output formats. Systems like Clarifai, Yahoo Open NSFW, and Amazon Rekognition provided probability scores in the range from 0 to 1 as outputs to quantify the likelihood of images being unsafe. In contrast, Microsoft Azure presented a binary true or false label to categorize such images. Google SafeSearch offered a more nuanced classification, with labels ranging from "UNKNOWN" to "VERY_UNLIKELY", "UNLIKELY", "POSSIBLE", "LIKELY", and "VERY_LIKELY", each indicating the varying degrees of probability for an image being deemed unsafe. Given this diversity in output formats, establishing a unified criteria for evaluation was crucial. Based on these varying methods of measuring whether or not an image is unsafe, we used the following thresholds to determine if an unsafe image is detected. For Clarifai, Amazon Rekognition, Azure, and Yahoo Open NSFW are logit values in the range $0-1$. The threshold value is set at 0.5 for unsafe images. For Google Vision AI, if the model output is "LIKELY" or "VERY_LIKELY", we assume that the model identifies unsafe images, which is a common criterion that has been discussed in previous work [59].

For the selected unsafe image detector, Google Vision AI, we provide different types of unsafe images as inputs and observe the detection results of the detector. Table 1 presents the results of this experiment. The results indicate that SOTA exhibit exemplary performance in detecting unsafe images depicting humans and anime characters, with all detectors achieving an average detection rate exceeding 90%. Notably, Google Vision AI demonstrates superior efficacy, achieving detection accuracies of 98% and 99% for human and anime sexually explicit images, respectively. Contrastingly, a pronounced decline in performance is observed when these models are tasked with identifying unsafe content within UGCG images. The detection efficacy of these sophisticated tools is markedly compromised in this context, with four of the tested models yielding detection rates below 20%. Google Vision

Image Type	State-of-the-Art Unsafe Image Detectors				
	Clarify	Yahoo Open NSFW	Amazon Rekog-nition	Micro-soft Azure	Google Vision AI
Sexually-explicit -human	88%	92%	98%	92%	98%
Sexually-explicit -anime	89%	81%	91%	90%	99%
Sexually-explicit -UGCG	13%	13%	17%	15%	67%

Table 1: Effectiveness of SOTA unsafe image detectors.

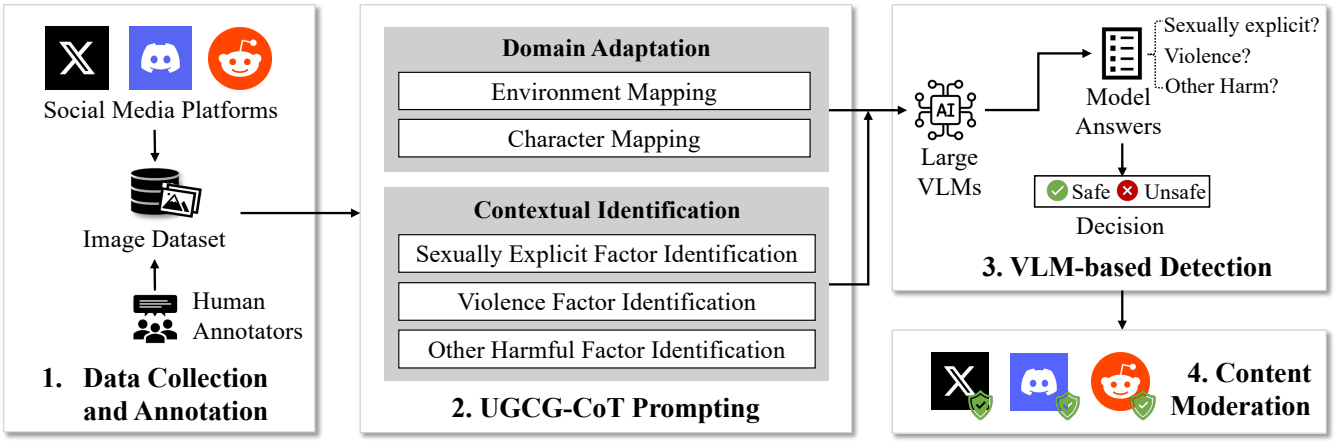


Figure 3: Overview of UGCG-GUARD.

AI, albeit the most effective among the evaluated models with a detection rate of 67%, falls short of the accuracy and reliability requisite for real-world content moderation applications. We believe that the superior performance of both human and anime images could be attributed to the similarity in the domain of these images, and also to the presence of training samples from both domains. Anime images, although animated, are semantically close to real human images, and are defined similarly to real humans. However, this is not the case for the UGCG images. As a result, this comparative analysis underscores the challenges in the detection of unsafe content within UGCGs: there is a marked domain shift in the case of UGCG images that renders existing systems severely limited.

5 UGCG-GUARD Design

5.1 Overview of UGCG-GUARD

The overview of our framework, UGCG-GUARD, is presented in Figure 3, which 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, utilizing the methodology described in Section 4.1. UGCG-GUARD incorporates human annotators to verify and label images based on the activities identified from our study of unsafe 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. 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.

5.2 Our Approach

5.2.1 Crafting UGCG-COT Prompts

Navigating the intricacies of UGCGs requires our prompts to be adept at tackling the dual challenges of domain adaptation and contextual identification by leveraging large VLMs to make moderation decisions effectively. However, VLMs cannot be directly used for this purpose. In the following, we discuss how our novel CoT reasoning-based prompting approach, UGCG-COT, which is illustrated in Figure 4. This approach combines conditional prompting and reasoning-based prompting strategies and can effectively leverage VLMs to perform domain adaptation and contextual identification, respectively.

