Artificial Intelligence and Social Computing, Vol. 163, 2025, 103–110
https://doi.org/10.54941/ahfe1006041

Interpretable AI-Generated Video Detection Using Deep Learning and Integrated Gradients

Joshua Weg, Taehyung Wang, and Li Liu

Department of Computer Science, California State University Northridge, CA, 91330, USA

ABSTRACT

The rise of generative AI has enabled highly realistic text-to-video models, raising concerns about misinformation and its impact on social media, news, and digital communications. AI-generated videos can manipulate public opinion, influence elections, and create false narratives, making robust detection methods essential for maintaining trust. Our research into video generation models revealed that diffusion transformers operate on noisy latent spaces, inspiring our classifier's architecture to analyze videos using the same structural units as generation models. This approach ensures adaptability to emerging AI techniques while maintaining high detection accuracy. Our explainable video classifier leverages deep learning, incorporating a convolutional encoder for latent representation, a patch vectorizer for feature extraction, and a transformer for final classification. Integrated Gradients (IG) provides transparency by highlighting the video elements that influenced the model's decision, enabling human-interpretable explanations. We successfully developed an AI model to identify AI-generated content and classify videos accordingly. Our design was informed by a deep understanding of state-of-the-art generative models, ensuring alignment with their underlying mechanisms. In addition to achieving high accuracy, we validated the model's ability to provide clear and interpretable explanations for its decisions.

Keywords: Explainable AI, Computer vision, Video generation, Feature importance, Content authentication

INTRODUCTION

Generative AI is entering a new phase in its video generation capabilities. Several text-to-video models, including OpenAI's SORA, can now generate highly realistic videos up to a minute long from a single prompt. As this technology advances, concerns about misinformation continue to grow. AI-generated videos enable users to create any narrative they choose, potentially misleading audiences who may struggle to distinguish between real-world video and AI-generated content. It is crucial to determine the authenticity of a video. However, labeling AI-generated content alone is insufficient; we must also provide clear evidence to support these classifications. For too long, AI models have been evaluated primarily on their accuracy, but it is equally important to demand transparency in their decision-making processes.

This paper presents an interpretable video classifier that distinguishes between AI-generated and real videos. The deep learning model is designed based on insights from Video Vision Transformers (ViViT) and Diffusion Transformers and is trained on a large dataset labeled as either AI-generated or real. While achieving high accuracy in classification is essential, we also employ Integrated Gradients (IG) to evaluate the model's ability to explain its decisions, ensuring greater transparency in video content verification.

RELATED WORKS

Explainable AI

Explainable AI (XAI) models are designed to be interpretable. Deep learning models, often referred to as 'black box' models, reveal only their inputs and outputs, while their internal decision-making processes remain unexplainable. This means that while these models can be trained to produce correct answers, they do not inherently provide explanations for their decisions.

This lack of transparency is problematic because explanations are essential for justifying decisions, improving processes, controlling actions, and discovering new approaches (Adadi and Berrada, 2018). If AI is to be trusted as an expert, it must be able to explain its reasoning.

In XAI, we focus on these questions (Adadi and Berrada, 2018) for AI:

- What features did the AI consider most important when making a decision?
- Can the AI provide alternative decisions or outcomes if parameters are slightly altered?
- How does the AI respond to changes in input data or under different scenarios?

Another objective of XAI is to encourage greater skepticism when engaging with AI systems. Human thinking is often described as operating through two systems. System one is fast and intuitive, relying on heuristics and shortcuts, while system two is slower and more deliberate, relying on logic and critical thinking.

A key issue with AI systems is that they can appear competent through system one processing, leading to a bias that assumes AI must be inherently intelligent and incapable of making mistakes. XAI aims to develop explainable techniques that engage more with system two thinking, which fosters skepticism. Encouraging skepticism increases the demand for AI systems to appear competent and prove their competence through clear and transparent explanations (Liao and Varshney, 2017).

ViViT and Diffusion Transformers

One approach to processing video data is using Video Vision Transformers (ViViT). This architecture applies the concept of patches to break videos

into chunks that can be vectorized, similar to text-based tokens. A key aspect of video processing is a specialized type of patch called a "tublet," also known as a space-time patch. While a traditional patch consists of a specific section of a single video frame, a tublet extends this section across multiple frames, capturing spatial and temporal information. Spatial information is contained within each frame, while temporal information is derived from frame changes. This space-time information is then encoded into vectors. Once a video is transformed into a series of these space-time vectors, a transformer classifier leverages its attention mechanism to analyze the footage and classify the content (Arnab et al., 2021).

