Zero and Few-Shot Detection of Sora2 AI-Generated Videos

Jaswin Hargun, and Mihir Shah

Abstract—The rapid rise of Sora2 has made the generation of highly realistic artificial videos widely accessible, intensifying concerns surrounding misinformation and erosion of visual trust. While numerous techniques exist for detecting AI generated videos, no existing work shows how these established techniques generalize to this new model. In this work, we construct a curated dataset of Sora2-generated videos and benchmark three state-of-the-art detection approaches: D3, NSG-VD, and Spatio-Temporal Anomaly Learning. Our experiments reveal consistently high false positive rates and a lack of clear decision boundaries, especially for videos with rapid transitions or stylized content. To address these shortcomings, we evaluate three new directions: ensembling of existing techniques, frequency-based extensions of D3, and a lightweight fully connected classification head trained on pretrained ViT embeddings. Our FCN-based method significantly outperforms all baseline approaches, achieving an F1 score of 0.84 despite requiring fewer than 40 training videos. These findings highlight both the difficulty of detecting modern AI-generated videos and the potential of few-shot learning with pretrained models to bridge the gap where training-free methods fail. Remaining failure cases underscore the need for future research into model-specific invisible signatures and more robust multimodal detection strategies.

Index Terms—Computational imaging, AI-generated video detection, vision transformers

1 Introduction

The release of Sora2 in September 2025 marked a significant milestone in text to video generation, rapidly reaching more than one million downloads within its first week and quickly becoming one of the most widely used video creation tools to date. Its success, however, has also intensified concerns about the role of AI generated video (AIGV) in enabling misinformation and disinformation. The videos produced by Sora2 often display a level of photorealism and temporal coherence that makes them difficult for viewers to distinguish from authentic footage. OpenAI provides 3 mechanisms to identify Sora2 generated content: a visible watermark, C2PA metadata, and reverse search tools for both images and audio. Of these, only the watermark is immediately visible to users. To a user casually viewing a Sora2 video downloaded from Sora2 and posted to other locations, only the watermark (which is added when a video is downloaded from the Sora app) can be used to immediately identify it as an AIGV. While OpenAI has taken steps to make it harder to remove this watermark, such as having it move throughout a video, this is far from sufficient. Watermark removal services are already widespread and can erase these indicators with minimal traces. Figure 1 shows an example of the output of such a watermark removal service. These developments highlight the need for robust, content based detection methods that operate directly on the visual signal rather than on metadata or auxiliary information that can be easily removed or altered. In this project, we aim to investigate the effectiveness of existing AI generated video detection techniques when applied to Sora2 content and to develop improved methods to better identify Sora2 videos. In particular, we focus on techniques that rely solely on visual data (rather than metadata or fingerprints which can be embedded by the video generation service). We also focus on techniques which can operate with limited training samples, an increasingly important requirement as new generative models continue to emerge faster than large scale labeled datasets can be constructed. We benchmark 3 existing techniques for AIGV detection which use different factors to make their predictions. One uses frame-wise vector embedding similarities generated by a pretrained model, another uses optical flow features, and the last aims to identify anomalous spatiotemporal dynamics. We evaluate zero-shot generalizability of these techniques to Sora2, applying no retraining. We then propose and test 3 new techniques. One simply ensembles existing techniques, aiming to compensate for each technique's weaknesses. The next modifies an existing benchmark, substituting vector embeddings for handcrafted frequency features. Finally, the last modifies a technique from literature, training a classification model directly on frame-wise predictions. By benchmarking state of the art detection approaches and exploring new visual cues and model driven features, our goal is to advance reliable and practical detection strategies for highly realistic AI generated videos. This work contributes to the broader effort to mitigate mis/disinformation risks that arise as generative video models rapidly improve in quality and accessibility.

2 Related Work

While Sora2 was released fairly recently, a wide variety of techniques for identifying AI generated videos have been developed for previously released models. Some focus on frequency features. Corvi et al aim to identify small-scale artifacts rather than visible errors by training a classification head on top of a pretrained DINOv2 model to identify

• *Jaswin Hargun and Mihir Shah are part of the Department of Computer Science at the University of Toronto.*
E-mail: jhargun@cs.toronto.edu or mihirshah@cs.toronto.edu

Fig. 1. Example of watermark removal.

differences in specific frequency bands characteristic to AI generated videos [1]. Vahdati et al train a CNN to identify AIGVs based on the noise residual produced by pretrained denoising models, though they point out that the noise residuals of AIGVs are much harder to differentiate from real data than AI images [2].