Conditional Prompting for Domain Adaptation. Domain adaption is a crucial methodology that facilitates AI/ML models in navigating and adapting to the transition between distinct domains, enhancing their performance and applicability across varied contexts [60, 61]. In this light, our domain-adaptation prompt is instrumental in aiding the model to delineate the nature of the provided images. We utilize conditional prompting for domain adaptation to instruct UGCG-GUARD to understand the specific characteristics of UGCG images in a zero-shot manner. More specifically, the purpose of conditioning in our approach is to refine the search space of VLMs, thereby enhancing the overall effectiveness of our approach. In our setup, we employed prompting as a method of conditioning, concentrating exclusively on images akin to those found in UGCGs. This strategy also aids in minimizing the influence of irrelevant training data on the model. This refinement is achieved through a two-step structure consisting

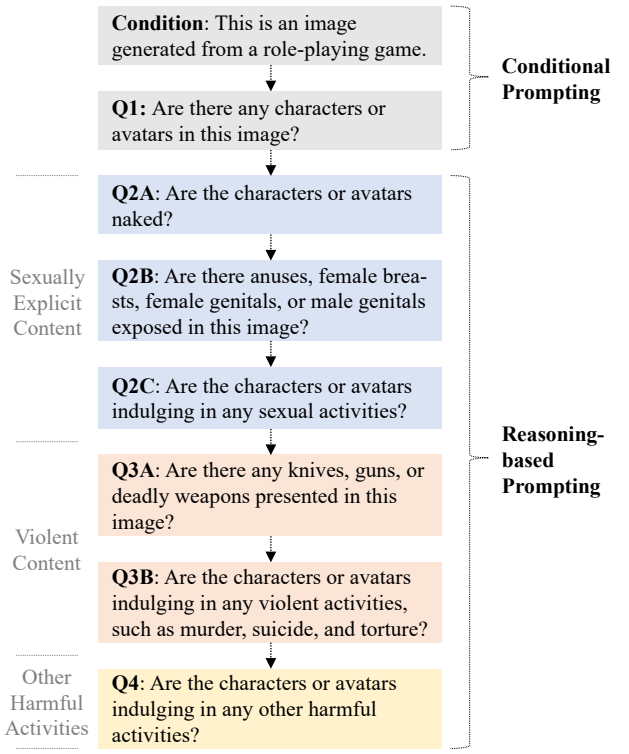


Figure 4: UGCG-CoT prompting.

of a condition and a guidance question. The condition articulated as *Condition*: *"This is an image generated from a role-playing game."* serves to anchor 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. This dual structure ensures that the model is not only informed of the simulated nature of the images but is also guided to focus on specific elements within them, facilitating a more refined and contextually appropriate analysis.

Reasoning-based Prompting for Contexual Identification. 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 enhance UGCG-GUARD ability to perform reasoning, consequently enabling the VLM to make contextual decisions [30] and ensuring that unsafe content can be identified and moderated. Our image filtering and annotation processes revealed a crucial observation: despite the pronounced domain shift characteristic of UGCG images, the attributes defining unsafe content remain analogous to those in traditional real-world images. Drawing inspiration from established methodologies in unsafe real-world image detection [54, 62–64], we meticulously crafted our contextual prompts. 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 end, through our comprehensive Contextual Identification reasoning prompts, we equip large VLMs with the enhanced capability to efficiently identify and flag unsafe UGCG images, ensuring a balanced approach that is both domain-specific and context-sensitive.

5.2.2 Leveraging Large VLM for Processing UGCG-CoT Prompts

We leverage VLMs to process the UGCG images in conjunction with the UGCG-CoT prompts, with specific criteria to ensure optimal performance. The first requirement for running our prompts is that the selected VLM should be comprehensively trained on an extensive array of varied vision-language tasks. Training on both vision and language modalities allows VLMs to understand the commonalities among various input domains, thus giving them domain adaptation capabilities. Furthermore, the incorporation of powerful encoders ensures the precise extraction and processing of complex factors and features inherent in UGCG images. Our second requirement stipulates that the model must possess enhanced reasoning capabilities. This is paramount due to the intricate nature of UGCG images and the nuanced, potentially unsafe content embedded within them that needs contextual detection. The adoption of the chain-of-thought methodology [30] ensures advanced reasoning capacity. This approach has been empirically validated to significantly elevate the decision-making proficiency of large language models (LLMs) [30]. In this scenario, the role of LLMs becomes pivotal. They act as decision-making models that process vision and language, facilitating reasoning-based decision-making based on both textual and visual features. Moreover, LLMs, enriched by their extensive training datasets, are essential for embodying CoT reasoning, helping UGCG-GUARD leverage the extensive knowledge based on these models to make accurate decisions. Thus, the synergistic combination of large VLMs, intensively trained across a spectrum of tasks and endowed with amplified reasoning faculties, emerges as a suitable choice to run our prompts.

Our system leverages a VLM to operationalize our prompting strategy in the following way. Given im-

age input X_{vision} and supplemented by the UGCG-CoT prompts, characterized as the language input $X_{language} \in \{X_{language}^{condition}, X_{language}^{Q_1}, X_{language}^{Q_{2A}}, ...\}$, the output is computed as,

$$\hat{y} = \text{argmax } p(y|X_{vision}, X_{language}). \quad (1)$$

Using our UGCG-CoT prompts, we decompose the primary problem of detection of unsafe images used in the illicit promotion of unsafe UGCGs into a series of sub-problems. This enables a structured and sequential approach to decision-making, where the final output $\hat{y}$ is a culmination of insights derived from intermediate states. To be specific, the steps are as follows.

Step 1. We first condition the VLM to enable domain adaptation from the real-world context to the simulated gaming environment, enhancing its ability to interpret UGCG images. In this process, the model integrates and processes the UGCG image, denoted as X, in conjunction with a specified condition C. This integration tailors the model's attention features, represented as X_C, ensuring they are attuned to the nuances of the UGCG content within the image.

Step 2. We then condition the VLM to achieve domain adaptation from identifying real humans to identifying human-like figures within the UGCG image, depicted as follows:

$$A_1 = \text{argmax } p(a|X_C, Q_1), \quad (2)$$

where a is an intermediate answer that the VLM could output, such as *Yes*, *No*, *N/A*, etc. In this step, UGCG-GUARD identifies valid characters. If detected, the framework proceeds to perform a context check; otherwise, it allows the input image to pass through unchanged without further evaluation.

Step 3. Next, we prompt the VLM to check if sexually explicit content is presented in the input image.

$$\begin{aligned} A_{2a} &= \text{argmax } p(b|X_C, Q_{2A}), \\ A_{2b} &= \text{argmax } p(c|X_C, Q_{2B}), \\ A_{2c} &= \text{argmax } p(d|X_C, Q_{2C}), \end{aligned} \quad (3)$$

where $b, c, d, ...$ are intermediate answers as a in Equaiton 2.

