Improving Detection and Understanding the Effects of AI-Generated Videos

Margie Ruffin[1], Khushi Patel[2], Samika Karumuri[2], Gang Wang[2]
[1]*Spelman College*
[2]*University of Illinois at Urbana-Champaign*
margieruffin@spelman.edu, {khuship4, samikak2, gangw}@illnios.edu

***Abstract*—Recently, social media has seen a significant uptick in the spread of AI-generated content, especially videos. Videos generated by high-quality models make it difficult for humans and detectors to distinguish between real and fake content, which may open the door to misuse and abuse. We conduct preliminary data collection from TikTok and analysis to set the stage for a two-pronged research approach. Our proposed research directions will (1) systematically evaluate and potentially advance diffusion-based AI video detection models and (2) examine the potential harms, if any, associated with viewing AI-generated videos.**

***Index Terms*—AI-Generated Media, Detection, Harm, Abuse**

1. Introduction

In 2019, the first AI-generated video, created by NVIDIA's GauGAN model [17], was released online. Following that release, in 2022, major technology companies, including Facebook and Google, revealed their own sophisticated text-to-video research models, but they were not made available to the public immediately. Runway ML made text-to-video generation widely accessible with the beta release of its diffusion-based Gen-1 model in early 2023 [8]. Since then, we have seen the release of even higher-quality models capable of producing photorealistic cinematic footage and audio, including OpenAI's Sora [16], Google Veo [1], and Pika Labs [10]. However, the emergence of these high-fidelity diffusion models has made their outputs increasingly difficult to detect which compromises our ability to moderate this content.

Internet platforms rely on removal, labeling, and age restrictions to curb abuse and disinformation [6], yet the complexity of moderation limits the accuracy of AI systems and increases the burden on human reviewers. Effective moderation requires both (1) more accurate detection of synthetic content and (2) a deeper understanding of the harms that exposure to such content may cause. Improving detection depends on developing and fine-tuning models using unbiased, realistic data, while understanding harms requires examining how viewers perceive and are affected by AI-generated videos. To address these needs, we propose two complementary research directions: **RD1** advances diffusion-based AI video detection models, and **RD2** investigates the potential harms associated with viewing AI-generated videos.

Search Tags	Synthetic (%)	Real (%)	Total Count (%)
AI_Generated_Election	96 (2.92%)	694 (21.07%)	790 (23.99%)
Kamala_Harris_AI	56 (1.70%)	867 (26.33)	923 (28.03%)
Trump_And_Elon	135 (4.10%)	678 (20.59%)	813 (24.69%)
Trump_Biden_AI	198 (6.01%)	569 (17.13%)	767 (23.29%)
Total	**485 (14.73%)**	**2,808 (85.27%)**	**3,293 (100%)**

TABLE 1. UNIQUE GEN-AI VIDEOS COLLECTED FROM TIKTOK BETWEEN THE 2024 U.S. PRESIDENTIAL ELECTION AND THE 2025 PRESIDENTIAL INAUGURATION, CATEGORIZED BY TYPE (SYNTHETIC OR REAL) AND SEARCH TAG.

2. Preliminary Work

To begin these proposed research projects, we conducted longitudinal data collection on TikTok from the 2024 United States Presidential Election through the 2025 Presidential Inauguration. TikTok is a video-first platform that allows users to upload their own content, like and comment on others' content, and share others' content. While other platforms have experienced an influx of synthetic videos, we focus on TikTok because of its tag-based search and its popularity among content creators. We developed a Selenium-based web crawler [14] to gather the videos and their metadata.