Others propose methods that focus on regeneration of images using generative models. Wang et al note that pretrained diffusion models show a "preference" for AI generated images, altering them less than real images when reconstructing, leading them to train a ResNet-50 model on the residuals between the original and reconstruction [3]. Liu et al build upon this work, adapting it for detection of AIGVs by training a CNN and LSTM on the sequence of residuals for individual frames [4].

Some techniques use vector embeddings created by pretrained models. Zheng et al use the difference of differences between vector embeddings of individual frames to distinguish AIGVs [5]. Battochio et al generate frame-level embeddings with a pretrained vision transformer, then use a Video Query Transformer to generate a single video embedding, and finally train a classification head to identify whether a given video is AI or real [6].

Finally, some methods aim to identify physically unrealistic features in the video. Some do this using optical flow to identify unrealistic motions in videos. Bai et al note that optical flow maps in AIGVs tend to be fairly different from those in real videos, leading them to generate optical flow maps using a pretrained RAFT model and train a 2-branched CNN to detect AIGVs (with one branch focusing on the image and the other focusing on the optical flow map) [7]. Ji et al propose a similar approach, though they use GMFlow to generate optical flow maps and a transformer-based 2-branched model [8]. Others use non-optical flow features. Sarkar et al train 3 classifiers to identify different types of inconsistencies in AI images: misaligned shadows, warped perspectives, and misaligned line segments [9]. Zhang et al instead compare the temporal gradient (determined by comparing frame differences) and spatial gradient (approximated with a diffusion model) [10].

TABLE 1
Summary of previous works described. Techniques are marked as requiring training if they require any task-specific training (except selection of a threshold for handcrafted features). The level is image if the technique wasn't evaluated on AIGVs, frame if it makes predictions on individual frames, and video if it uses temporal information (beyond simply averaging framewise predictions).

Approach	Requires Training	Level	Category
Corvi et al [1]	Yes	Frame	Frequency Based
Vahdati et al [2]	Yes	Frame	Frequency Based
Wang et al [3]	Yes	Image	Reconstruction Based
Liu et al [4]	Yes	Video	Reconstruction Based
Zheng et al [5]	No	Video	Embedding Based
Battochio et al [6]	No	Video	Embedding Based
Bai et al [7]	Yes	Video	Optical Flow Based
Ji et al [8]	Yes	Video	Optical Flow Based
Sarkar et al [9]	Yes	Image	Geometry Based
Zhang et al [10]	No	Video	Physics Based

Table 1 shows a summary of the techniques described. While these techniques have shown promise, with many demonstrating impressive performance on a variety of AI generated image and/or video models, each of them presents some issues. First, almost all require at least some training. This means unless they can show zero-shot generalizability to AIGV models which they were not trained on, the release of a new AIGV model may mean retraining is required. Given the rapid pace of progress in this field, this retraining requirement could represent an ongoing difficulty to real-world utilization. Particular types of approaches also have their own challenges. The optical flow, geometry, and physics based techniques look for physically improbable or impossible motion or views in AI generated content. As generative models improve, such errors may become less common, limiting the applicability of such techniques. Frequency, reconstruction, and embedding based techniques may be more robust to such model improvements. They look at differences which are invisible to the human eye, meaning fixing such differences may not be a priority when developing new AIGV models. However, this also means that such techniques could stop working without any visible changes to a generative model's outputs.

We select 3 techniques to benchmark, which we explain in more detail. Each of the techniques show good generalizability to a broad range of AIGV models, including the first version of Sora. These benchmarks were also selected due to their variety of approaches and code and model weight availability. One of our proposed methods also builds directly upon a previous technique, which we explain in more detail.

2.1 D3

The authors present D3, a training-free AIGV detector [5]. They use a pretrained visual encoder to extract embeddings for each frame and compute the similarity (via L2 or cosine similarity) of consecutive frames' embeddings. They then

note that empirically, real videos tend to have significant volatility in inter-frame similarities while AIGVs display a flatter pattern. To capture this second-order trend, they take the standard deviation of these scores across an entire video. Videos with high standard deviations (meaning high volatility in similarities) are classified as real, while those with low standard deviations are classified as AI generated. The authors show good performance across a variety of AIGV models, including an average precision of 0.9991 on Sora.