Step 4. Then, we prompt the VLM to identify if violent content is presented in the input image.

$$\begin{aligned} A_{3a} &= \text{argmax } p(e|X_C, Q_{3A}), \\ A_{3b} &= \text{argmax } p(f|X_C, Q_{3B}). \end{aligned} \quad (4)$$

Step 5. In addition, we prompt the VLM to output whether other harmful activities are shown in the image.

$$A_4 = \text{argmax } p(g|X_C, Q_4). \quad (5)$$

Step 6. The final decision regarding the safety of a UGCG image is determined by aggregating all previous outputs from the VLM. If any of the contextual questions receive a positive answer *i.e.*, yes, the UGCG image is flagged as unsafe. This decision is represented as follows:

$$\begin{aligned} A_{\text{all}} &= \{A_{2a}, A_{2b}, A_{2c}, A_{3c}, A_4\}, \\ \hat{y} &= \begin{cases} \text{unsafe}, & \text{if } \exists a_i \in A_{\text{all}} : a_i = \text{yes}, \\ \text{safe}, & \text{otherwise.} \end{cases} \end{aligned} \quad (6)$$

5.2.3 Illicit Online Image Promotion Moderation

In this section, we outline the deployment of UGCG-GUARD for real-world social media platforms, as detailed in Algorithm 1. UGCG-GUARD employs the large VLM to evaluate promotional images of UGCGs systematically. Initially, it ascertains the presence of a valid character in the image. Subsequently, the framework assesses the content to identify sexually explicit or violent elements. Additionally, UGCG-GUARD scrutinizes the image for other potential safety concerns and annotates any identified issues. In the end, UGCG-GUARD aggregates the responses from the large VLM to determine the safety status of the promotional image. If an image is identified as "unsafe", UGCG-GUARD will flag the image and issue a warning stating why. Otherwise, UGCG-GUARD will approve the image for sharing on social media platforms. The comprehensive assessment ensures that images are flagged

Algorithm 1: Illicit UGCG Promotional Image Safety Analysis Using Large VLM

Input: UGCG image X_{vision}, UGCG-CoT prompts $T = \{C, Q_{\text{dom}}, Q_{\text{sex}}, Q_{\text{violent}}, Q_{\text{harm}}, Q_{\text{dec}}\}$, Inference Function F, Large Vision Language Model (M);
Output: Safety prediction $\hat{y}$ with reason

```
                                        // Domain Adaptation
M_{X_C} ← M(X, C)
y_1 ← F_domain(M_{X_C}, Q_dom)
if y_1 = "no" then
    return y_1
end if
                                        // Contextual Identification
for q ∈ Q_sex ∪ Q_violent ∪ Q_harm do
    A_q ← F(M_{X_C}, q)
    if A_q = "yes" then
        continue
    end if
end for
                                        // Decision Making
if any A_q = "yes" then
    ŷ = 'unsafe'
    return ŷ,
      Flag the image and issue a warning with the reason.
end if
else
    ŷ = "safe"
    return ŷ, Approve the image for sharing.
end if
```

appropriately, enhancing the safety and quality of content on social media platforms.

6 Implementation and Evaluation

In this section, we first discuss the implementation of multiple components of our system, followed by experiments to evaluate our approaches to identify illicit promotional images of unsafe UGCGs from different perspectives. Our evaluation goals are summarized below.

- Understand the effectiveness of UGCG-GUARD by comparing it against existing baseline detectors. (§ 6.3)
- Analyzing the capability of UGCG-GUARD to address the challenge of the shift from traditional unsafe image to the UGCG input domain by comparing it with state-of-the-art object detection methods. (§ 6.4)
- Investigating the effectiveness of the conditioning process of UGCG-GUARD. (§ 6.5)
- Examining the effectiveness of the contextual identification process of UGCG-GUARD. (§ 6.6)
- Running UGCG-GUARD on "in-the-wild" samples from diverse social media platforms. (§ 6.7)
- Examining the limitations of traditional vision models for detecting unsafe UGCG images. (§ 6.8)

6.1 Implementation Details

In this section, we discuss the implementation specifics of UGCG-GUARD. We utilized the InstructBLIP model, *instructblip-vicuna-13b* [46], as our preferred large VLM for the large-scale execution and evaluation of UGCG-COT prompts. We deployed UGCG-GUARD with a High-Performance Computing (HPC) system equipped with two 40GB Nvidia A100 graphics cards. Most of our evaluation experiments were conducted using the labeled dataset described in Section 4.1, with an additional 322 illicit promotional UGCG images gathered from Reddit and Discord. This dataset formed as a test case that represented "in-the-wild" UGCG images, providing insights into the model's performance in real-world scenarios, as detailed in Section 6.7. In this experiment, we also built and evaluated UGCG-GUARD based on the GPT-4V(ision) model to demonstrate the generalizability of our framework. The other specific parameter settings, excluding those set to default values, are discussed in the evaluation.

6.2 Baselines

We evaluated UGCG-GUARD with several pertinent baselines to establish a comparative evaluation: (1) Google Vision AI, a SOTA commercial approach renowned for its efficacy; (2) A Convolutional Neural Network (CNN) model [65], trained with large-scale image data exceeding 60 gigabytes, comprises a diverse collection of real-world and animated NSFW images renowned for its benchmark performance [66]. In our experiment, we name it as "NSFW-CNN". (3) A general prompt for unsafe image detection, employing the same large VLM as UGCG-GUARD, to evaluate the impact of different prompt strategies on detection performance.

6.3 Comparison Against Baselines

In this experiment, 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 [23] as a representative of cutting-edge commercial systems for unsafe image detection. Additionally, we considered the NSFW-CNN [66], a CNN model built upon MobileNetV2 architecture [67], 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 model, InstructBLIP, 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 [68], was phrased as follows: "Is this image unsafe? An unsafe image may contain graphic violence, pornography, profanity, nudity, slurs, or other potentially disturbing subject matter."

The results of our experiment are depicted in Table 2. The NSFW-CNN pre-trained model demonstrates the worst performance, as evidenced by its accuracy of 0.5 and F1 score of 0.47. This can be attributed to its high dependency on the dataset it was trained on. The images of UGCGs are very different compared to both real-world and animated visuals, presenting that the traditional pre-trained AI/ML models, such as the NSFW-CNN model, can not be sufficiently equipped to handle due to the training data limitations. Google Vision AI and InstructBLIP with a general prompt ("InstructBLIP-

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 -GEN	0.6	0.88	0.42	0.57
UGCG-GUARD	0.94	0.98	0.91	0.94

Table 2: Comparing UGCG-GUARD against the baselines.

