ARTIFACT EVALUATED
usenix ASSOCIATION
AVAILABLE

ARTIFACT EVALUATED
usenix ASSOCIATION
FUNCTIONAL

ARTIFACT EVALUATED
usenix ASSOCIATION
REPRODUCED

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**[*]

[*] UB University at Buffalo
The State University of New York

[†] UTSA
The University of Texas
at San Antonio™

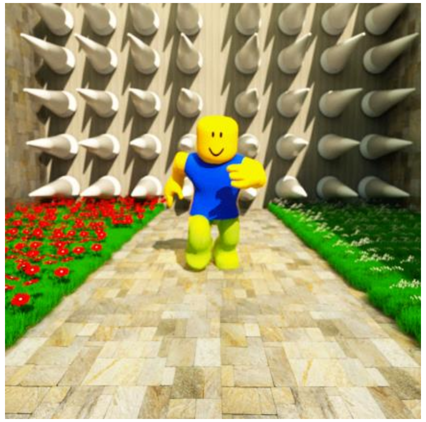
[‡] CLEMSON
UNIVERSITY

[§] N Northeastern
University

Disclaimer: This presentation contains sensitive images that could be disturbing to some members of the audience

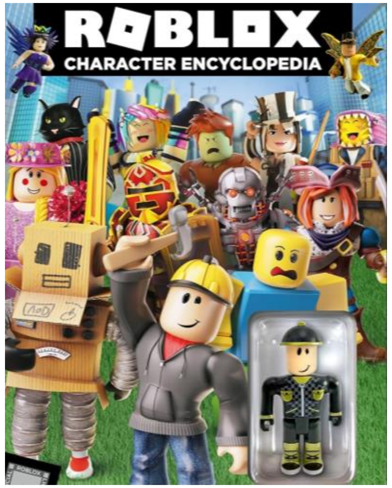
User Generated Content Games (UGCGs)

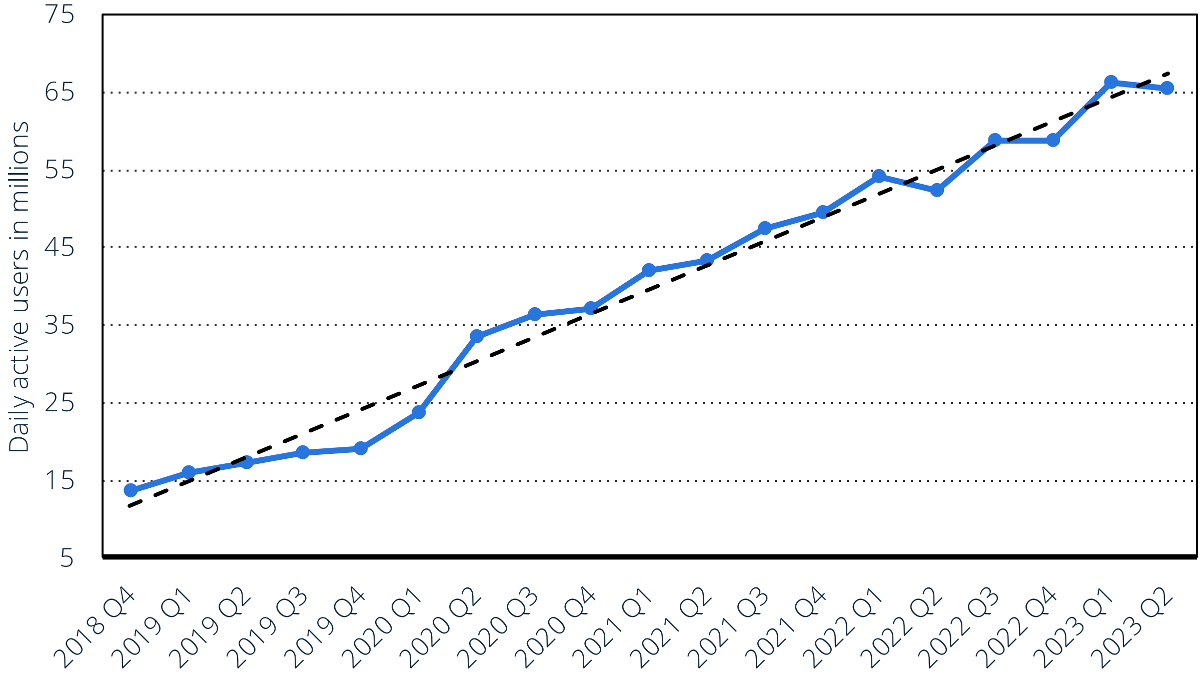
- UGCGs are video games that allow players to create, modify, and share their own content within the games



Growth of UGCGs

- **Roblox**, as one of the most popular UGCG platforms, has experienced continuous growth in its online user base