2.2 NSG-VD

The authors present NSG-VD, a physics driven framework for detecting AIGVs, built around a new statistic termed the Normalized Spatiotemporal Gradient (NSG) [10]. This metric captures the ratio between spatial probability gradients and temporal density variations, enabling the detection of inconsistencies in video dynamics when viewed through the lens of probability flow conservation. The method leverages pre trained diffusion models to approximate spatial gradients and incorporates motion aware temporal modeling without relying on explicit motion representations such as optical flow. NSG features are extracted for each video, and the Maximum Mean Discrepancy is employed to quantify the difference between the NSG distribution of a test sample and that of authentic video data, where larger divergences signal generated content. The authors show good performance across a variety of AIGV models, achieving an F1 score of 0.8713 on Sora even when not trained on Sora videos.

2.3 Spatio-Temporal Anomaly Learning

The authors present a 2-branch model to detect AIGVs [7]. They note that even high-quality generated videos that are indistinguishable from real videos to humans tend to have temporal discontinuities which can be seen in optical flow maps as blurry contours. Therefore, they propose a model which consists of two ResNet50 classifiers: one which takes a random crop of a frame as input and one which takes a random crop of an optical flow map generated by a pretrained RAFT model. These output two independent classifications indicating whether the frame is AI generated. These independent classifications are then combined with a weighted summation to get a single frame-wise classification. The mean of these frame-wise classifications is used as the video classification. The authors show good performance across a variety of AIGV models, achieving an AUC-ROC of 92.1% on Sora (though they only have 48 Sora videos and don't report any other metrics, meaning this result could be inflated by the severe class imbalance).

2.4 Advance Fake Video Detection via Vision Transformers

While this is not a method we benchmark, one of our proposed techniques is based on this paper. The authors propose a visual only detection framework based on Vision Transformers that adapts image level forensic analysis to the video domain [6]. A pretrained vision transformer (ViT) is used to generate vector embeddings for each frame in a video. A positional encoding is then added to these frame-wise embeddings, which are then passed to a Video Q-Former which outputs a single, video-level vector embedding. A classification head is then used to learn to classify videos as real or AI generated based on these video-level embeddings. The authors train on 503 real videos and 2515 AIGVs, showing good performance across a number of AIGV models (though they do not evaluate their technique on Sora videos).

TABLE 2
Summary of proposed techniques. Techniques are marked as requiring training if they require any task-specific training (except selection of a threshold for handcrafted features). The level is frame if it makes predictions on individual frames, and video if it uses temporal information (beyond simply averaging framewise predictions).

Approach	Requires Training	Level
Ensembling (summation)	No	Video
Ensembling (CatBoost)	Yes	Video
D3 with Frequency Features	No	Video
Classification Head	Yes	Frame

3 Proposed Method

We first create a dataset containing a mix of real and Sora2 videos. We then choose three benchmark techniques, which we evaluate using this dataset. Afterwards, we test three techniques designed to improve upon the benchmarks. A summary of the proposed methods is shown in Table 2.

3.1 Dataset

We use the first 1000 real videos from the GenVideo dataset (video IDs 1-1000), a dataset created by Zheng et al which combined data from multiple previous publications [5]. Only the first 1000 videos were used to reduce processing time and class imbalance while still retaining a diverse dataset. We then supplement these with 60 Sora2 videos. Half of these videos were taken from the new page while the other half were taken from the leaderboard of most popular videos (which will presumably be higher quality on average than the first 30 videos).

These videos are downloaded both directly from the app, which adds a watermark, and using savesora.com, a free online tool which detects, removes, and infills the watermark before downloading (while preserving original resolution). While this risks adding artifacts that could confound our analysis, a visual inspection showed a lack of clearly visible artifacts. Cropping the videos was considered as an alternative to avoid the risk of introducing invisible artifacts, but this was not pursued for two main reasons. First, due to the constant motion of the watermark throughout the video, cropping would result in many short (< 2 second) cuts of the video, with different aspect ratios for each of these crops, making it difficult to properly evaluate the methods tested. Second, those aiming to create mis/disinformation with Sora2 are more likely to use infilling tools than cropping, meaning a method capable of identifying AIGVs with artifacts present would still have utility.