Diffusion Transformers (DiT) are used to generate AI videos, combining the strengths of the generative diffusion process with the parallel processing benefits of transformer architectures. A Diffusion Transformer also utilizes latent representations during training. A latent representation is a compressed version of an image that reduces computational requirements. The diffusion process begins with a noised latent, passing through a series of DiT blocks that denoise it. Finally, the latent space is transformed into pixel space, generating the video frames (Peebles and Xie, 2023).

WHY HAVE EXPLAINABILITY FOR CLASSIFYING VIDEOS

We need to understand how the model can classify these videos. The model should show its work. Otherwise, a model that can return a label with no explanation might as well be an educated guess. A label based on a guess can have a significant social impact. Consider this model and its ability to identify AI videos. This tool aims to prevent misinformation by labeling videos as fake and stopping them from spreading as accurate. This is a well-intentioned goal, but there can be unintended effects. Every real video that is misclassified as AI-generated can impact content creators by causing a loss of public trust. This can harm their livelihood. When systems only provide labels as feedback, it is difficult for those creators to challenge or correct the mislabeling. Models that show evidence can be held to a higher standard because their mistakes are more transparent. To achieve this transparency, we used a feature importance approach.

DEVELOPING THE EXPLAINABLE TRANSFORMER CLASSIFIER FOR AI-GENERATED CONTENT

Design of the Explainable Transformer Classifier

DiT models inspire our classifier design. DiTs begin with noisy latents, which are partitioned into spacetime patches. These patches are then processed into denoised latents, which are subsequently transformed into pixel-based frames for video reconstruction.

Building on this understanding, we developed a ViViT-based transformer (Figure 1) that compresses video frames into latent space, segments them into spacetime patches, and processes these patches through transformer layers to

determine whether the video is AI-generated. Our classifier consists of three distinct modules.

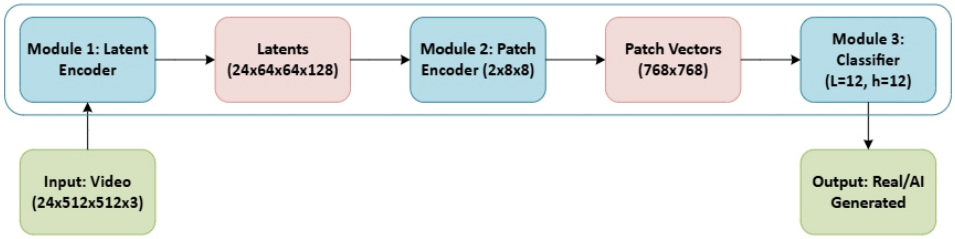

Figure 1: Architecture of the explainable transformer classifier for AI-generated content.

The first module is the latent encoder (Figure 2), which compresses video frames into latent space, reducing their size to one-eighth of the original. This reduces computational requirements in the later stages of the model. The module consists of three convolutional layers, each halving the spatial dimensions while increasing the number of channels from 32 to 128. Each channel captures specific features that are encoded into the latent space.

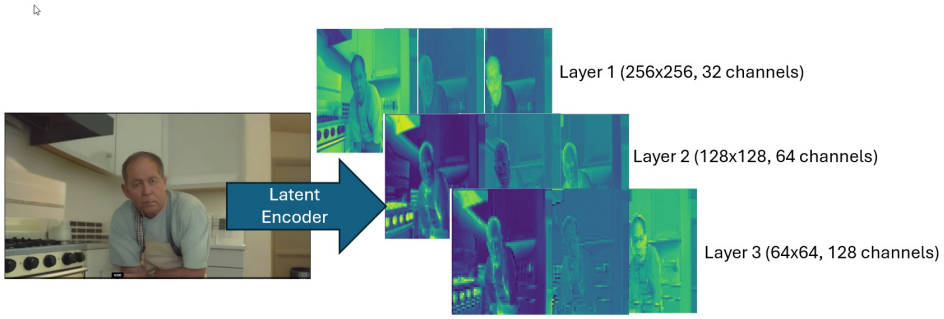

Figure 2: Module 1 latent encoder-three CNNs process each frame to generate learned features in latent space.

The second module is the patch encoder (Figure 3). It partitions the latent representation into $2\times8\times8$ patches, which the transformer vectorizes and processes. This module consists of three additional convolutional layers that transform the patches into one-dimensional vectors of size 768.

The third module is the transformer classifier (Figure 4), following the design outlined in Arnab et al. (2021). The sequence of vectorized patches is fed into 12 transformer layers to determine whether the video is AI-generated. Each layer consists of 12 attention heads. This module serves as the core of the classifier, leveraging the attention mechanism to analyze features from the previous modules and identify patterns indicative of AI-generated content (Vaswani et al., 2017).