First Labeling Task. The United States (U.S.) presidential election is often a highly contentious period during which public figures face criticism, parody, and praise. We focus on this time frame to capture a diverse dataset of real and synthetic election-related content. The search tags we selected, *AI_Generated_Election*, *Kamala_Harris_AI*, *Trump_And_Elon*, and *Trump_Biden_AI*, were chosen to target key political actors (Joe Biden, Kamala Harris, and Donald Trump), as well as Elon Musk, given his prominent public alignment with Trump at the time. Some tags include the "AI" qualifier because we specifically aimed to capture AI-generated videos. Given the millions of daily TikTok uploads [11] and the absence of reliable platform-applied labels, narrowing the collection to videos containing the keyword "AI" was necessary to feasibly identify potentially synthetic content. We crawled TikTok from December 23rd, 2024, through January 24th, 2025 (business days only), for

a total of 24 days. Although TikTok's Terms of Service discourage web scraping [22], it was necessary to collect data at the scale required for this study and is consistent with prior work examining potentially abusive platform behaviors [4], [24]. To protect privacy and minimize server impact, we filtered personal information and used a slow, rate-limited crawling process [13].

In total, we collected $3,293$ videos along with their captions, IDs, and uploader information. A team of three researchers split up and manually coded each video as *real* or *synthetic* (AI-generated), applying a conservative best-judgment approach as distinguishing between the two is becoming an increasingly difficult task. Videos were labeled "synthetic" if they exhibited dreamlike or hyper-realistic characteristics [7], [20]. According to this scheme, 14.73% of the collected videos were AI-generated: 2.92% from the *AI_Generated_Election* tag, 1.70% from *Kamala_Harris_AI*, 4.10% from *Trump_And_Elon*, and 6.01% from *Trump_Biden_AI*. See Table 1 for a full breakdown by search tag.

Second Labeling Task. Following the release of publicly available text-to-video Generative AI models, social media platforms saw an influx of synthetic content. So much so that in 2024, several of them announced the rollout of their own "AI-Generated Video Detectors." In several press releases, social media platforms such as YouTube, Meta, and TikTok announced they would begin labeling synthetic videos and images as "AI-generated" [5], [15], [21]. With our focus remaining on TikTok, we found that none of the videos we collected had the "AI-generated" label. A further inspection revealed that, at the time, videos were labeled only if the uploader applied the label. This observation motivates a closer look at how platforms govern content.

TikTok's content moderation is governed by its community guidelines. In 2024, TikTok updated the Integrity and Authenticity section of its guidelines to include Edited Media and AI-Generated Content (AIGC) [23]. These guidelines detail what is allowed on the platform and expressly state what is not. Three researchers developed a codebook based on TikTok's Integrity and Authenticity guidelines, focusing on the Misinformation and AIGC subsections. An initial round of independent coding yielded substantial intercoder agreement (Fleiss's Kappa = 0.62), after which disagreements were resolved and definitions refined. The final codebook produced four primary codes, *Misinformation*, *AIGC*, *AI-Mixed Media*, and *Non-violation*, and 19 subcodes (8 under Misinformation and 11 under AIGC).

We applied this codebook to the 485 videos labeled as synthetic from the first labeling task. This task revealed that 343 videos fell into one of 11 AIGC subcategories and 8 videos fell into one of 8 misinformation subcategories. These 351 out of the 485 videos violate TikTok's community guidelines. Given the remaining 134 videos, which do not TikTok's community guidelines 2 of the videos belonged to the AI-Mixed Media category, and 132 videos belonged to the Non-violation category. A full breakdown of the codebook can be seen in Table 2 in the Appendix.

3. RD1: Improving AI-Video Detection

Earlier text-to-video Gen AI models would leave behind visual artifacts that would help a viewer determine if what they were seeing was indeed real or not [9], [12]. That is no longer always the case, as visual clues are fading. The increasing quality and variety of models make it difficult for human moderators and detectors to distinguish between fake and real with sufficient accuracy. Detectors in particular struggle with this task because many are trained and tested on synthetic videos explicitly created for that task, rather than on videos collected *from the wild* [2].

To address these gaps, we must first identify the capabilities and limitations of the detection models. To do so, we plan to perform a systematic benchmark of AI-generated video detection techniques. Using our current dataset and additional videos that we plan to collect to create a more robust testing and training set, we will assess the accuracy, precision, and recall of state-of-the-art diffusion-based video classification models. One such technique is LAVID, an Agentic LVLM Framework [12]. Once the gaps, if any, relative to our benchmark are identified, we will work to close them by iteratively refining the most successful approaches.