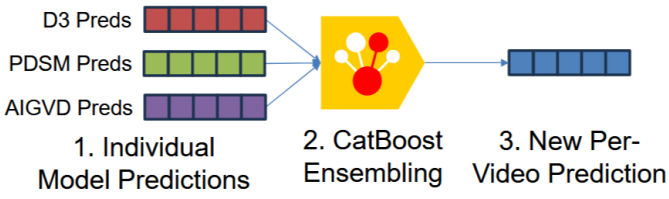

Fig. 2. Ensembling approach

3.2 Benchmark Ensembling

The first new method was combining the outputs of the three benchmark methods. The goal was to use each benchmark method to account for the weaknesses of the other two. First, the scores output by each method were simply normalized to the range [0,1] and summed to get one combined score. Next, 5 different ensembling techniques were tested: logistic regression, random forest, k-nearest neighbors, CatBoost, and XGBoost, with CatBoost showing the best performance. This technique is shown visually in figure 2.

3.3 D3 with Frequency Features

The second new method was an adaptation of the technique used by Zheng et al. Rather than comparing the difference of differences between vector embeddings of frames, the idea was to compare the difference of differences between vectors of handcrafted frequency features. While previous work had shown that frequency features could be useful for identifying AIGVs, these techniques required training [1], [2]. Our proposed technique takes the FFT of an image and calculates meaningful features. Then, we compute the cosine similarity between features for consecutive frames, take the standard deviation of these similarities, and threshold the standard deviation to classify videos. The frequency features used include:

1) Radial energy distribution: the FFT was split into 20 radial bins (with each Fourier coefficient falling into its bin based on distance from the center of the shifted FFT). The value of each of these bins is the sum of coefficients that fall into it. These features represent the energy present at lower vs higher sinusoidal frequencies.
2) Angular energy distribution: the FFT was split into 12 angular bins (with each Fourier coefficient falling into its bin based on its angular position in the shifted FFT). The value of each of these bins is, as before, the sum of coefficients. These features represent the energy present at different sinusoidal orientations.
3) The standard deviation of the radial and angular energy distribution: the standard deviation of the radial and angular energy distributions was calculated. A higher radial standard deviation indicates that the FFT energy is concentrated in frequency bands, while a higher angular standard deviation indicates that the FFT energy is concentrated in certain sinusoidal directions.

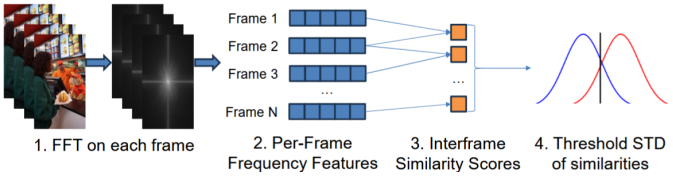

Fig. 3. Frequency feature based approach

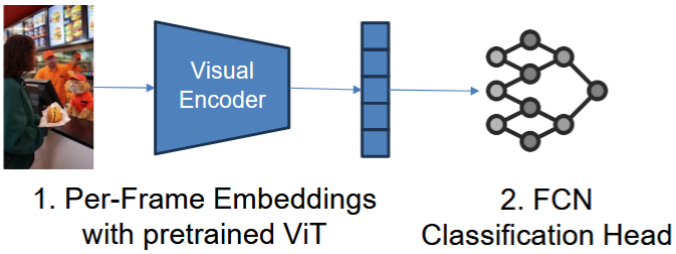

Fig. 4. Classification head based approach

4) The mean, standard deviation, kurtosis, and skew of the FFT: these simple statistical features can provide additional information about the frequencies present.
5) The "peakiness" of the FFT: the number of peaks (or frequencies with coefficient magnitudes above the 95th percentile) was calculated. This feature represents whether there are a few dominant frequency peaks. A higher value indicates a more widely distributed energy, while a lower value indicates highly concentrated energy (meaning there is a stronger periodic component to the image).

While the previous techniques that used frequency features required training a classification model, our proposed method requires no training beyond the choice of a threshold for the standard deviation. This technique is shown visually in figure 3.

3.4 Classification Head

Finally, we develop a technique similar to the one described by Battochio et al [6]. We generate per-frame vector embeddings using a pretrained vision transformer, then train a fully connected network to predict whether a frame is from a real or Sora2 video. Note that our technique is simpler. Whereas Battochio et al use a video query transformer to convert a series of frame-level vector embeddings into a single video-level embedding, we train directly on the frame embeddings. This gives us many training examples per video rather than just 1, allowing us to use a much smaller dataset (60 AIGVs rather than over 2000). This technique is shown visually in figure 4.