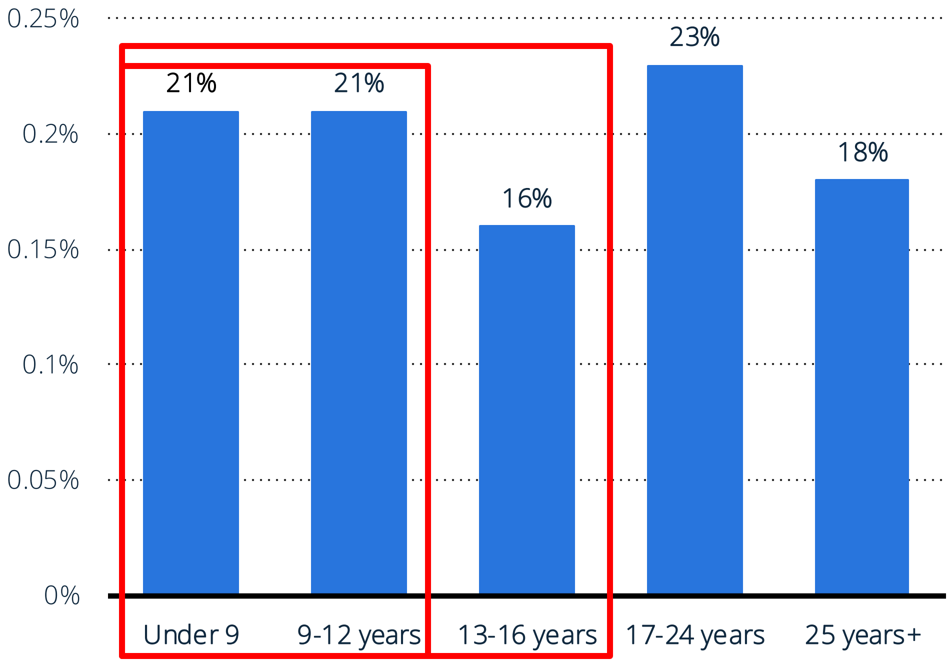


DAU of Roblox games worldwide from 4th quarter 2018 to 2nd quarter 2023

https://www.statista.com/statistics/1192573/daily-active-users-global-roblox/

UGCG Among Children

- UGCGs on Roblox are predominantly used by **children** and **adolescents**



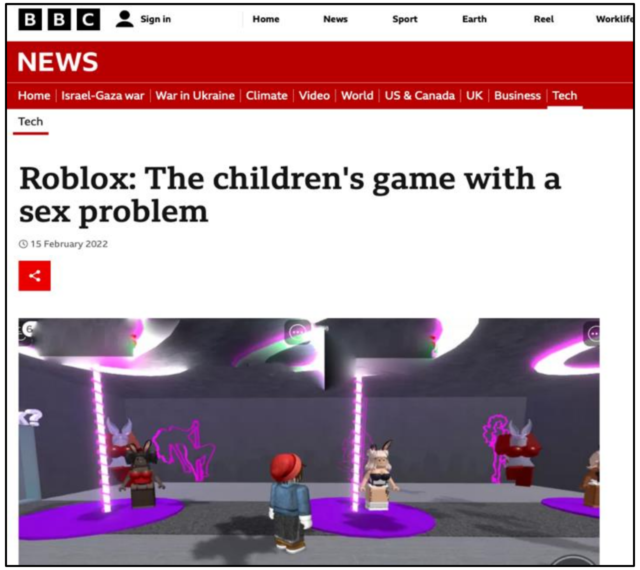
Distribution of Roblox audiences worldwide as of December 2023, by age group

https://www.statista.com/statistics/1190869/roblox-games-users-global-distribution-age/

- **58%** of its user base is under 16 years old
- with a substantial **42%** comprising children who are **under 13 years old**

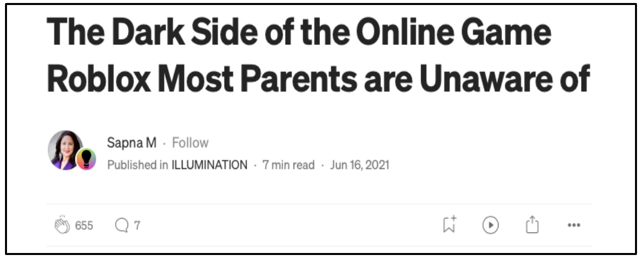
The Dark Side of UGCGs

- Sex, Violence, Scams, and bullying…


BBC
Sign in
Home
News
Sport
Earth
Reel
Worklife
NEWS
Home | Israel-Gaza war | War in Ukraine | Climate | Video | World | US & Canada | UK | Business | Tech
Tech

