Detection of Sora2 AI-Generated Videos

Jaswin Hargun, Mihir Shah

Department of Computer Science, University of Toronto

Motivation

Sora2 has made it possible to easily generate extremely convincing AI videos. This presents a novel mis/disinformation threat, particularly since watermarks are easily removed by online tools. We aim to identify methods to identify Sora2 videos, with a particular focus on video-specific techniques and techniques that require minimal data.

Related Work

A) Evaluated Benchmark References

D3: Training-Free AI-Generated Video Detection Using Second-Order Features

- **Summary:** This method exploited residual spatiotemporal inconsistencies by using the difference of differences between consecutive frame embeddings from a pretrained encoder.
- **Limitation:** The performance heavily relied on the specific artifacts captured by the pretrained encoder and the consistency of those second-order features.

Physics-Driven Spatiotemporal Modeling for AI-Generated Video Detection

- **Summary:** This paper used temporal derivatives and spatial gradients to identify physically impossible or anomalous video dynamics.
- **Limitation:** The reliance on explicit physics-based anomalies became less reliable as advanced generators like Sora2 reduce obvious physical errors.

AI-Generated Video Detections via Spatio-Temporal Anomaly Learning

- **Summary:** This approach used a two-branch approach analyzing both raw frames and optical flow to exploit common errors in flow estimation.
- **Limitation:** The detector's effectiveness was constrained by the accuracy of the external optical flow computation and, like the previous method, is vulnerable to reduction of flow errors in newer models.

B) Build Upon

Advanced Fake Video Detection via Vision Transformers

- **Summary:** This paper trained a small FCN on ViT-derived embeddings to classify AI-generated images, which we adapted to identify frames in Sora2 videos.
- **Limitation:** The approach is image-centric and struggles to capture the temporal inconsistencies unique to video generation. It also requires retraining to be applied to new models, making it less easy to generalize.

References

[1] C. Zheng et al., “D3: Training-Free AI-Generated Video Detection Using Second-Order Features. ” 2025. [Online]. Available: https://arxiv.org/abs/2508.00701

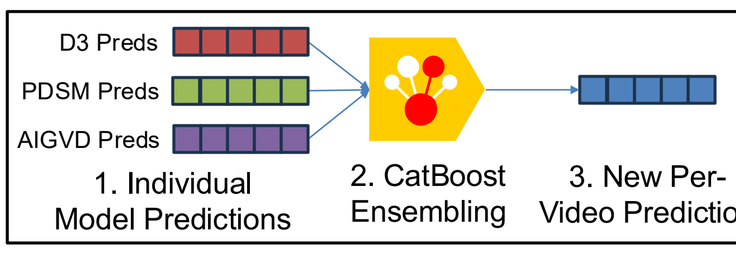
[2] S. Zhang et al., “Physics-Driven Spatiotemporal Modeling for AI-Generated Video Detection.” 2025. [Online]. Available: https://arxiv.org/abs/2510.08073

[3] J. Bai, M. Lin, and G. Cao, “AI-Generated Video Detection via Spatio-Temporal Anomaly Learning. ” 2024. [Online]. Available: https://arxiv.org/abs/2403.16638

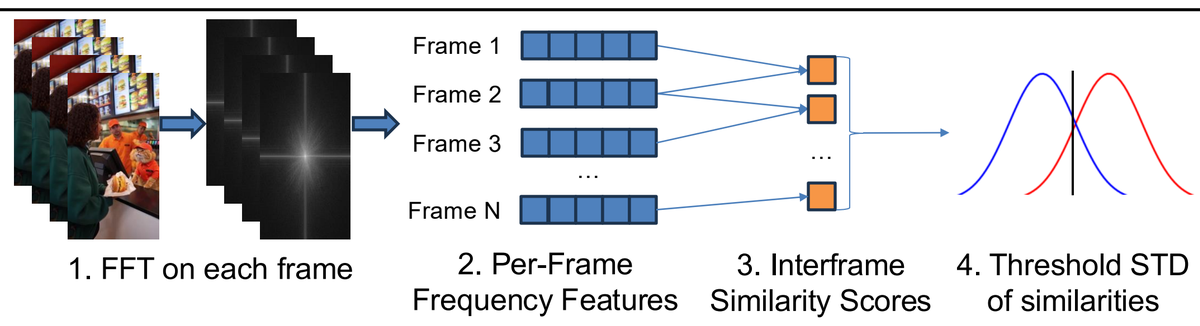
[4] J. Battocchio, S. Dell’Anna, A. Montibeller, and G. Boato, “Advance Fake Video Detection via Vision Transformers, ” in Proceedings of the 2025 ACM Workshop on Information Hiding and Multimedia Security, 2025, pp. 1–11. doi: 10.1145/3733102.3733129.