Note that due to our small dataset, we saw issues of overfitting. Therefore, we use a slightly different classification head than the one described in literature. Notably, we apply a very aggressive dropout of 50% to each of our dense layers. The specifics are shown in Table 3.

4 Experimental Results

Below are the experimental results of each of the benchmark and proposed techniques evaluated on our dataset. For the

TABLE 3
Classification head architecture.

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

ensembling and classification head approaches, the scores are reported on a validation set consisting of 25% of the real and 25% of the Sora2 videos, with the rest of the data used to train the classifier. Since the benchmarks and D3 with frequency features approach require no training, the results are calculated using the entire dataset.

4.1 Benchmark Techniques

We see that the benchmarks struggle to differentiate between real and Sora2 videos. Figure 5 shows the predictions of each benchmark, with a score of 0 representing a prediction that the video is real and 1 representing a prediction that the video is AI. There is a clear difference in the distribution of predictions between real and AI videos for all benchmarks, with AI videos consistently getting higher scores. This indicates that these techniques can identify some differences between real and AI videos. However, none of these techniques have cleanly separable distributions. This means any threshold chosen will have a large proportion of false positives. Figure 6 shows the precision-recall curves for these methods. These curves show extremely low precision at almost any threshold, highlighting the high false positive rate.

Notably, we see that the NSG-VD technique performs significantly worse than the other two methods. This is likely due to the advances of Sora2, which is much better at making physically realistic videos than previous video generation models. However, we note that the optical flow based method performs significantly better. This may be because the CNN model trained in this method can identify small-scale errors in optical flow, whereas NSG-VD thresholds a global spatiotemporal gradient, making it less sensitive to small errors.

A qualitative analysis of the errors shows that each of the benchmark techniques struggles to identify AIGVs containing cartoonish depictions. This may be because the techniques were originally designed to identify attempts to create realistic AIGVs. The benchmarks also struggle with videos that have many rapid cuts. This is likely because older AIGV models created continuous videos without cuts. Since each of the benchmark techniques relies on identifying temporal inconsistencies, such cuts present a major challenge.

4.2 Proposed Techniques

The F1 scores of each of the benchmark and proposed techniques are shown in Table 4. Note that due to the severe class imbalance of our dataset (1000 negative classes and 60 positive classes), F1 score is reported rather than metrics such as accuracy.

TABLE 4
F1 Scores of each benchmark and proposed technique.

Approach	F1
D3	0.3889
NSG-VD	0.2257
Spatio-Temporal Anomaly Learning	0.3784
Sum of Benchmarks	0.4000
Ensemble of Benchmarks (CatBoost)	0.4769
D3 with Frequency Features	0.2905
Classification Head	0.8372

TABLE 5
Precision, recall, and F1 scores of each ensembling technique.

Approach	Precision	Recall	F1
Logistic Regression	0.2488	0.8833	0.3872
Random Forest	0.7056	0.3167	0.4292
K-Nearest Neighbors	0.4819	0.2667	0.3344
CatBoost	0.3323	0.8333	0.4724
XGBoost	0.4427	0.4167	0.4287

4.2.1 Benchmark Ensembling

Figure 7 shows the results of summing the normalized scores of each technique. While a separation in the distributions of real and AI scores is observed, it is not significantly better than any individual technique. The maximum F1 score of this technique is only 0.4, while the average precision is only 0.412, compared to 0.3784 and 0.342 respectively for the 3rd benchmark technique.

Using a more complex ensembling model leads to improved results. As mentioned previously, 5 ensembling techniques were tested. The precision, recall, and F1 scores of each of these methods are shown in table 5 (note that since these methods return a binary classification rather than a probability, scores and precision-recall curves are not shown). CatBoost achieves the highest F1, followed closely by XGBoost and Random Forest. However, while CatBoost achieves a high recall at the cost of low precision, Random Forest does the opposite and XGBoost has a balanced precision and recall. Therefore, depending on one's priorities, it may make sense to use a model other than CatBoost.

