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 Urbana-Champaign

ConPro 2026 Workshop
Thursday, May 21, 2026

Our Overarching Problem

AI-generated video evolution is outpacing our ability to detect and moderate it.

2019

First AI video released (NVIDIA GauGAN)

2023

Public text-to-video access (Runway Gen-1)

2024+

Sora, Veo, Pika — photorealistic & cinematic

The Moderation Gap

- Higher fidelity makes synthetic content harder to detect. **Detectors trained on lab-controlled data often fail on real-world videos.**
- Platforms depend on user self-labeling — which almost never happens.

Our Approach: Two Complementary Research Directions

Effective moderation requires both better detection AND a deeper understanding of harm.

RD1

Improve AI-Video Detection

- Systematically benchmark diffusion-based detection models
- Test on real-world data collected in the wild — not lab-generated videos
- Identify accuracy gaps and iteratively refine the best-performing approaches

RD2

Understand AI-Video Harms

- Create or possibly extend existing harm frameworks to AI-generated content
- Quantify type and severity of harms viewers encounter
- Use findings to inform smarter, prioritized content moderation

Both directions are grounded in our preliminary TikTok dataset

Preliminary Work: Data Collection

We built a real-world dataset of AI-generated political videos from TikTok during the 2024 U.S. election cycle.

Why TikTok?

- Video-first platform with high synthetic content exposure
- Tag-based search enables targeted, reproducible collection
- High daily upload volume and broad creator base
- Prominent in prior platform moderation research

Collection Method

Selenium-based web crawler —
rate-limited, privacy-preserving, consistent with prior web measurement research.

Collection Window

Dec 23, 2024 → Jan 24, 2025 (24 business days)

Search Tags

- AI Generated Election
- Kamala Harris AI
- Trump And Elon
- Trump Biden AI

Targeted: Biden, Harris, Trump, Musk → key political actors + 'AI' keyword to surface synthetic content.

Labeling Task 1: Synthetic vs. Real

Three researchers independently hand-labeled 3,293 videos. Nearly 1 in 7 was AI-generated.

Search Tag	Synthetic	Real	Total
AI_Generated_Election	96 (2.9%)	694 (21.1%)	790 (24.0%)
Kamala_Harris_AI	56 (1.7%)	867 (26.3%)	923 (28.0%)
Trump_And_Elon	135 (4.1%)	678 (20.6%)	813 (24.7%)
Trump_Biden_AI	198 (6.0%)	569 (17.1%)	767 (23.3%)
TOTAL	**485 (14.7%)**	**2,808 (85.3%)**	**3,293**

Table 1. Unique Gen-AI videos from TikTok (Dec 2024 – Jan 2025), by tag and type.

14.7%

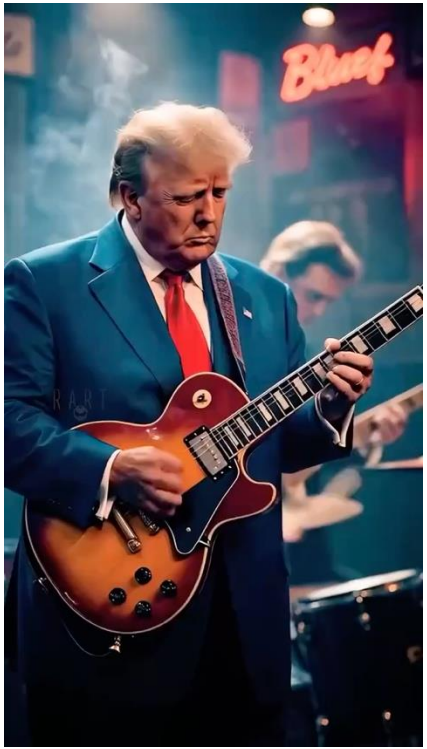
of collected videos were AI-generated

Conservative labeling

Videos labeled synthetic only when exhibiting dreamlike or hyper-realistic visual artifacts.

Labeling Task 1: Synthetic vs. Real

Conservative labeling Videos labeled synthetic only when exhibiting dreamlike or hyper-realistic visual artifacts.