New Technique

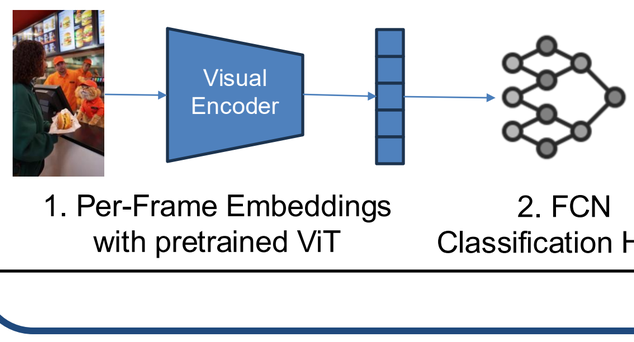
We created a dataset of 60 Sora2 Videos. We then evaluate the 3 benchmark video detection techniques described in “Related Work” to see how they generalize to Sora2. We then trial 3 new techniques that build upon them.



First, we try ensembling the 3 tested models to obtain a more reliable prediction. We tested a few methods and found CatBoost performed best.



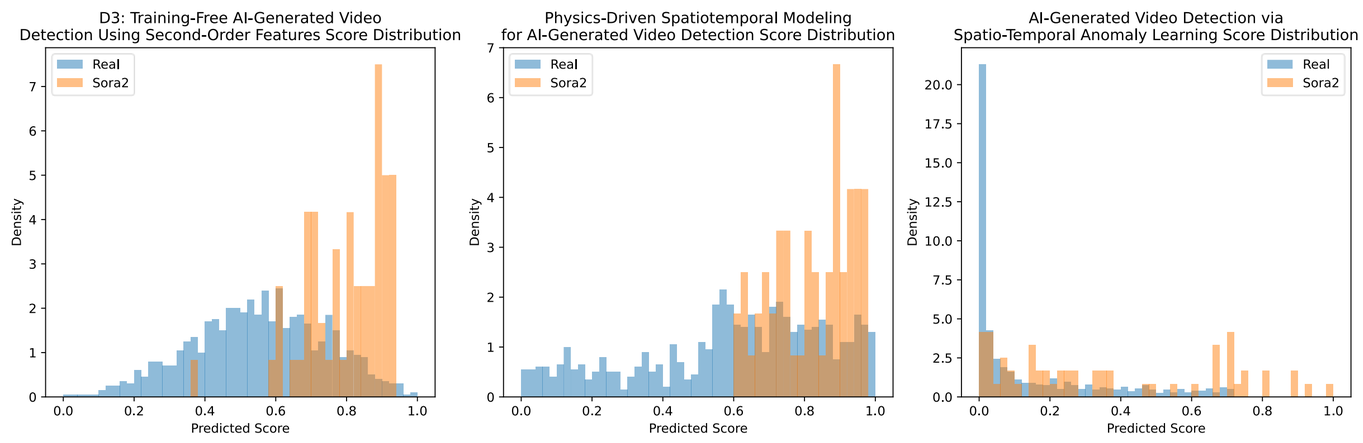
Next, we try adapting D3 but comparing frequency features rather than embeddings from a pretrained model. This was inspired by the presence of unnatural high-frequency artifacts in the background of many videos.



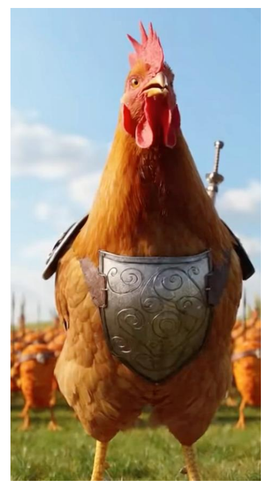
Finally, we try training a fully connected classification head using the embeddings created by a pretrained ViT as described in [4].

Experimental Results

The benchmark techniques struggle with very high false positive rates. The score distributions shown below highlight why. While the distribution of AI videos tends to be towards the left, there is no clear delineating boundary that can be applied.



The benchmark techniques particularly struggle on videos with rapid transitions and those with cartoonish depictions taking up a large portion of the video (e.g. chicken warriors as shown on the left).



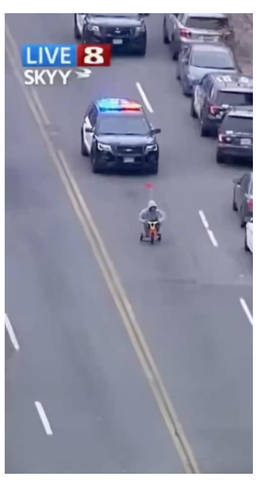
The FCN head vastly outperforms other methods, while the ensemble performs best out of the untrained methods.

Approach	Max F1
D3	0.3889
Physics-Driven Spatiotemporal Modeling	0.2257
Anomaly Learning Score Distribution	0.3784
Ensemble of above 3 (CatBoost)	0.4769
D3 with frequency features	0.2905
FCN head	0.8372

Although it would be ideal to avoid any finetuning, even an ensemble of general-purpose techniques failed to achieve high performance. Considering the substantial performance gains and relatively low data requirement (<40 videos), training a small classification head appears to be a worthwhile tradeoff.

Layer	Output Nodes
Pretrained ViT	1000
Dense+ReLU	512
Dropout (50%)	n/a
Dense+ReLU	128
Dropout (50%)	n/a
Dense+Sigmoid	1

FCN Head Architecture



Notably, some of the videos that were uncaught by any method are also indistinguishable from real videos to the human eye (example shown on the right). Such videos may require extremely tailored techniques that take advantage of invisible signatures unique to specific models.