4.2.2 D3 with Frequency Features

Figure 8 shows the results of adapting D3 with frequency features. It achieves an average precision of 0.215 and F1 score of 0.2905, slightly worse than the average precision of 0.264 and significantly worse than the F1 score of 0.3889 of D3.The score distributions help explain this result. Recall that the original paper treats low standard deviations of inter-frame similarity scores as a sign of AI generated videos. With our technique, most real videos and all AI videos have similarly high predicted scores, which means they have low standard deviations. This could be caused by a few factors. First, many of the frequency features used (e.g. energy across radial bands) are likely

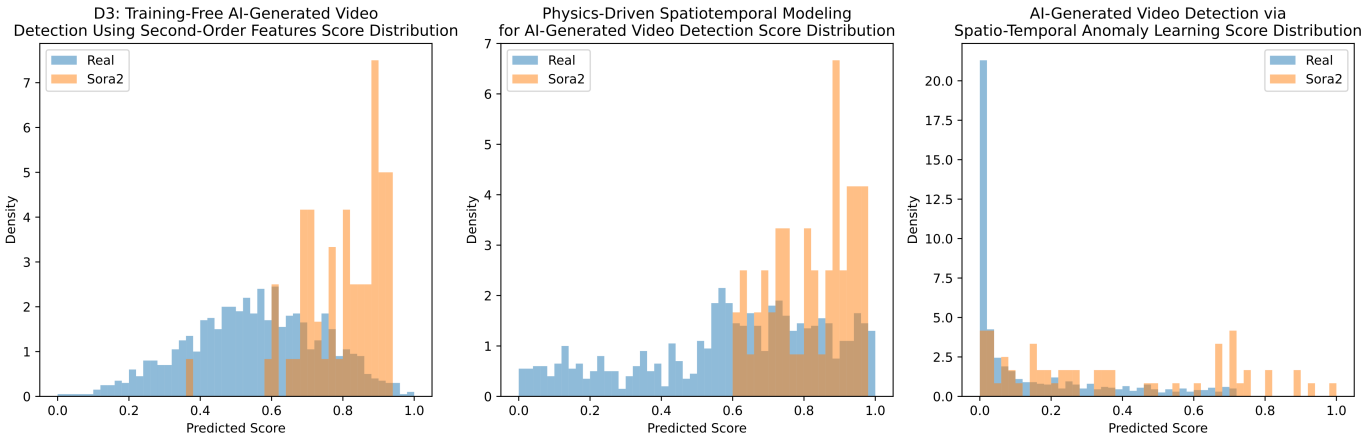

Fig. 5. Score distributions for each of the benchmark techniques.

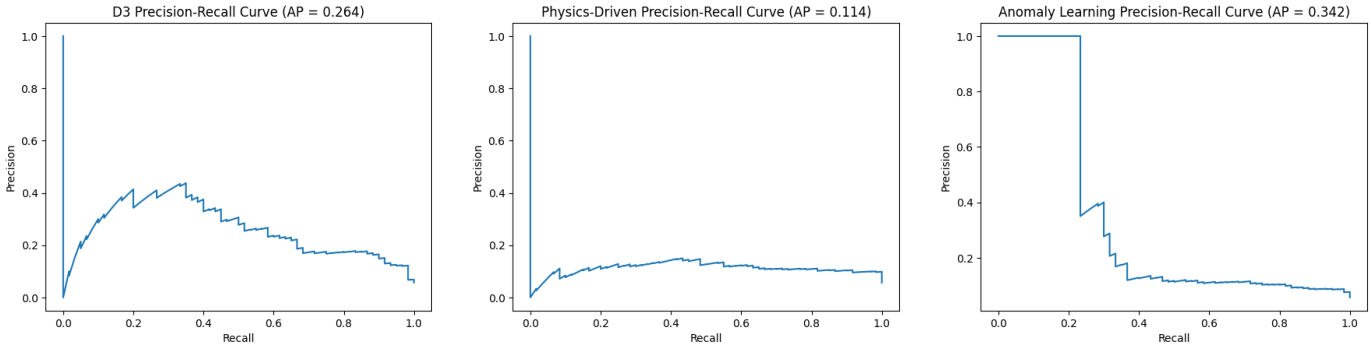

Fig. 6. Precision-Recall curves (with average precision score) for each of the benchmark techniques.