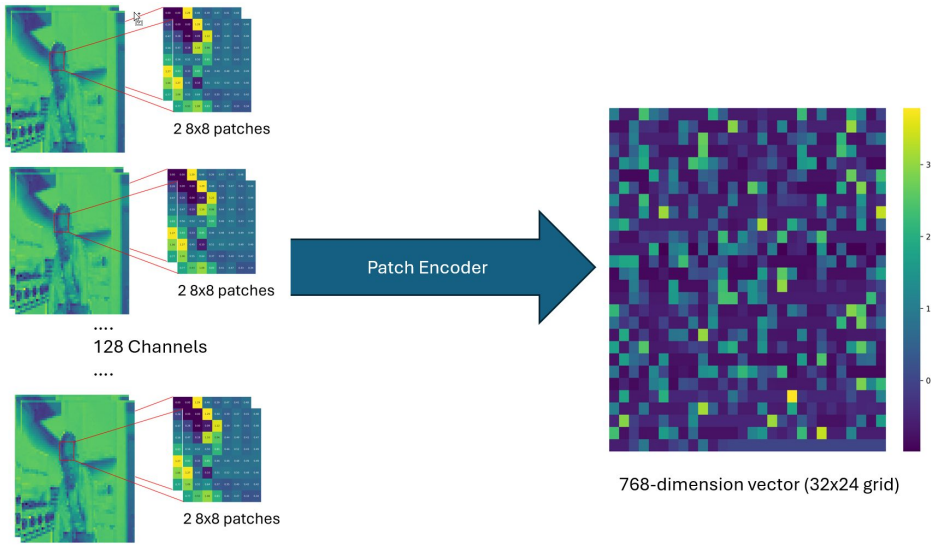

Figure 3: Module 2 patch encoder-two latent frames are divided into 128 × 8 × 8 sections and converted into 768-long vector.

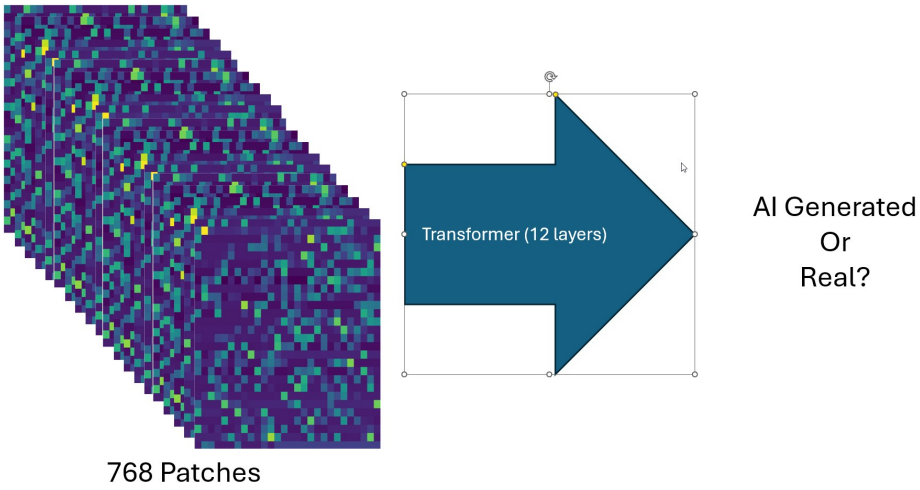

Figure 4: Module 3 transformer classifier—my classifier processes all 768 patches to determine whether they are AI or real.

Training the Explainable Transformer Classifier

To train this classifier, we utilized the GenVideo dataset, a collection of videos labeled as real or AI-generated from various models and sources (Chen et al., 2024). We randomly selected 20,000 videos from the dataset for our experiment: 10,000 real videos from a collection called Youku and 10,000 AI-generated videos from different models, including VideoCrafter, SVD, and Pika (Chen et al., 2024). These videos were then split into training and testing sets using an 80–20 split.

The training environment was an AWS SageMaker accelerated computing instance, ml.g5.12xlarge, equipped with four Nvidia A10G GPUs, each with 24 GB of VRAM. The training was conducted for 10 epochs, with the training set batched into sets of 24. The model extracted 14 million unique patches. Cross-entropy loss was used to measure the model's performance, and the Adam optimizer was employed to adjust the model's weights.

Training Results and Performance Evaluation

The confusion matrix, shown in Figure 5, visually represents the model's classification performance, revealing that the classifier accurately identified most AI-generated and real videos, with minimal misclassifications between the two categories. This demonstrates the model's strong ability to distinguish between authentic and synthetic video content, further supported by its excellent performance during training, achieving an 85% validation accuracy on the testing dataset in its best-performing epoch.