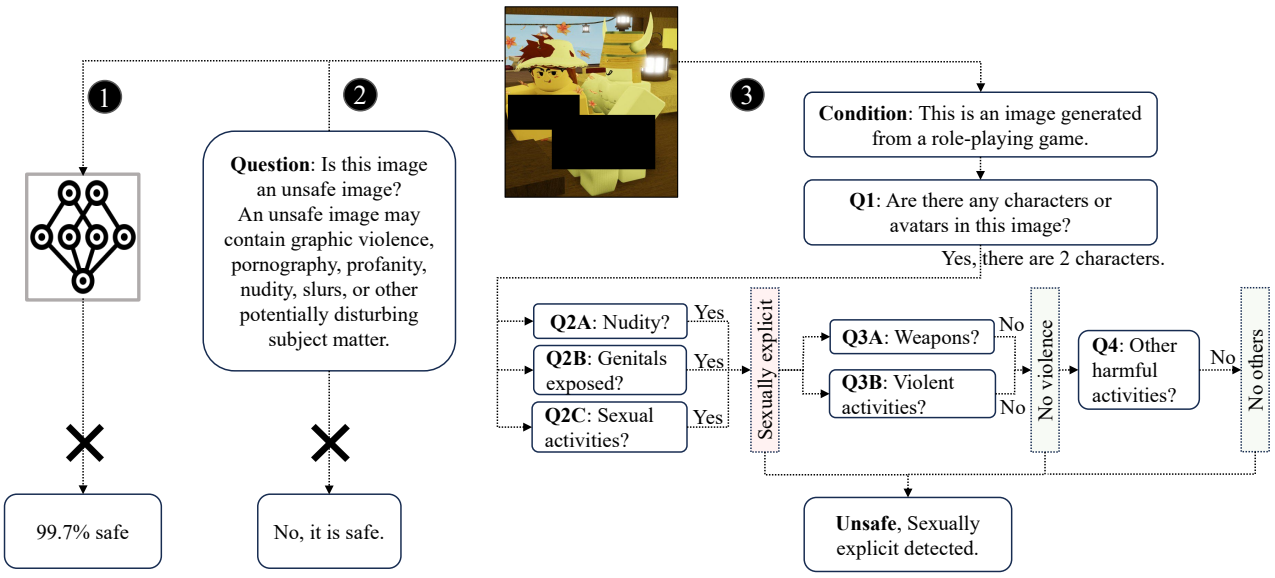


Figure 5: ❸ UGCG-GUARD for image-based illicit promotion of UGCG detection compared to ❶ traditional unsafe image detection models and ❷ large VLMs with general prompting strategy.

GEN") exhibit comparable results, with accuracy levels of 0.68 and 0.6, respectively. However, these systems are not suitable for practically flagging such content due to their limited efficacy. InstructBLIP-GEN excels in the precision score, indicating its strength in minimizing false positives. However, its recall is the lowest, pointing towards the limitation in identifying unsafe images of UGCGs. Google Vision AI has a more "balanced" performance, although it still achieves limited recall and F1 scores. UGCG-GUARD distinctly outperforms all the baselines, with an impressive accuracy of 0.94 and an F1 score of 0.94, indicating its effectiveness in flagging images-based illicit promotion of unsafe UGCGs. The results underscore the efficacy of employing chain-of-thought reasoning [30] in identifying unsafe UGCGs, as illustrated in the sample depicted in Figure 5. Unlike the general zero-shot prompting strategy that probes the large VLM once, UGCG-COT adopts a multi-step decision-making approach, integrating model conditioning and contextual questions. This method capitalizes on the large VLM's capabilities to enhance the detection of illicit promotional images of UGCGs through a structured reasoning process. It ensures that the final decision is not just an immediate output, but is derived from a comprehensive evaluation, resulting in increased precision and recall.

6.4 Capabilitiy of UGCG-GUARD in Domain Shift

In this experiment, our goal was to evaluate UGCG-GUARD's capability to interpret images of UGCGs, distinguishing them from the realm of traditional unsafe images. We focused on the response to question Q2 within UGCG-GUARD, detailed in Section 5.2.1. This approach allowed us to examine the adaptability of UGCG-GUARD in identifying the presence of characters or avatars amidst the domain shift to UGCG imagery. We assessed UGCG-GUARD's adaptability to this new domain by comparing it with four renowned object detection tools: Yolo [69], SSD [70], Faster RCNN [71], and Google Vision AI. These tools are capable of detecting objects that they're trained on and assign a confidence score to each prediction. In our study, we focused on the object "person" due to the nature of our task, in which unsafe images predominantly depict a persona. A threshold of 0.5 was established for the prediction scores to evaluate the performance of each model in identifying valid human-like figures in UGCG images. Three distinct datasets were employed to scrutinize the efficacy of each method. These include real-world and animated sexually explicit datasets, each containing 1,000 images, as detailed in Section 4.2.2, and an additional 1,000 sexually explicit images from our dataset. Each image in this custom dataset has been manually verified to contain human-like figures, ensuring consistency in our evaluation criteria.

Image Type	Yolo	SSD	Faster RCNN	Google Vision AI	UGCG-GUARD
Sexually-explicit-human	97.4%	98.8%	99.9%	89.6%	99.9%
Sexually-explicit-anime	97%	99%	99.9%	89.2%	100%
Sexually-explicit-UGCG	3.2%	18.2%	12.4%	12.8%	98.2%

Table 3: Comparing SOTA object detection tools.

Table 3 presents the results. Each technique demonstrates effectiveness in detecting real-world and anime people, with Faster RCNN and InstructBLIP achieving exemplary performance. However, their proficiency diminishes significantly when applied to UGCG sexually explicit images, with none of the object detection tools surpassing an 18.2% detection rate. In contrast, UGCG-GUARD is the only effective tool capable of identifying human-like figures in unsafe UGCG images, showing a remarkable 98.2% detection rate. The stark contrast in detection rates points out the significant shift in the input domain when transitioning from real-world and animated images to UGCGs, and the exceptional performance of UGCG-GUARD highlights the important role that large VLMs can play in adeptly navigating and overcoming this challenge, marking a notable advancement in adapting to and mitigating the complexities introduced by the diverse content in UGCGs.