Roblox: The children's game with a sex problem

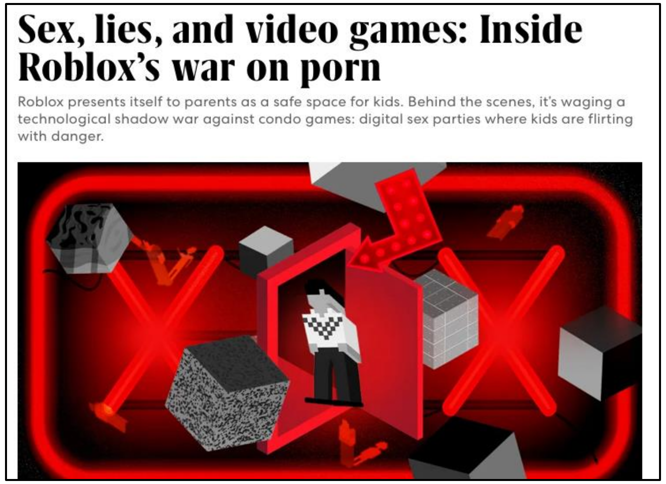
15 February 2022



The Dark Side of the Online Game Roblox Most Parents are Unaware of

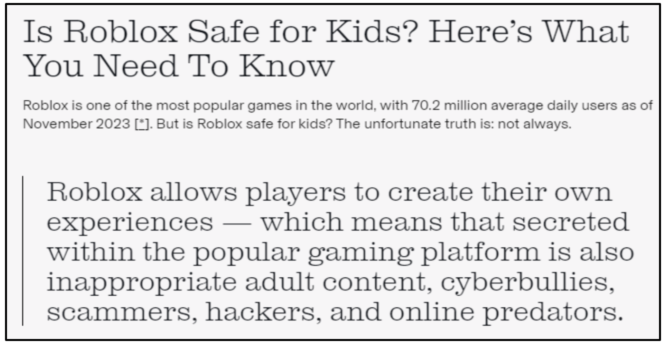
Sapna M · Follow

Published in ILLUMINATION · 7 min read · Jun 16, 2021



Sex, lies, and video games: Inside Roblox's war on porn

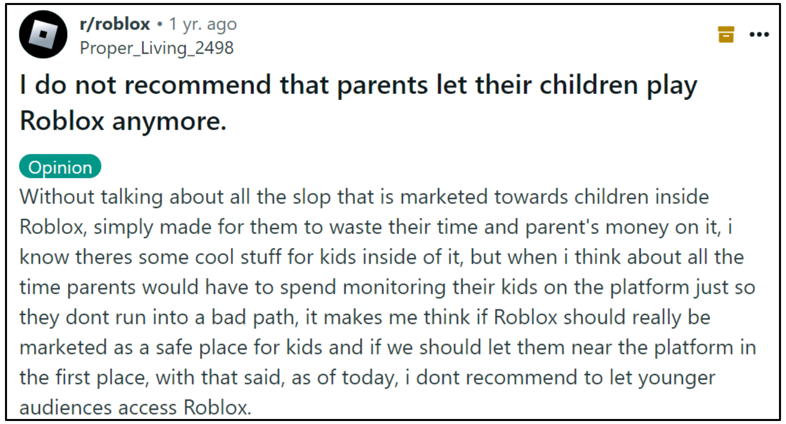
Roblox presents itself to parents as a safe space for kids. Behind the scenes, it's waging a technological shadow war against condo games: digital sex parties where kids are flirting with danger.



Is Roblox Safe for Kids? Here's What You Need To Know

Roblox is one of the most popular games in the world, with 70.2 million average daily users as of November 2023 [*]. But is Roblox safe for kids? The unfortunate truth is: not always.

> Roblox allows players to create their own experiences — which means that secreted within the popular gaming platform is also inappropriate adult content, cyberbullies, scammers, hackers, and online predators.

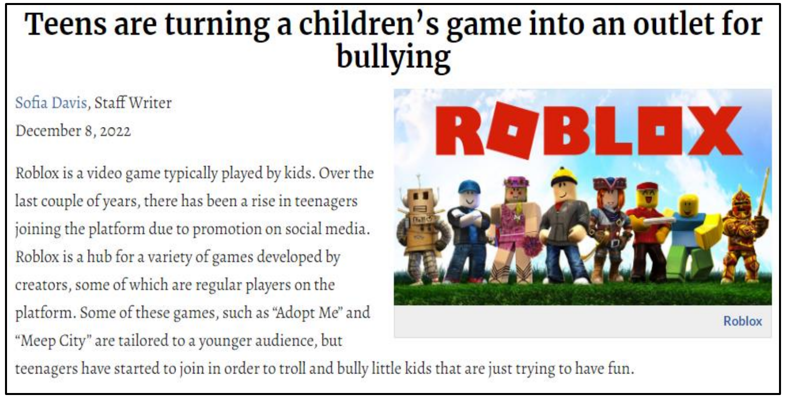


r/roblox • 1 yr. ago
Proper_Living_2498

I do not recommend that parents let their children play Roblox anymore.

Opinion

Without talking about all the slop that is marketed towards children inside Roblox, simply made for them to waste their time and parent's money on it, i know theres some cool stuff for kids inside of it, but when i think about all the time parents would have to spend monitoring their kids on the platform just so they dont run into a bad path, it makes me think if Roblox should really be marketed as a safe place for kids and if we should let them near the platform in the first place, with that said, as of today, i dont recommend to let younger audiences access Roblox.



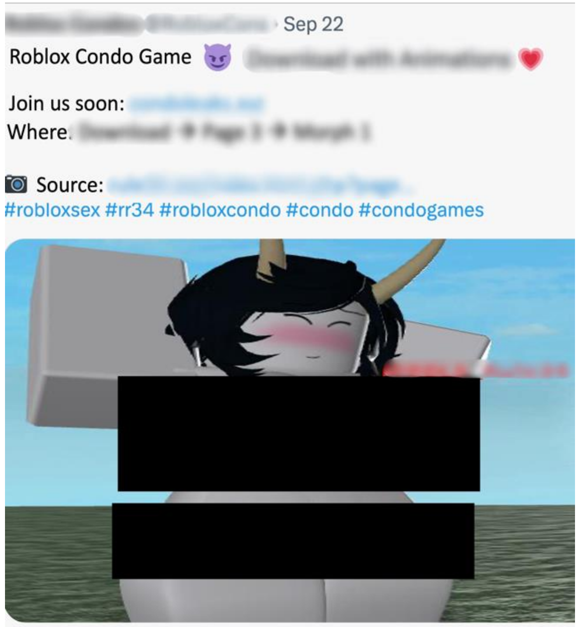
Teens are turning a children's game into an outlet for bullying

Sofia Davis, Staff Writer

December 8, 2022

Roblox is a video game typically played by kids. Over the last couple of years, there has been a rise in teenagers joining the platform due to promotion on social media. Roblox is a hub for a variety of games developed by creators, some of which are regular players on the platform. Some of these games, such as "Adopt Me" and "Meep City" are tailored to a younger audience, but teenagers have started to join in order to troll and bully little kids that are just trying to have fun.