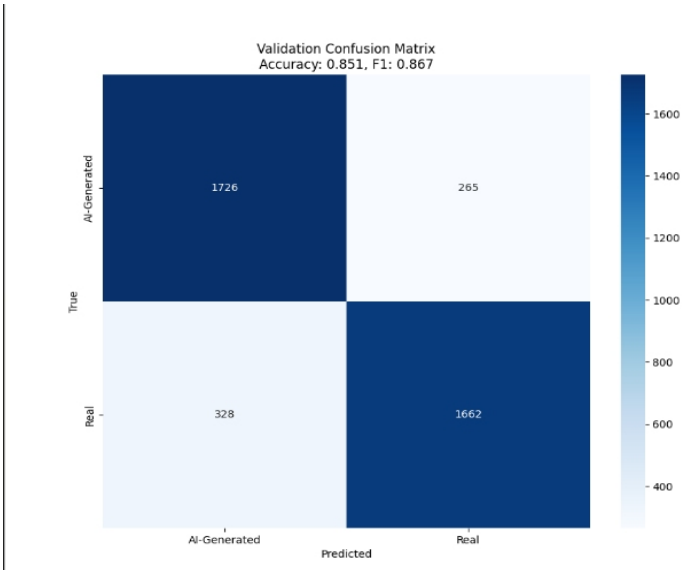

Figure 5: Confusion matrix of the explainable transformer classifier.

Table 1 provides a detailed analysis of the model's performance, including precision, recall, F1 scores for each class, overall accuracy, and macro-averaged metrics. The model achieves high precision and recall, with F1 scores of 0.86 for both AI and Real classes, reinforcing its ability to correctly identify AI-generated and real videos while minimizing false positives and negatives. Overall, it effectively and accurately classifies real and AI-generated videos.

Table 1: Performance analysis of the explainable transformer classifier.

	Precision	Recall	F1-Score
Al-Generated	0.84	0.867	0.853
Real	0.862	0.835	0.849
Accuracy	0.851	0.851	0.851
Macro avg	0.851	0.851	0.867
Weighted avg	0.851	0.851	0.867

Assessing the Interpretability of the Explainable Transformer Classifier

Accuracy is not our sole concern. After training the classifier, we applied Integrated Gradients (IG) to assess its interpretability. IG measures feature importance at the pixel level, scoring all pixels in the videos to determine their significance in the model's classification of AI-generated or real videos. Figure 6 shows a frame from a video that was accurately classified as AI-generated. IG is ideal because, unlike other methods, it satisfies Sensitivity, Implementation Invariance, and Completeness (Sundararajan et al., 2017). These properties ensure that features are comprehensively measured. This method uses a baseline (all-black frames) applied over the video during classification. The classification is then evaluated over a series of steps, where gradients are computed and summed along a straight-line path from 0 (all baseline) to 1 (no baseline) (Sundararajan et al., 2017).

Figure 6: Explainable AI classifier example: the image on the left is from an AI-generated video, and the image on the right shows the activations from the classifier using IG.

CONCLUSION

This paper shows an innovative tool that detects AI-generated video content and demonstrates its results to end users. The AI model this tool created is based on functions and processes of the state-of-the-art generative models.

The tool achieved high accuracy and has been evaluated with Integrated Gradients (IG) to validate the model's ability to explain the process of its determinations. The tool also shows its ability to determine hybrid videos - real videos with added AI-generated elements. With IG, the model can detect AI content and exhibit strong activations on synthetic regions, which is critical for developing more robust tools to handle complex videos.

ACKNOWLEDGMENT

The authors want to express their gratitude to Wayne Smith, a professor from David Nazarian College of Business and Economics of California State University Northridge for his input to this research.

REFERENCES

Adadi, A. Berrada, M. (2018) Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access, Volume: 6. https://doi.org/10.1109/ACCESS.2018.2870052.

Arnab, A. Dehghani, M. Heigold, G. Sun, C. Lucic, M. Schmid, C. (2021). ViViT: A Video Vision Transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). https://doi.org/10.1109/ICCV48922.2021.00676.

Chen, H. Hong, Y. Huang, Z. Xu, Z. Gu, Z. Li, Y. Li, H. (2024). DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark. arXiv preprint arXiv:2405.19707.

Liao, Q. V. Varshney, K. R. (2022). Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv preprint arXiv:2110.10790v5.

Peebles, W. Xie, S. (2023). Scalable Diffusion Models with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV 2023). https://doi.org/10.1109/ICCV51070.2023.00387.

Sundararajan, M. Taly, A. Yan, Q. (2017). Axiomatic Attribution for Deep Networks. arXiv preprint arXiv:1703.01365v2.

Vaswani, A. Shazeer, N. Parmar, N. Uszkoreit, J. Jones, L. Gomez, A. N. Kaiser, Ł. Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).