6.5 Effectiveness of Conditioning Prompts

In this experiment, we investigated the impact of condition prompts in UGCG-COT, *i.e. Condition* and Q_1 as discussed in Section 5.2.1. To be specific, without the conditioning process, UGCG-GUARD probes the large VLM directly with contextual identification questions and derives conclusions based solely on their responses.

The results can be observed in Table 4. Without the conditional prompts, UGCG-GUARD exhibits notable performance, displaying an accuracy of 0.79 and an F1 score of 0.82, which significantly outperforms the baselines outlined in Table 2. The incorporation of conditioning, however, elevates UGCG-GUARD's performance exponentially. The precision escalates to an impressive 0.98, while recall advances to 0.91, denoting a substantial enhancement in identifying and capturing unsafe content of UGCGs. The results highlight the significance of the conditional prompts in narrowing the search space of large VLMs, enabling them to focus on knowledge pertinent to UGCG or similar images. This focus significantly boosts the overall efficacy of UGCG-GUARD in detecting unsafe content. This enhancement can also potentially be attributed to the combined vision and language training embedded in the large VLMs [43, 44, 72]. In contrast, visual models are solely trained on images. This dual training paradigm of large VLMs allows for a more comprehensive understanding and interpretation of content, leading to increased accuracy and efficiency in the detection of illicit UGCG images.

Methods	Accuracy	Precision	Recall	F1
UGCG-GUARD without conditioning	0.79	0.91	0.74	0.82
UGCG-GUARD with conditioning	0.94	0.98	0.91	0.94

Table 4: Comparison of UGCG-GUARD w/ and w/o the conditioning process.

Prompt Ablations		Detection Rate
Conditional Prompts +	only Q2A	86.5%
	only Q2B	56.6%
	only Q2C	42%
	only Q2A & Q2B	94.6%
	only Q2A & Q2C	93.3%
	only Q2B & Q2C	73.3%
	Q2A & Q2B & Q2C	98.2%

Table 5: Ablation study for reasoning-based prompts.

6.6 Effectiveness of Reasoning-based Prompts

In this experiment, we assessed the effectiveness of reasoning-based prompts for contextual identification. To illustrate this, we specifically focused on the application of these prompts in identifying sexually explicit content. Note that we ensured the consistency of our approach by maintaining conditional prompts for the domain adaptation process throughout all evaluations. We then applied specific reasoning-based prompts for identifying 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?", and Q2C: "Are the characters or avatars indulging in any sexual activities?" to the same dataset of 1,000 sexually explicit UGCG images as mentioned in Section 4.2.2.

We further conducted an ablation study to evaluate the effectiveness of our reasoning-based prompting strategy. As shown in Table 5, we initially assessed the detection rates for each sexually explicit contextual identification prompt: Q2A, Q2B, and Q2C, with results of 86.5%, 56.6%, and 42%, respectively. These results indicate that while nudity is common in sexually explicit UGCG images, reliance on a single prompt for detection is inadequate. Subsequently, we evaluated the impact of removing each prompt individually. Removing Q2A decreased the detection rate from 98.2% to 73.3%. Eliminating Q2B made the detection rate drop to 93.3%, and upon removing Q2C, the rate decreased to 94.6%. The results demonstrate the remarkable effectiveness of UGCG-GUARD's reasoning-based prompts. Together, these prompts establish a comprehensive reasoning process, effectively guiding large VLMs toward accurate final decisions.

6.7 Running UGCG-GUARD on Unlabeled Samples "In-the-Wild"

We conducted an experiment on the *unlabeled* samples from two other social media platforms, Reddit and Discord, to simulate an "in-the-wild" running scenario that leverages our ap-

proach to control the real-world image-based illicit promotion of UGCGs. On Reddit, we employed the keywords associated with unsafe UGCGs, as identified in our preceding study detailed in Section 4.1, and manually collected a total of 112 images used for the illicit promotion of UGCGs. Of these, 33 were classified as safe, while 79 contained unsafe content. In the case of Discord, we obtained illicit promotional images of UGCGs utilizing a server listing platform [73]. This platform enabled our entry into a variety of Roblox game servers on Discord, leading to the discovery of 210 instances for image-based illicit promotions of UGCGs. Among these, 92 images were classified as unsafe, while the remaining 118 were safe. Subsequently, we evaluated UGCG-GUARD with these in-the-wild unsafe UGCG images and underwent an evaluative process, benchmarked against three pre-existing models: Google Vision AI [23], Clarifai [24], and NSFW-CNN [66]. The recent unveiling of GPT-4V(ision) [47, 74] has garnered significant attention, demonstrating exceptional performance in various studies [75, 76]. Our framework, UGCG-GUARD, with its adaptable architecture, can also be deployed based on different large VLMs such as GPT-4V. In this experiment, we assessed the generalizability of UGCG-GUARD by integrating it with both InstructBLIP and GPT-4V models.

Table 6 illustrates the results of the "in-the-wild" experiment. In this table, "R" stands for Reddit, "D" stands for Discord. Clarifai underperforms on both Reddit and Discord datasets, with F1 scores significantly lower than 0.5, indicating its limited capability in identifying unsafe UGCG images effectively. Compared to other baselines, NSFW-CNN achieved better performance with an average F1 score of 0.58. However, it still falls short of the feasibility required for practical applications. Despite Google Vision AI achieving comparatively decent results on both datasets with an average F1 score of 0.78, its low recall, especially for UGCG images from Reddit, suggests that this popular commercial detector is currently insufficient to meet the challenges posed by such illicit online image promotions. Both InstructBLIP-based UGCG-GUARD and GPT-4V-based UGCG-GUARD achieved significantly better results, outperforming other baselines. UGCG-GUARD integrated with InstructBLIP achieved an impressive average accuracy of 0.92 and an average F1 score of 0.93

Detectors	Accuracy		Precisioin		Recall		F1	
	R	D	R	D	R	D	R	D
Clarifai	0.44	0.73	1	1	0.22	0.27	0.36	0.43
NSFW-CNN	0.57	0.78	1	1	0.4	0.41	0.57	0.58
Google Vision AI	0.71	0.87	0.98	0.96	0.59	0.74	0.74	0.83
UGCG-GUARD (InstructBLIP)	0.91	0.93	0.96	0.88	0.92	0.98	0.94	0.92
UGCG-GUARD (GPT-4V)	0.88	0.9	1	0.97	0.83	0.79	0.91	0.88

Table 6: The "in-the-wild" experiment.