Illicit Promotion of Unsafe UGCGs

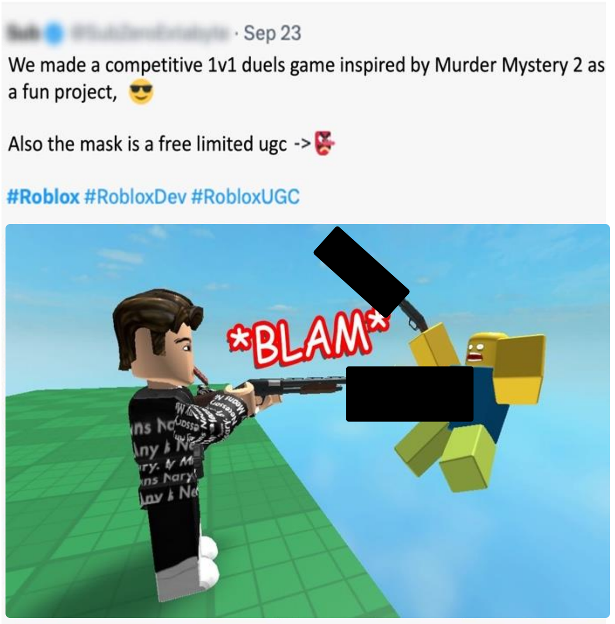


[redacted] · Sep 22

Roblox Condo Game 😈 [redacted] ❤️

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

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



[redacted] · Sep 23

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

Also the mask is a free limited ugc -> 👺

#Roblox #RobloxDev #RobloxUGC

- Being prevalent on social media platforms such as X, Reddit, Discord, etc
- Using unsa e UGCGs with benig
- Rarely mod or even warned

Many children were recruited in unsafe UGCGs by viewing such illicit online image promotions

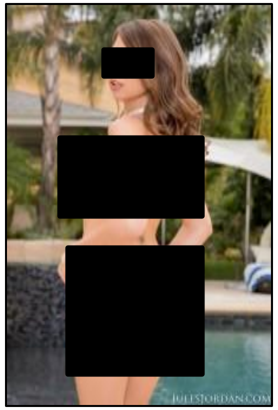
Illicit Promotion of Unsafe UGCGs

- UGCG Image Dataset
  - Used hashtags identified in self-reported stories gathered from Common Sense Media
  - Collected from X since 01/01/2020 to 12/31/2022
  - 38,182 tweets with images
  - 2,924 valid UGCG images in 4,000 randomly picked images
    - **1,621** Sexually explicit images
    - **202** violent images
    - **1,101** Safe images

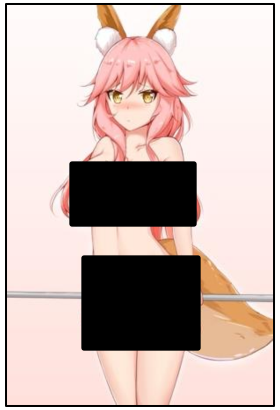
Common Sense Media – Roblox: https://www.commonsensemedia.org/website-reviews/roblox/user-reviews/adult

Can We Use Existing Unsafe Image Detectors?

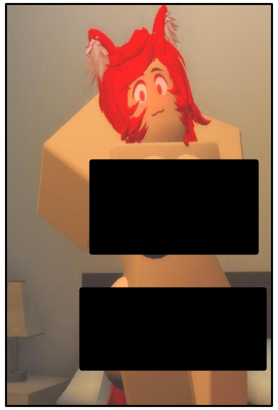
- Comparison of three different unsafe image datasets with the five state-of-the-art unsafe image detectors