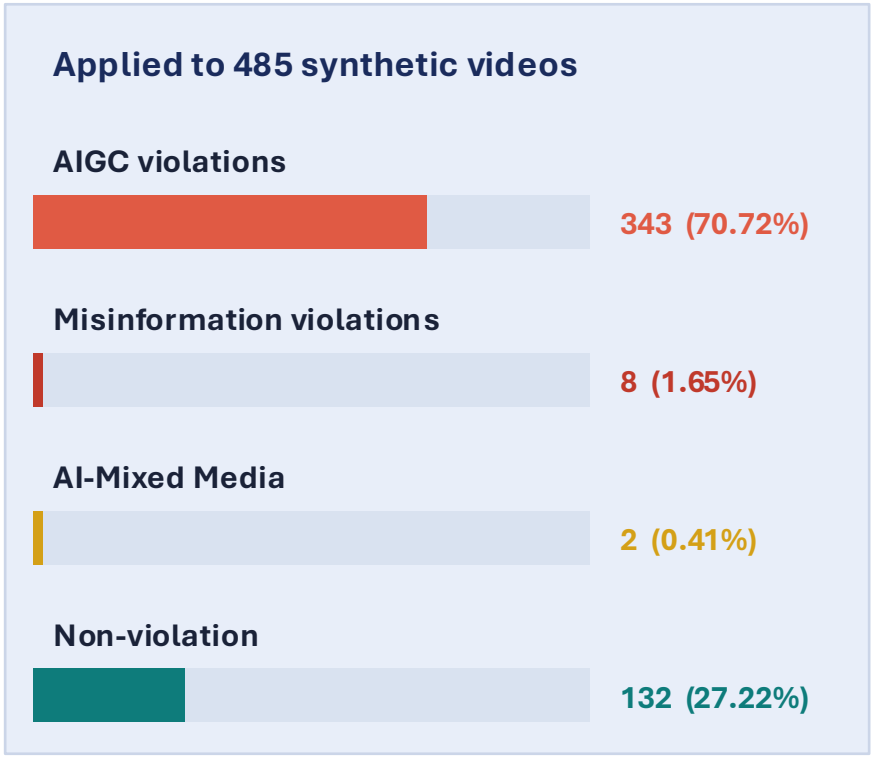


Labeling Task 2: Community Guideline Violations

Codebook Development

- Grounded in TikTok's Integrity & Authenticity guidelines — Misinformation and AIGC subsections
- Three researchers developed codes independently, then reconciled
- Initial inter-coder agreement: Fleiss's κ = 0.62, disagreements were resolved
- Final codebook: 4 primary codes, 19 subcodes

72% of synthetic videos violate TikTok's community guidelines — yet none were labeled by the platform.



RD1: Improving AI-Video Detection

The core challenge

- Most detection models are trained and benchmarked on synthetically curated datasets — not real-world videos collected in the wild.
- Increasing model variety means detectors face a distribution mismatch problem.
- Our dataset provides a rare real-world benchmark opportunity.

Our plan

Systematic benchmark of state-of-the-art models → identify gaps → iteratively refine top performers using our dataset.

Candidate Detection Models

Model	Code	Status
DeMamba	GitHub	arXiv
UNITE	None (Google)	Published
AIGVDet	GitHub	arXiv
DVID	GitHub	arXiv
LAVID (Agentic LVLM)	None (Unreleased)	arXiv

RD2: Understanding AI-Video Harms

Why this matters

- TikTok's own policy acknowledges labeled content can still cause harm
- No AI-specific harm framework has been applied to real video datasets
- Harm is platform- and context-specific — what's entertainment for one viewer is offensive to another

Our approach

Evaluate existing harm frameworks → develop an AI-specific extension or original framework → quantify type and possibly severity of harms across our dataset.

Harm signals from labeling data

Falsely depicts public figures
312 videos

Conspiracy theories (named individuals)
4 videos

Political endorsement/condemnation
2 videos

Minors depicted realistically
1 video

Crisis event misrepresentation
1 video

Open Questions for the Community

01

Data collection at scale

Our conservative labeling found 14.7% AI-gen hits. What methodology would maximize synthetic video yield without sacrificing ecological validity?

02

Detection evaluation scope

Given the rapidly evolving model landscape, how should we bound our benchmark? What evaluation metrics matter most to practitioners?

03

Build vs. adapt harm frameworks

Should we extend existing frameworks (e.g., 4Cs, Content, Contact, Conduct, Compulsion) or build an AI-content-specific one from scratch? What are the tradeoffs?

Conclusion & Takeaways

What we've done

- Collected 3,293 real-world TikTok videos during a high-stakes political window
- Identified 485 AI-generated videos via conservative manual labeling
- Found zero platform-applied labels — confirming a moderation failure
- Codified that 72% of synthetic videos violate TikTok's own guidelines

What's next

- Benchmark diffusion-based detection models on real-world data
- Expand dataset and fine-tune top-performing detectors
- Build an AI-specific harm framework
- Use findings to recommend prioritized moderation practices

Thank You!

Margie Ruffin, Khushi Patel, Samika Karumuri, Gang Wang

margieruffin@spelman.edu