Figure 6: Samples of sexually explicit images used in the ablation study.

across the Reddit and Discord datasets, highlighting its exceptional ability to identify unsafe UGCG images. Similarly, UGCG-GUARD integrated with GPT-4V also showed strong results, with an average accuracy of 0.89 and an F1 score of 0.9, confirming its effective adaptability across different large VLMs. The overall performance, with an average F1 score of 0.91, validates the possibility of UGCG-GUARD for real-world deployment, and it also underscores the generalizability in detecting unsafe UGCG images within various resources.

6.8 Comparison Against Traditional Vision Models

In this experiment, we study the limitations of traditional, vision models for the detection of unsafe UGCG images. We fine-tune a ResNet-based CNN model [77] utilizing 80% of the samples from our annotated UGCG image dataset and the rest for testing. We conducted an ablation study, where we removed the sensitive regions of 20 unsafe UGCG, as depicted in Figure 6, and analyzed them with both the ResNet model and our system. These images should ideally not be considered sensitive since sensitive regions in these images have been removed. This enables us to discern if the ResNet model was effectively identifying unsafe UGCG images contextually, thereby ensuring its effectiveness in detecting such content reliably, or whether they are training on unrelated visual artifacts, like skin color patterns in sensitive images.

Initially, we tested the ResNet model on unaltered samples, achieving 0.9 accuracy and a 0.87 F1 score, competitive with UGCG-GUARD. Subsequently, we processed 20 modified images through our system, accurately identifying 18 as "safe" due to the removal of explicit content. However, the CNN model labeled all modified images as "unsafe". These outcomes underscore a critical observation: the traditional ResNet model, although adept in certain contexts, tends to overfit, becoming overly sensitive to specific visual cues that are not directly associated with unsafe content and is not practically suitable due to a high false positive rate. It highlights

an inherent limitation in its capacity to distinguish between genuine unsafe elements and incidental visual patterns. This revelation underscores the necessity for a more refined and nuanced architecture like UGCG-GUARD that is adept at detecting unsafe images in a contextual manner.

7 Discussion

Limitations. Our study has several limitations. First, our analysis is based on reviews obtained from Common Sense Media. Although this platform offers a rich array of insights from young users and parents, facilitating the identification of four distinct unsafe topics within UGCGs, it exclusively features English content. This linguistic limitation potentially narrows the scope of uncovered unsafe content in other languages and cultures. Broadening the research to incorporate additional review and community platforms, especially those in diverse languages, could yield a more nuanced and comprehensive understanding of the unsafe content prevalent in UGCGs. The evaluation of "in-the-wild" scenarios was constrained by a manually collected dataset, which has a limited number of data from Reddit and Discord. To enhance the generalizability of UGCG-GUARD, efforts will be enhanced towards amassing a broader dataset from various popular social media platforms. It will not only solidify the robustness of UGCG-GUARD but also ensure its efficacy and adaptability for real-world digital environments. Another limitation is that the focus of our study is currently confined to UGCGs within Roblox. Numerous other platforms exist, such as Minecraft [78], Terasology [79], and LEGO Worlds [80], which enable users to create their content and attract a significant population of children and adolescents. It is plausible that unsafe activities prevalent in Roblox might also be present in these platforms, and potentially manifest distinct characteristics that necessitate specific moderation strategies. Therefore, extending the analysis to include these platforms could provide a more holistic understanding of the safety challenges associated with UGCGs. Additionally, we observed that illicit promotions of unsafe UGCGs occasionally employ diverse modalities, including GIF images and short videos. We believe that once we collect enough data for these varied content formats, the incorporation and evaluation could fortify our analysis, offering a more comprehensive assessment of the multifaceted nature of the illicit promotion of unsafe UGCGs.

Ethical Considerations. In our work, we annotated the illicit promotional images of unsafe UGCGs by three of the authors, and no additional workers were recruited in the whole study. Our data collection task was approved by IRB. Every author participant in the process understands the unsafe content before our task. In our paper, we ensured the removal of mentions to user accounts so that no user information could be traced via public social media. In addition, we will take all the necessary steps, such as only sharing the data with verified researchers.

In-game Content Moderation Our present work is centered on identifying and moderating the image-based illicit promotion of unsafe UGCGs, a crucial endeavor given the associated risks posed to online users, particularly children [2]. However, we believe that the insights gleaned from our current research can shed light on the expansion of future work on in-game unsafe content moderation. By moderating both the game promotions and UGCGs themselves, we can minimize the dangers posed by the UGCGs as comprehensively as possible, thereby protecting users, especially young individuals, shielding them from exposure to unsafe content and activities.

8 Conclusion and Future Work

In this work, we have embarked upon an initial journey to understand and detect the threat of illicit promotion of unsafe UGCGs. We have conducted an insight study to understand the limitation of existing unsafe image detectors and present the urgent need for effective moderating mechanisms. With the studies, we proposed a novel framework UGCG-GUARD to practically address the problem of image-based illicit UGCG promotions. Our evaluation demonstrates that UGCG-GUARD can effectively capture the unsafe images used in the illicit promotion of unsafe UGCGs.

In the future, we plan to extend our framework to adapt it for moderating not only image-based illicit promotions but also in-game unsafe content. Furthermore, the evolving landscape of Virtual Reality (VR) presents both novel opportunities and challenges. We envision an extension of our work into the VR domain, adapting and enhancing our methodologies to address the special and complex safety challenges that arise in these immersive environments.

Acknowledgements

This material is based upon work supported in part by the National Science Foundation (NSF) under Grant No. 2228617, 2129164, 2120369, 2245983, 2237238, 2329704, 2112878 and a National Centers of Academic Excellence in Cybersecurity grant No. H98230-22-1-0307.

References

[1] GenieLabs. User-Generated Content (UGC): The Future of Gaming. https://medium.com/@GenieLabs./user-generated-content-ugc-the-future-of-gaming-1dfc4c41d526, 2023. Accessed: 2023-9-23.