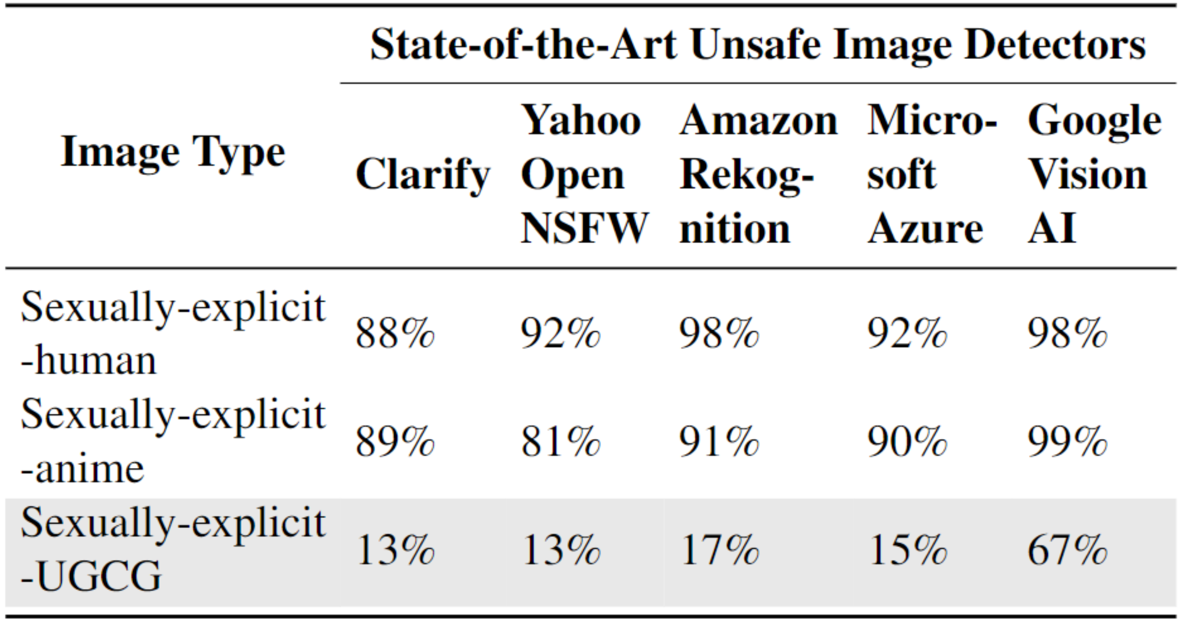
Sexually-explicit-human



Sexually-explicit-anime



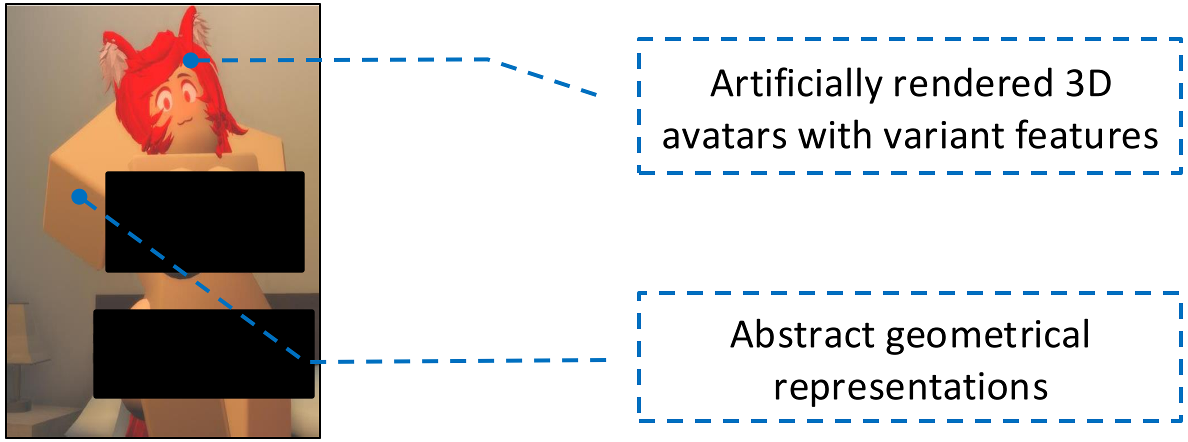
Sexually-explicit-UGCG



	State-of-the-Art Unsafe Image Detectors				
Image Type	Clarify	Yahoo Open NSFW	Amazon Rekognition	Microsoft 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%

Challenges in Detecting Unsafe UGCG Images

- Challenge 1: Limited Training Data
  - No large-scale training dataset of unsafe UGCG images that can be used by existing detectors
- Challenge 2: Complex Context
  - Unsafe UGCG images are very different from traditional unsafe images



Artificially rendered 3D avatars with variant features

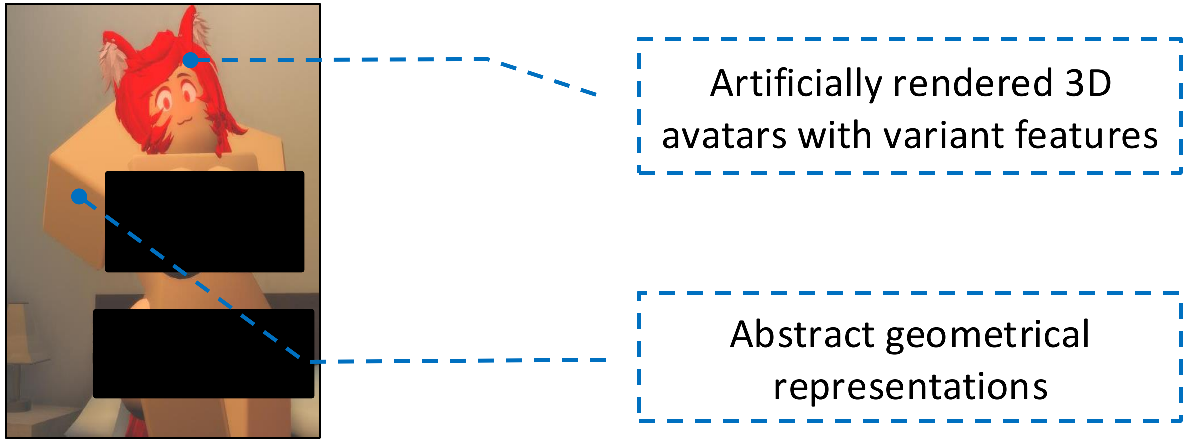
Abstract geometrical representations

Using Large Vision-Language Models (VLMs)

