PRESERVING FORGERY ARTIFACTS: AI-GENERATED VIDEO DETECTION AT NATIVE SCALE

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of video generation models has enabled the creation of highly realistic synthetic media, raising significant societal concerns regarding the spread of misinformation. However, current detection methods suffer from critical limitations. They often rely on preprocessing operations like fixed-resolution resizing and cropping, which not only discard subtle, high-frequency forgery artifacts but can also cause distortion and significant information loss. Furthermore, these methods are frequently trained and evaluated on outdated datasets that fail to capture the sophistication of modern generative models. To address these challenges, we introduce two key contributions: a new large-scale dataset and benchmark, as well as a novel detection framework. We present a comprehensive dataset of over 140K videos from 16 state-of-the-art open-source and leading commercial generators. In addition, we curate Magic Videos Testset, featuring ultra-realistic videos produced through a meticulous generation and filtering pipeline. In addition, we propose a novel detection framework built on the Qwen2.5-VL Vision Transformer, which processes videos at their native spatial resolution and temporal duration. This native-scale approach preserves high-frequency details and spatiotemporal inconsistencies that are often lost during conventional preprocessing. Extensive experiments show that our method achieves state-of-the-art performance across multiple benchmarks. Our work underscores the importance of native-scale processing and establishes a robust new baseline for AI-generated video detection.

1 INTRODUCTION

Artificial Intelligence-Generated Content (AIGC) has advanced rapidly, revolutionizing the creation of high-quality text Yang et al. (2024); DeepSeek-AI (2024), image Esser et al. (2024); Labs (2024), audio Kreuk et al. (2023); Copet et al. (2023) and video Brooks et al. (2024). Among these advancements, video generation has seen particularly significant progress, evolving from foundational models like Stable Diffusion Rombach et al. (2022) to more advanced architectures such as Diffusion Transformers (DiTs) Peebles & Xie (2023); Brooks et al. (2024), as well as proprietary commercial products Pika Labs (2023); Jimeng AI (2024); Kuaishou (2024). These developments have pushed the boundaries of deepfake technologies Yang et al. (2022), enabling large-scale creation of fully AI-generated videos. However, the emergence of near-photorealistic synthetic videos pose serious threats to privacy, reputation, and public trust Wang et al. (2024), underscoring the urgent need for effective detection and mitigation strategies against disinformation and misinformation.

Deepfake detection Yan et al. (2023) and AI-generated image detection Wang et al. (2020); Zhu et al. (2023) have made significant progress in identifying manipulated content. However, existing deepfake detection methods Qian et al. (2020); Xu et al. (2023); Oorloff et al. (2024); Nguyen et al. (2024) often face generalizability issue as they primarily focus on detecting facial forgeries. Meanwhile, approaches for detecting images generated by Generative Adversarial Network (GAN) and diffusion models Wang et al. (2020; 2023c); Tan et al. (2024); Luo et al. (2024) are typically restricted to static media, leaving general spatiotemporal forgery detection largely unaddressed.

Recent studies have begun to develop more robust solutions for AI-generated image and video detection Yan et al. (2025); Li et al. (2025); Song et al. (2024); Chen et al. (2024b); Kundu et al. (2025b). A significant and shared limitation among these methods is the conventional preprocessing of resiz-

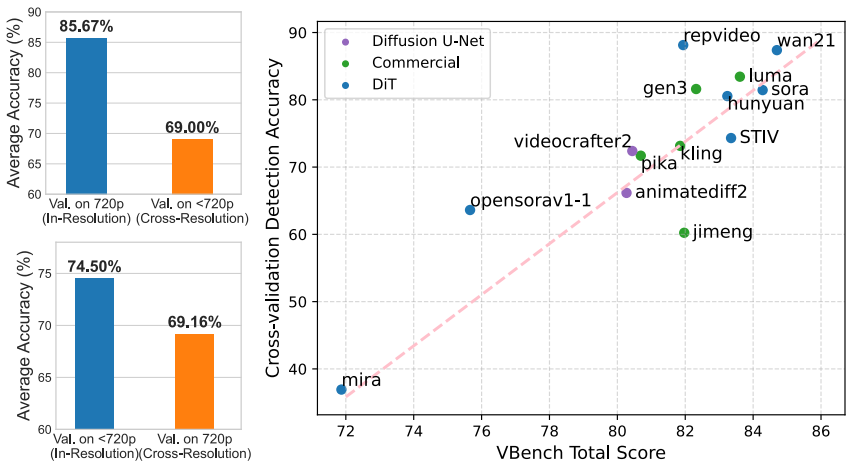


Figure 1: The impact of video resolution and generation quality on detection. **Left:** Models trained on 720p videos (top) and others on <720p videos (bottom) both show a significant drop when validated on a different resolution than they were trained on. **Right:** A strong positive correlation (Pearson ρ=0.86) exists between a generator's quality (VBench score) and cross-validation detection accuracy, indicating that generated videos with higher-quality can assist the training of more effective detectors. These findings highlight the necessity of a general framework that can handle diverse input resolutions and generated artifacts.