[2] Yubo Kou and Xinning Gui. Harmful Design in the Metaverse and How to Mitigate it: A Case Study of User-Generated Virtual Worlds on Roblox. 2023.

[3] Statista. Distribution of Roblox audiences worldwide as of December 2022, by age group, 2023. Accessed: 2023-10-10.

[4] Medium. "From the Devs": How to Promote Your Game Effectively, as Told By jjwood1600, 2018. Accessed: 2023-9-20.

[5] Roblox Developer. How to market your game?, 2022. Accessed: 2023-9-20.

[6] Youtube. 3 FREE WAYS to PROMOTE your Roblox GAME and make it POPULAR, 2023. Accessed: 2023-9-20.

[7] Roblox Developer. Best free ways to advertise my game, 2021. Accessed: 2023-9-20.

[8] Reddit. Game Advertising:r/roblox, 2022. Accessed: 2023-9-20.

[9] Twitter, 2023. Accessed: 2023-09-20.

[10] Reddit. https://www.reddit.com. Accessed: 2023-09-20.

[11] Discord. https://www.discord.com. Accessed: 2023-09-20.

[12] BBC News. Roblox: The children's game with 150 million players. https://www.bbc.com/news/technology-60314572, 2023. Accessed: 2023-10-10.

[13] Cheyenne Macdonald. Outrage after 7-year-old's Roblox avatar is 'gang raped' by other players in the virtual world, 2018. Accessed: 2023-10-10.

[14] Melissa M. McDonald, Andrew M. Defever, and Carlos David Navarrete. Killing for the greater good: Action aversion and the emotional inhibition of harm in moral dilemmas. *Evolution and Human Behavior*, 38(6):770–778, 2017.

[15] Bastiaan Vanacker and Don Heider. Ethical harm in virtual communities. *Convergence*, 18(1):71–84, 2012.

[16] Jessica Wolfendale. My avatar, my self: Virtual harm and attachment. *Ethics and information technology*, 9:111–119, 2007.

[17] Hacker News. Both of my kids played a lot of Roblox until we banned it. . https://news.ycombinator.com/item?id=20622129, 2019. Accessed: 2023-9-20.

[18] Sapna M. The Dark Side of Roblox Every Parent Should Know. https://medium.com/illumination/the-dark-side-of-roblox-every-parent-should-know-93bf066b16c0, 2023. Accessed: 2023-9-23.

[19] Safety Features: Chat, Privacy & Filtering. https://en.help.roblox.com/hc/en-us/articles/203313120-Safety-Features-Chat-Privacy-Filtering. Accessed: 2023-10-12.

[20] Roblox. Safety & Civility at Roblox. https://en.help.roblox.com/hc/en-us/articles/4407444339348-Safety-Civility-at-Roblox. Accessed: 2023-10-10.

[21] Code of Conduct. https://us.battle.net/support/en/article/42673. Accessed: 2023-10-12.

[22] Rules of Conduct. https://www.swtor.com/legalnotices/rulesofconduct. Accessed: 2023-10-12.

[23] Google. Google Vision AI. https://cloud.google.com/vision/. Accessed: 2023-10-10.

[24] Clarifai. Clarifai. https://www.clarifai.com/. Accessed: 2023-10-10.

[25] Amazon. Amazon Rekognition. https://aws.amazon.com/rekognition/. Accessed: 2023-10-10.

[26] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Q Vera Liao, Yunfeng Zhang, and Chenhao Tan. Human-ai collaboration via conditional delegation: A case study of content moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2022.

[27] Microsoft. Microsoft Azure. https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/. Accessed: 2023-10-10.

[28] Yahoo. Yahoo Open NSFW. https://github.com/yahoo/open_nsfw, 2016. Accessed: 2023-10-10.

[29] Common Sense Media. Common Sense Media: Ratings, Reviews, and Advice., 2023. Accessed: 2023-09-20.

[30] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, 2023.

[31] IGN. User Generated Content in Games. https://www.ign.com/games/feature/user-generated-content. Accessed: 2023-10-16.

[32] Nicolas Ducheneaut and Robert Moore. The Social Side of Gaming: A Study of Interaction Patterns in a Massively Multiplayer Online Game. pages 360–369, 11 2004.

[33] James Grimmelmann. The virtues of moderation. *Yale JL & Tech.*, 17:42, 2015.

[34] Jialun Aaron Jiang, Peipei Nie, Jed R Brubaker, and Casey Fiesler. A Trade-off-centered Framework of Content Moderation. *arXiv preprint arXiv:2206.03450*, 2022.

[35] Ella Steen and Kathryn Yurechko and Daniel Klug. You can (not) say what you want: Using algospeak to contest and evade algorithmic content moderation on tiktok. *Social Media + Society*, 9(3):20563051231194586, 2023.

[36] Won Ik Cho and Soomin Kim. Google-trickers, Yaminjeongeum, and Leetspeak: An Empirical Taxonomy for Intentionally Noisy User-Generated Text. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 56–61, Online, November 2021. Association for Computational Linguistics.

[37] Christian Platzer, Martin Stuetz, and Martina Lindorfer. Skin sheriff: a machine learning solution for detecting explicit images. In *Proceedings of the 2nd International Workshop on Security and Forensics in Communication Systems*, SFCS '14, page 45–56, New York, NY, USA, 2014. Association for Computing Machinery.

[38] Kan Yuan, Di Tang, Xiaojing Liao, XiaoFeng Wang, Xuan Feng, Yi Chen, Menghan Sun, Haoran Lu, and Kehuan Zhang. Stealthy porn: Understanding real-world adversarial images for illicit online promotion. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 952–966. IEEE, 2019.

[39] Rashid Tahir, Faizan Ahmed, Hammas Saeed, Shiza Ali, Fareed Zaffar, and Christo Wilson. Bringing the Kid back into YouTube Kids: Detecting Inappropriate Content on Video Streaming Platforms. *2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 464–469, 2019.

[40] Kostantinos Papadamou, Antonis Papasavva, Savvas Zannettou, Jeremy Blackburn, Nicolas Kourtellis, Ilias Leontiadis, Gianluca Stringhini, and Michael Sirivianos. Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children. In *International Conference on Web and Social Media*, 2019.

[41] Weiming Hu, Ou Wu, Zhouyao Chen, Zhouyu Fu, and Stephen J. Maybank. Recognition of Pornographic Web Pages by Classifying Texts and Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:1019–1034, 2007.