- Challenge 1: Limited Training Data
  - No large-scale training dataset of unsafe UGCG images that can be used by existing detectors
- Challenge 2: Complex Context
  - Unsafe UGCG images are very different from traditional unsafe images

Zero-/few-shot learning capabilities

Reasoning capabilities


[redacted]
[redacted]

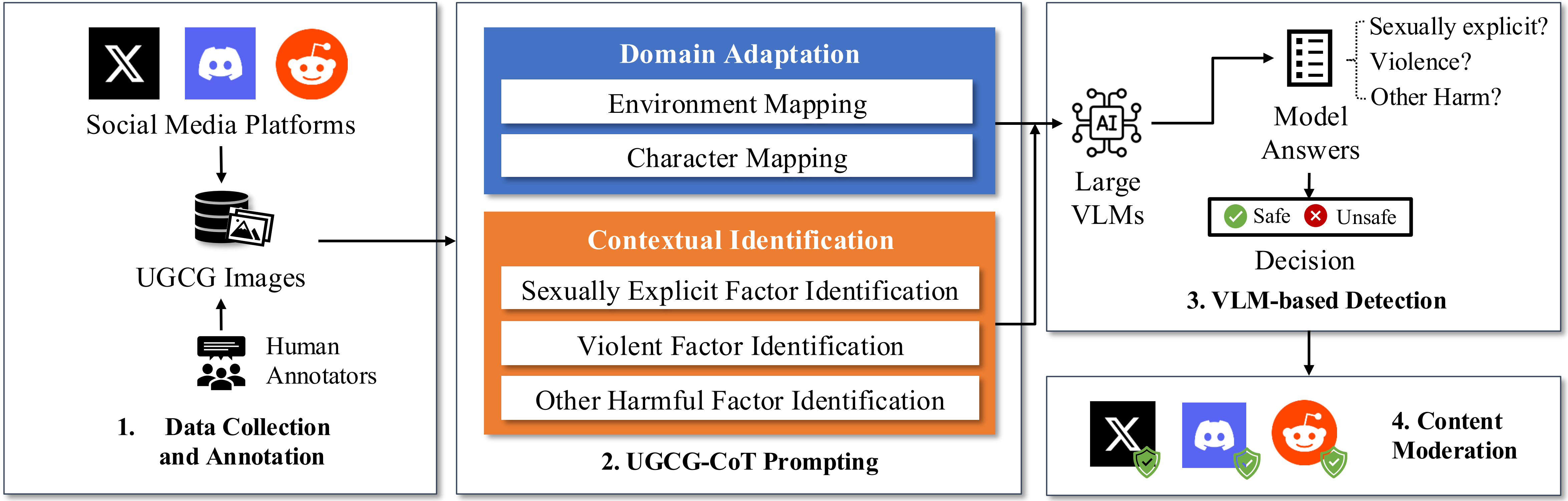
Artificially rendered 3D avatars with variant features

Abstract geometrical representations

InstructBLIP

GPT-4V

UGCG-GUARD



Social Media Platforms

UGCG Images

Human Annotators

1. Data Collection and Annotation

Domain Adaptation

Environment Mapping

Character Mapping

Contextual Identification

Sexually Explicit Factor Identification

Violent Factor Identification

Other Harmful Factor Identification

2. UGCG-CoT Prompting

Large VLMs

Model Answers

Sexually explicit?

Violence?

Other Harm?