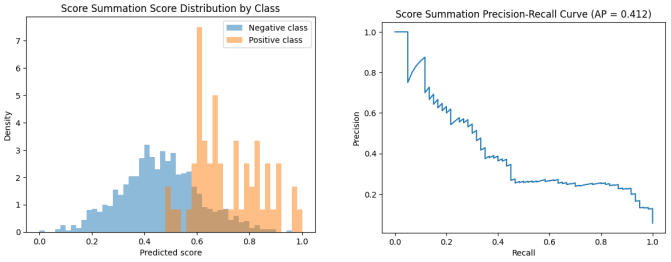

Fig. 7. Classification scores and precision-recall curve for summation of benchmarks.

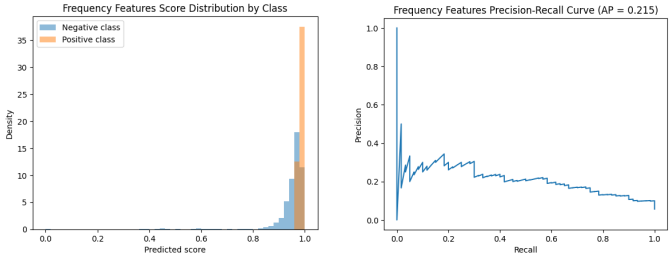

Fig. 8. Classification scores and precision-recall curve for D3 with frequency features.

to remain relatively constant between consecutive frames. Second, only 40 features were used, which means the feature vectors were significantly shorter than the vector embeddings used originally. In high dimensional space, random vectors typically have lower similarity scores, meaning it is easier for trained embeddings to produce more meaningful high similarities for related items and low similarities for unrelated ones. Finally, embedding vectors are much more capable of capturing semantic differences between frames. All of these factors mean frequency feature similarities are less likely to display high standard deviations, making it harder to identify real videos. Notably, this method showed some promising performance, outperforming the NSG-VD benchmark in both average precision and F1 score.

4.2.3 *Classification Head*

Figure 9 shows the results of training a classification head on frame-wise ViT embeddings. These results are by far the best of any method shown. Almost all real frames have a low score, and the right threshold achieves an average precision of 0.901 and F1 score of 0.8372. This performance is especially impressive when one considers the simplicity of this technique. By using only ViT frame-wise embeddings rather than video query transformer video embeddings, we can learn from a very small amount of Sora2 videos.

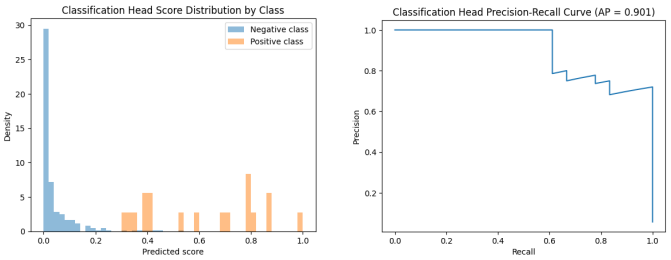

Fig. 9. Classification scores and precision-recall curve for classification head.

5 Discussion

5.1 Limitations and Future Work

Our approach has a few limitations. The primary limitation is dataset size. We used 60 Sora2-generated videos and 1,000 real-world videos, which may not fully capture the breadth of Sora2's generative capabilities. Sora2 can generate a wide range of scenes, motions, and subjects, and a dataset of 60 fake videos cannot represent the model's full distribution. On top of making evaluation difficult, it presents a risk of overfitting. For example, our classification head had to incorporate a very aggressive dropout rate of 0.5 to counteract this. Performance when trained on a small dataset is valuable information, as a technique capable of learning on very little data will be easier to adapt as new AIGV models are released. However, future work could include the creation of a larger Sora2 dataset and ablation studies to determine exactly how the performance of each technique varies with differing amounts of data.

Second, the generalizability of these methods was not thoroughly examined. First, we evaluate our proposed methods only on our Sora2 dataset. This means the performance of these techniques may differ when applied to AIGVs created by different models. This could be addressed in the future by evaluating proposed techniques on a mix of existing datasets as well as the Sora2 dataset provided. On top of this, our evaluation was done only on videos with the watermark removed by inpainting tools, which could introduce artifacts. This is a particular concern with the classification head, which uses ViT embeddings which are difficult to interpret. It is possible the model is simply learning to pick up on differences in embeddings generated for real and fake videos, such as the "preferences" shown in previous work [3]. However, it could also be identifying invisible artifacts from the inpainting process. Future work could test performance with multiple different watermark removal services to better determine whether these artifacts are playing a major role in the results.