[42] Farman Ali, Pervez Khan, Kashif Riaz, Daehan Kwak, Tamer Abuhmed, Daeyoung Park, and Kyung Sup Kwak. A Fuzzy Ontology and SVM–Based Web Content Classification System. *IEEE Access*, 5:25781–25797, 2017.

[43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021.

[44] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation, 2022.

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning, 2023.

[46] 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.

[47] OpenAI. GPT-4V(vision) System Card. Technical report, OpenAI, 2023.

[48] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal Chain-of-Thought Reasoning in Language Models, 2023.

[49] Jiaxin Ge, Hongyin Luo, Siyuan Qian, Yulu Gan, Jie Fu, and Shanghang Zhang. Chain of Thought Prompt Tuning in Vision Language Models, 2023.

[50] Nishant Vishwamitra, Keyan Guo, Farhan Tajwar Romit, Isabelle Ondracek, Long Cheng, Ziming Zhao, and Hongxin Hu. Moderating New Waves of Online Hate with Chain-of-Thought Reasoning in Large Language Models. In *2024 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2024.

[51] Fan Huang, Haewoon Kwak, and Jisun An. Chain of Explanation: New Prompting Method to Generate Quality Natural Language Explanation for Implicit Hate Speech. In *Companion Proceedings of the ACM Web Conference 2023*. ACM, apr 2023.

[52] Pujan Paudel, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. LAMBRETTA: Learning to Rank for Twitter Soft Moderation, 2022.

[53] Maarten Grootendorst. Keyword Extraction with BERT, 2023. Accessed: 2023-10-10.

[54] Dinh Duy Phan, Thanh Thien Nguyen, Quang Huy Nguyen, Hoang Loc Tran, Khac Ngoc Khoi Nguyen, and Duc Lung Vu. LSPD: A Large-Scale Pornographic Dataset for Detection and Classification. *International Journal of Intelligent Engineering and Systems*, 15(1), 2022.

[55] Alex Clark and Contributors. Pillow (PIL Fork). https://github.com/python-pillow/Pillow, 2023.

[56] Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms, 2011.

[57] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[58] Alex Kim. NSFW Data Scraper. https://github.com/alex000kim/nsfw_data_scraper, 2021. GitHub repository.

[59] Mazal Bethany, Andrew Seong, Samuel Henrique Silva, Nicole Beebe, Nishant Vishwamitra, and Peyman Najafirad. Towards Targeted Obfuscation of Adversarial Unsafe Images using Reconstruction and Counterfactual Super Region Attribution Explainability. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 643–660, Anaheim, CA, August 2023. USENIX Association.

[60] Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep-Learning Systems for Domain Adaptation in Computer Vision: Learning Transferable Feature Representations. *IEEE Signal Processing Magazine*, 34(6):117–129, 2017.

[61] Gabriela Csurka. *A Comprehensive Survey on Domain Adaptation for Visual Applications*, pages 1–35. Springer International Publishing, Cham, 2017.

[62] Chunna Tian, Xiangnan Zhang, Wei Wei, and Xinbo Gao. Color pornographic image detection based on color-saliency preserved mixture deformable part model. *Multimedia Tools and Applications*, 77:6629–6645, 2018.

[63] Corey H Basch, Jan Mohlman, and Charles E Basch. An assessment of violent imagery in advertisements on city buses in Manhattan, New York City. *Health Promotion Perspectives*, 10(2):162–165, 2020.

[64] Wikipedia. Graphic violence. https://en.wikipedia.org/wiki/Graphic_violence, 2023. Accessed: 2023-10-11.

[65] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

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

[67] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks, 2019.

[68] Wikipedia. Not safe for work — Wikipedia, The Free Encyclopedia, 2023. Online; accessed 2023-10-15.

[69] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection, 2016.

[70] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In *Computer Vision – ECCV 2016*, pages 21–37. Springer International Publishing, 2016.

[71] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2016.

[72] Alara Dirik and Sayak Paul. A Dive into Vision-Language Models, 2023. Accessed: 2023-10-15.

[73] Top Roblox Servers on Discord, 2023. Accessed: 2023-10-10.

[74] OpenAI. ChatGPT, 2023.

[75] James Gallagher and Piotr Skalski. First Impressions with GPT-4V(ision). 2023. Accessed: 2023-10-15.

[76] Indish Marketer. GPT-4 Vision Access in ChatGPT! Full Tour & Impressive Results! 2023. Accessed: 2023-10-15.

[77] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 2015.

[78] Minecraft. Video Game, 2011. Available: https://www.minecraft.net/.

[79] Terasology. http://terasology.org/. Accessed: 2023-9-20.

[80] Lego worlds. https://www.lego.com/en-us/kids/games/videogames/lego-worlds-85e9a7d2143a4e6d8f1a2c4a36b3b14d. Accessed: 2023-10-8.

A Hashtags for UGCG Image Collection

Here we provide the complete list of hashtags we used for UGCG image data collection in Table 7.

Hashtags
#RobloxUGC, #RobloxCondo, #RobloxSex #RobloxR34, #RobloxR63, #RobloxF**k, #RobloxPorn, #rr34, #rr63, #rosex, #legosex, #robloxlewd, #robloxlewdsex, #robloxnsfw #RobloxKiller, #RobloxBlood, #RobloxMurderMystery, #RobloxMurderMystery2, #RobloxMM, #RobloxMM2, #RobloxPiggy, #RobloxWar, #RobloxWarrior, #RobloxCombatWarrior, #RobloxWelcomeToPhantom, #RobloxWelcome2Phantom, #RobloxW2P, #RobloxArenal, #RobloxBattleGround

Table 7: List of hashtags.

B UGCG Image Annotation Codebook

We provide the codebook used for annotating the UGCG image data collections, depicted in Table 8.

Codebook for Illicit Promotional Images of UGCGs
Are there any characters or avatars in this image?
Are the characters or avatars naked or are any private parts exposed in this image?
Are the characters or avatars indulging in any sexual activities?
Are any knives, guns, or deadly weapons presented in this image?
Are the characters or avatars indulging in any violent activities?
Are the characters or avatars indulging in any bullying activities?
Is there any scam content in this image?
Are the characters or avatars indulging in any other harmful activities?

Table 8: Codebook for illicit promotional images of UGCGs.