4. RD2: Understanding AI Video Harms

TikTok's AI-generated content policy acknowledges that such media can still cause harm even when properly labeled [19], yet no structured harm framework has been applied to evaluate the effects of viewing AI-generated content online. Prior work shows that harmful material is regularly shared on social platforms [18]. They also agree that the harmful content focuses on specific sub-problems and varies by platform [3]. Using the videos we have collected, we aim to assess what harms may occur and how severe they are.

We plan to consider existing harm frameworks and develop one focused on AI that quantifies the types of harm a viewer might encounter and their severity in our collection of videos and their ground-truth labels. Alongside captions and comments from shared videos, a deeper understanding of the harms viewers may face when encountering unmoderated media could help improve moderation practices.

5. Conclusion

With much of the data collection complete, the proposed two-pronged research approach will further the study of AI-generated videos, which have experienced rapid growth and are likely to remain prevalent. Upon completion of these research tasks, we expect to identify methods to improve the detection of AI-generated content using diffusion models and address gaps in current moderation practices. Even when detected and properly labeled, these videos can still cause harm; by understanding the nature and severity of those harms, content moderators can prioritize harmful content and create safer platforms.

References

[1] Google Veo 3.1. Gemini AI video generator powered by Veo 3.1. https://gemini.google/overview/video-generation/, 2026.

[2] Mubarak Alrashoud. Deepfake video detection methods, approaches, and challenges. *Alexandria Engineering Journal*, 2025.

[3] Arnav Arora, Preslav Nakov, Momchil Hardalov, Sheikh Muhammad Sarwar, Vibha Nayak, Yoan Dinkov, Dimitrina Zlatkova, Kyle Dent, Ameya Bhatawdekar, Guillaume Bouchard, and Isabelle Augenstein. Detecting harmful content on online platforms: What platforms need vs. where research efforts go. *ACM Computing Surveys*, 2023.

[4] Muhammad Abu Bakar Aziz and Christo Wilson. Johnny Still Can't Opt-out: Assessing the IAB CCPA Compliance Framework. In *Proc. of PETS*, 2024.

[5] Monika Bickert. Our Approach to Labeling AI-Generated Content and Manipulated Media. https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/, 2024.

[6] Daria Dergacheva, Vasilisa Kuznetsova, Rebecca Scharlach, and Christian Katzenbach. One Day in Content Moderation: Analyzing 24 h of Social Media Platforms' Content Decisions through the DSA Transparency Database. *Centre for Media, Communication and Information Research (ZeMKI)*, 2023.

[7] Pandora Dewan. Can you guess which of these images were made by AI? https://www.newsweek.com/guess-images-artificial-intelligence-midjourney-machine-learning-1807051, 2023.

[8] Will Douglas Heaven. The original startup behind stable diffusion has launched a generative AI for video. https://www.technologyreview.com/2023/02/06/1067897/runway-stable-diffusion-gen-1-generative-ai-for-video/, 2023.

[9] Rebecca Heilweil. Sora is even fooling human deepfake detectors. https://www.fastcompany.com/91428173/sora-human-deepfake-detectors-openai-sam-altman.

[10] Pika Labs. Pika AI Generation Model. https://pika.art/login, 2026.

[11] Robert A. Lee. TikTok usage statistics 2025: User growth, engagement, and more. https://sqmagazine.co.uk/tiktok-usage-statistics/, 2026.

[12] Qingyuan Liu, Yun-Yun Tsai, Ruijian Zha, Victoria Li, Pengyuan Shi, Chengzhi Mao, and Junfeng Yang. LAVID: An agentic LVLM framework for diffusion-generated video detection. http://arxiv.org/abs/2502.14994, 2025.

[13] Alan Mislove and Christo Wilson. A Practitioner's Guide to Ethical Web Data Collection. In *The Oxford Handbook of Networked Communication*. 2020.

[14] Baiju Muthukadan. Selenium with Python — Selenium Python Bindings 2 documentation. https://selenium-python.readthedocs.io/, 2025.