In addition, the computational cost of analyzing long videos presents a serious challenge for the development of a real-time detection pipeline. This makes the deployment of an AIGV detection system to places like a web browser difficult. Without real time, seamless detection of AI generated videos, it is difficult to protect users from being unknowingly influenced by mis/disinformation efforts using AIGVs. Future work could focus on

Finally, there are plenty of other sources of information or confounding factors that could be used and combatted respectively. For example, Sora2 generates not just video but also audio, which is often noticeably desynchronized (particularly when people are talking, with the audio not matching lip movements). Multi-modal features such as this could be used to better identify AIGVs. Our dataset also focused on full resolution Sora2 videos. Future studies could explore improved resistance to compression, re-encoding, or adversarial editing. Such transformations could be used in real world mis/disinformation campaigns, and they may weaken the reliability of detection systems, so developing techniques robust to these transformations is an important next step.

5.2 Conclusion

The rapid progress of video generation systems such as Sora2 introduces important social, ethical, and security considerations. As generated videos become more realistic, people may find it increasingly challenging to recognize which content is genuine. This raises concerns about misinformation, identity misuse, political manipulation, and reputational harm.

We demonstrate that zero-shot techniques which showed promise on AIGVs generated by older models do not generalize well to Sora2 videos. However, we show that few-shot methods hold major promise. Detection systems like these are essential for preserving trust in digital media. However, they require ongoing retraining and evaluation with the ever continuing development of generative models. At the same time, further advances in generation may reduce the visibility of artifacts that current detectors rely on, necessitating continuing development of new techniques for AIGV detection. This ever increasing challenge also highlights not just the importance of further research but the need for coordination among researchers, technology companies, and policy makers to build standards for transparency, verification, and responsible deployment of generative video technology. Understanding these broader implications is important for ensuring that powerful video generation tools support creativity and innovation while protecting the reliability of information in the digital world.

Acknowledgments

The authors would like to thank Professor David Lindell for his advice and guidance during our project, particularly in the literature review and ideation stage. We would also like to thank the Department of Computer Science for the support in presenting our work, including compensating poster costs and providing a venue for presentation.

References

[1] R. Corvi, D. Cozzolino, E. Prashnani, S. D. Mello, K. Nagano, and L. Verdoliva, "Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation," 2025. [Online]. Available: https://arxiv.org/abs/2506.16802

[2] D. S. Vahdati, T. D. Nguyen, A. Azizpour, and M. C. Stamm, "Beyond deepfake images: Detecting ai-generated videos," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024, pp. 4397–4408.

[3] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," 2023. [Online]. Available: https://arxiv.org/abs/2303.09295

[4] Q. Liu, P. Shi, Y.-Y. Tsai, C. Mao, and J. Yang, "Turns out i'm not real: Towards robust detection of ai-generated videos," 2024. [Online]. Available: https://arxiv.org/abs/2406.09601
[5] C. Zheng, R. suo, C. Lin, Z. Zhao, L. Yang, S. Liu, M. Yang, C. Wang, and C. Shen, "D3: Training-free ai-generated video detection using second-order features," 2025. [Online]. Available: https://arxiv.org/abs/2508.00701
[6] J. Battocchio, S. Dell'Anna, A. Montibeller, and G. Boato, "Advance fake video detection via vision transformers," 2025. [Online]. Available: https://arxiv.org/abs/2504.20669
[7] 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
[8] L. Ji, Y. Lin, Z. Huang, Y. Han, X. Xu, J. Wu, C. Wang, and Z. Liu, "Distinguish any fake videos: Unleashing the power of large-scale data and motion features," 2024. [Online]. Available: https://arxiv.org/abs/2405.15343
[9] A. Sarkar, H. Mai, A. Mahapatra, S. Lazebnik, D. A. Forsyth, and A. Bhattad, "Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now," 2024. [Online]. Available: https://arxiv.org/abs/2311.17138
[10] S. Zhang, Z. Lian, J. Yang, D. Li, G. Pang, F. Liu, B. Han, S. Li, and M. Tan, "Physics-driven spatiotemporal modeling for ai-generated video detection," 2025. [Online]. Available: https://arxiv.org/abs/2510.08073