✓ Safe ✗ Unsafe

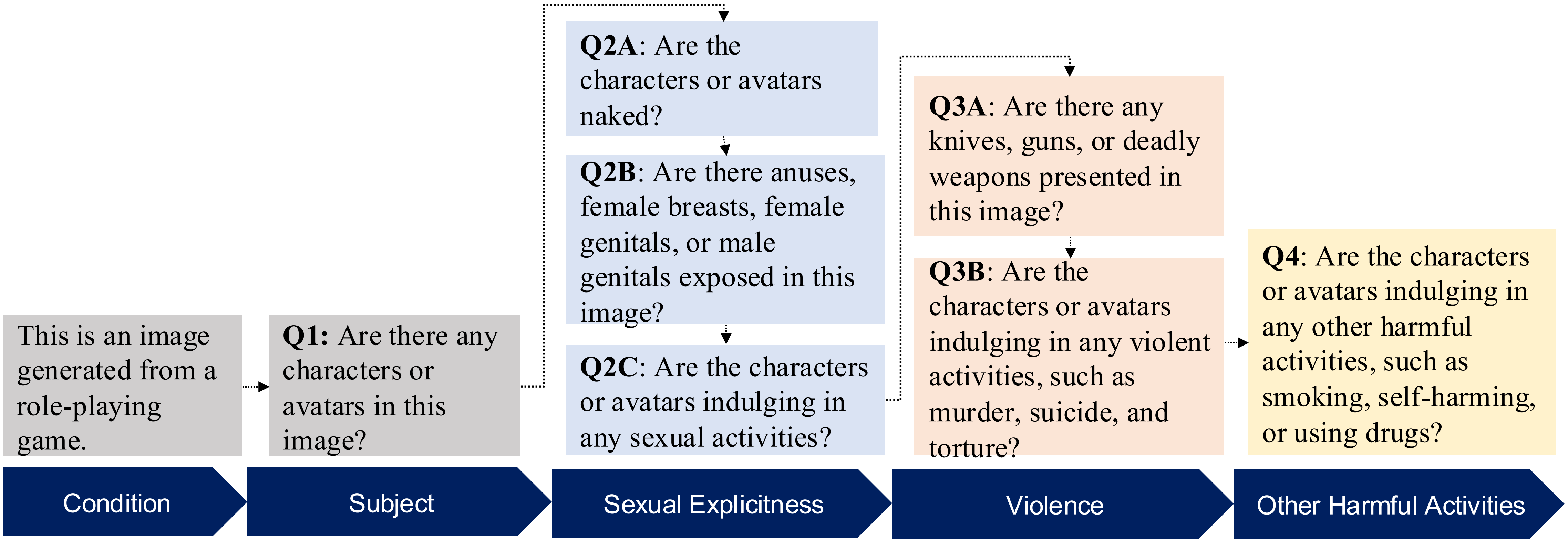
Decision

3. VLM-based Detection

4. Content Moderation

UGCG-CoT

- A Chain-of-Thought (CoT) reasoning-based prompting strategy



This is an image generated from a role-playing game.

Q1: Are there any characters or avatars in this image?

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?

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?

Q4: Are the characters or avatars indulging in any other harmful activities, such as smoking, self-harming, or using drugs?