[15] Jennifer Flannery O'Connor and Emily Moxley. Our approach to responsible AI innovation. https://blog.youtube/inside-youtube/our-approach-to-responsible-ai-innovation/, 2023.

[16] OpenAI. Sora is here. https://openai.com/index/sora-is-here/, 2025.

[17] Isha Salian. Stroke of genius: GauGAN turns doodles into stunning, photorealistic landscapes. https://blogs.nvidia.com/blog/gaugan-photorealistic-landscapes-nvidia-research/, 2019.

[18] Gautam Kishore Shahi and William Kana Tsoplefack. Mitigating harmful content on social media using an interactive user interface. In *Social Informatics*, 2022.

[19] TikTok Support. About AI-generated content. https://support.tiktok.com/en/using-tiktok/creating-videos/ai-generated-content, 2026.

[20] Anna Louie Sussman. 5 telltale signs that a photo is AI-generated. https://insight.kellogg.northwestern.edu/article/ai-photos-identification, 2024.

[21] TikTok. New labels for disclosing AI-generated content. https://newsroom.tiktok.com/en-us/new-labels-for-disclosing-ai-generated-content, 2019.

[22] TikTok. Terms of service | TikTok. https://www.tiktok.com/legal/page/us/terms-of-service/en, 2026.

[23] TikTok. TikTok Community Guidelines | Overview. https://www.tiktok.com/community-guidelines/en/integrity-authenticity#3, 2026.

[24] Christo Wilson, Bryce Boe, Alessandra Sala, Krishna P.N. Puttaswamy, and Ben Y. Zhao. User interactions in social networks and their implications. In *Proc. of EUROSYS*, April 2009.

Acknowledgments

This work was supported by the NSF Graduate Research Fellowship Program under Grant No. 21-46756 and the Mastercard Impact Fund by way of the Atlanta University Center's Data Science Initiative.

Appendix

The release of statements that alluded to the labeling of AI-generated videos on TikTok's platform prompted an investigation into how the platform was governed. According to its Community Guidelines, videos that aren't marked as "AI-generated" by the uploader are subject to removal under these policies. Researchers use these guidelines to develop a codebook to quantify which videos remain on the platform that violate the policies. The codebook contains primary codes and subcodes. All videos have a primary label that speaks to their main theme. Some videos have a secondary label because they contain more than one theme. See Table 2 for a full breakdown of codes.

Primary Code	Subcode	Primary Label Freq.	Secondary Label Freq.
Non-Violation	–	132	0
Misinformation	Climate change misinformation	0	0
	Conspiracy theories that are violent or hateful	0	0
	Conspiracy theories that name and attack individual people	4	1
	Health misinformation	0	0
	Misrepresenting authoritative sources	0	0
	Poses a risk to public safety	0	0
	Unverified claims (that may cause panic) related to an emergency or unfolding event	1	1
	Other Misinfo (please explain)	3	2
AI-Generated Content (AIGC)	Shares or shows fake authoritative sources	0	0
	Shares or shows crisis events	0	0
	Falsely shows public figures in certain contexts	312	0
	Likeness of adult private figure (without permission)	0	0
	Realistic-appearing people under the age of 18	1	1
	Misleading AIGC or edited media that falsely shows content made to seem as if it comes from an authoritative source	0	0
	Misleading AIGC or edited media that falsely shows a crisis event, such as a conflict or natural disaster	1	0
	Misleading AIGC or edited media public figure who is being degraded or harassed, or engaging in criminal or anti-social behavior	0	0
	Misleading AIGC or edited media public figure who is taking a position on a political issue, commercial product, or a matter of public importance	2	0
	Misleading AIGC or edited media public figure who is being politically endorsed or condemned by an individual or group	0	0
	Other AGIC (please explain)	27	1
AI-Mixed Media	–	2	0
Total		**485**	**6**

TABLE 2. **TikTok Community Guidelines Codebook** — We show the primary codes, subcodes, and their frequencies. All videos have a primary label that speaks to their main theme. Some videos have a secondary label because they contain more than one theme.