Condition

Subject

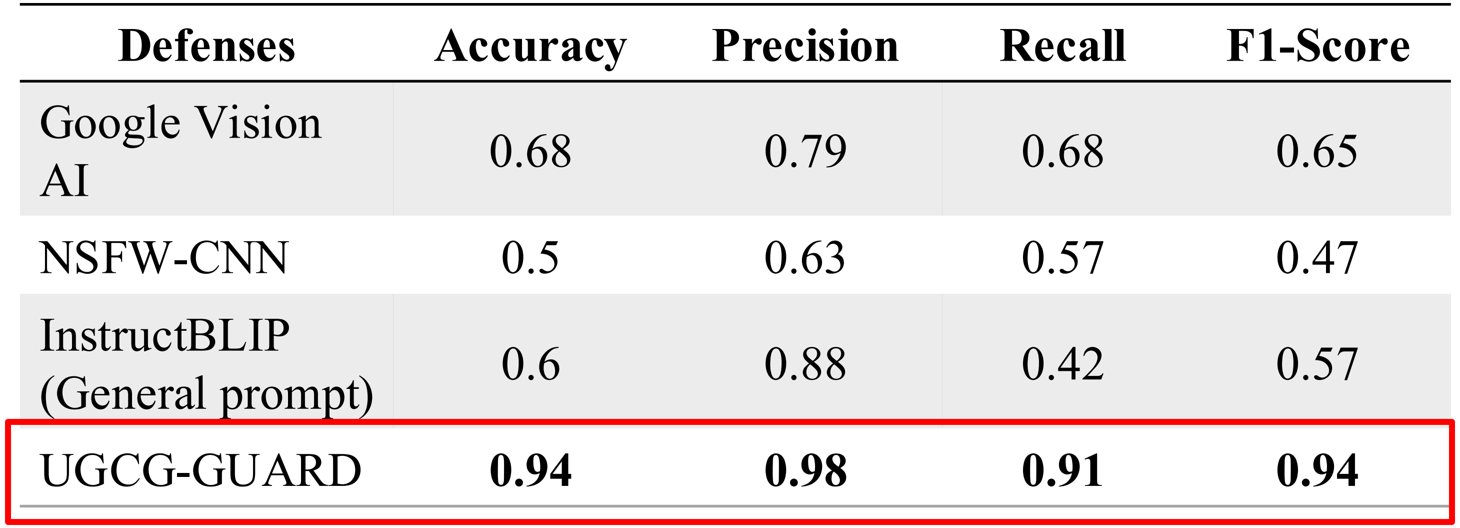
Sexual Explicitness

Violence

Other Harmful Activities

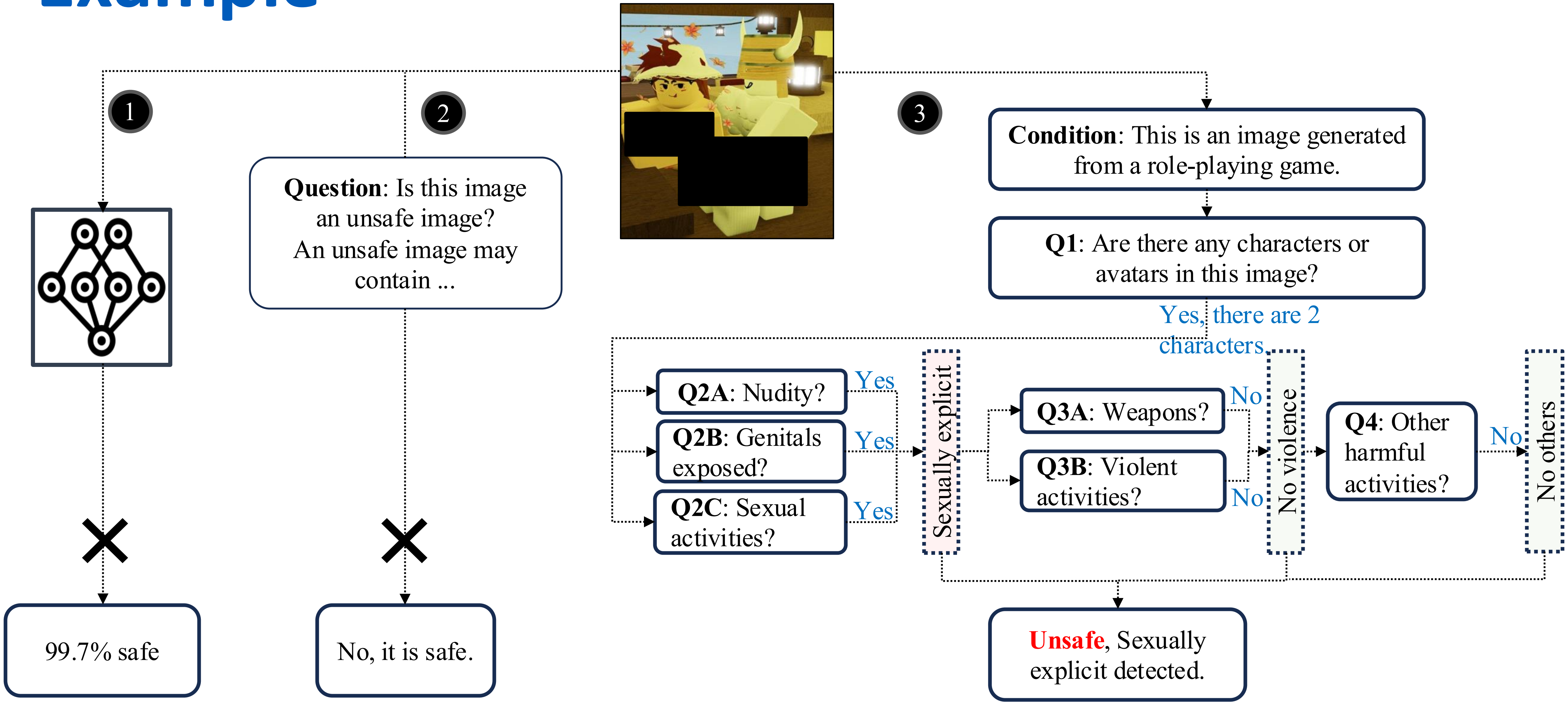
Evaluation

- UGCG-Guard vs. Baselines



Defenses	Accuracy	Precision	Recall	F1-Score
Google Vision AI	0.68	0.79	0.68	0.65
NSFW-CNN	0.5	0.63	0.57	0.47
InstructBLIP (General prompt)	0.6	0.88	0.42	0.57
UGCG-GUARD	**0.94**	**0.98**	**0.91**	**0.94**

Example



[redacted]

1

99.7% safe

2

Question: Is this image an unsafe image? An unsafe image may contain ...

No, it is safe.

3

Condition: This is an image generated from a role-playing game.

Q1: Are there any characters or avatars in this image?

Yes, there are 2 characters.

Q2A: Nudity? Yes

Q2B: Genitals exposed? Yes

Q2C: Sexual activities? Yes

Sexually explicit

Q3A: Weapons? No

Q3B: Violent activities? No

No violence

Q4: Other harmful activities? No

No others

Unsafe, Sexually explicit detected.

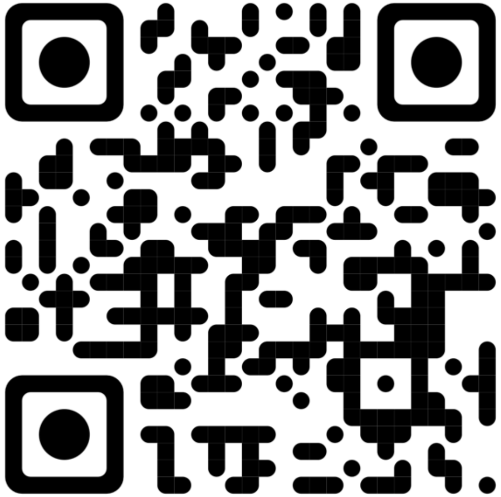
UGCG-CoT for unsafe UGCG image decision-making (3) compared to traditional unsafe image detection model (1) and LVLM with general prompting (2)

Evaluation

- “In-the-Wild”

Detectors	**Accuracy**		**Precision**		**Recall**		**F1-Score**	
	Reddit	Discord	Reddit	Discord	Reddit	Discord	Reddit	Discord
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	**0.93**	0.79	0.91	0.88

Available Online!



Paper

Code

Dataset

Conclusion and Future Work

- Conclusion
  - A comprehensive study to understand the threat of illicit image promotion for unsafe UGCGs
  - Examining the capabilities of the existing detection tools
  - A novel framework to address the problem of illicit image promotion for unsafe UGCGs
- Future work
  - Multi-platform UGCGs
  - In-game unsafe content moderation
  - Unsafe UGCGs in Virtual Reality (VR) environment

Thank you !

Keyan Guo,
Ph.D. Candidate

keyanguo@buffalo.edu

University at Buffalo,
Buffalo, NY,
United States