ing Yan et al. (2025) or cropping Li et al. (2025) input frames to a fixed resolution (e.g., 224x224). Forgery detection methods often rely on two types of features, subtle artifacts and high-level semantics Cheng et al. (2025). However, this fixed-resolution preprocessing degrades both types of features. Resizing distorts the original aspect ratio, misleading detectors into learning superficial distribution differences rather than robust and generalizable forgery features Rajan et al. (2025). Cropping, meanwhile, can discard important content outside the selected area, thereby destroys global semantic cues of high-resolution content. Furthermore, both downsampling approaches degrade the subtle, pixel-level artifacts that are critical for identifying synthetic media and capturing fine-grained inconsistencies Corvi et al. (2025).

Furthermore, progress in AI-generated video detection is hampered by the use of outdated synthetic data sources. Existing datasets Chen et al. (2024b); Song et al. (2024) are predominantly composed of videos generated by earlier models, which typically exhibit low resolution, limited quality, and short durations. As a result, detection models trained on these datasets experience a significant performance drop when evaluated on modern super-realistic synthetic videos. To better understand these challenges, we conduct cross-validation experiments using existing detectors on a synthetic videos dataset sourced from 14 generative models. Our preliminary results reveal two critical insights, as illustrated in Figure 1. First, we observe a significant performance drop when detectors are evaluated on videos with different resolution. Second, detection performance is positively correlated with the quality of the video generators, meaning that stronger detectors require training on higher-quality, more realistic synthetic videos. These findings further highlight the importance of constructing a high-quality and diverse dataset, as well as a training framework capable of effectively handling videos with diverse resolutions, durations and generative sources.

In response to the limitations of existing methods, we propose a unified framework that supports training and evaluation on videos with diverse resolutions and generative sources. First, we curate a high-quality and diverse video dataset sourced from 15 representative video generation models for training and develop a meticulously crafted pipeline to synthesize high-quality, human-indistinguishable videos for evaluation, termed Magic Videos. Second, we design a native-resolution training framework based on the Qwen2.5-VL Vision Transformer Bai et al. (2025), which unifies image and video modeling and enables the model to natively process videos with arbitrary spatial resolutions and temporal lengths. By removing the constraints of fixed-size downsampling preprocessing, our method achieves strong generalization capabilities to capture general spatiotemporal forgery artifacts. Extensive experiments on a wide range of benchmarks (Genvideo Chen et al. (2024b), DVF Song et al. (2024) and our proposed Magic Videos) demonstrate that our model is ro-

bust and achieves state-of-the-art performance in detecting AI-generated videos. Our Contributions are summarized as follows:

- We introduce a new high-quality and diverse dataset, sourcing videos from 18 state-of-the-art generative models, ensuring that both training and evaluation are aligned with the current generative quality of AIGC.
- We propose a novel native-resolution training framework built upon the Qwen2.5-VL Vision Transformer, which processes videos in their original spatial resolutions and temporal lengths, preserving crucial forgery artifacts often lost during conventional resizing or cropping.
- Through extensive experiments, we demonstrate that our method achieves state-of-the-art performance and robust generalization across a wide range of benchmarks, setting a new standard for AI-generated video detection.

2 Related Work

2.1 Video Generative Models

Diffusion models Ho et al. (2020); Song et al. (2022); Rombach et al. (2022) have significantly enhanced the quality and controllability of image generation, inspiring researchers to extend these techniques to video generation tasks. Early workSinger et al. (2022) propose incorporating motion dynamics into pre-trained text-to-image generation models. More recent studies Chen et al. (2024a); Guo et al. (2024); Blattmann et al. (2023); Wang et al. (2023a); Wei et al. (2024) leverage latent-based diffusion models Rombach et al. (2022) to generate short dynamic videos from text or image inputs. With the growing popularity of Diffusion Transformers (DiTs) Peebles & Xie (2023) in image generation Labs (2024), DiT and its variants Esser et al. (2024) have been widely proposed for video generation tasks Ma et al. (2024b); Zheng et al. (2024); Brooks et al. (2024); Yang et al. (2025); Kong et al. (2024); Team (2025); Polyak et al. (2024). Besides Diffusion based methods, Generative Adversarial Network (GAN) Shen et al. (2023); Wang et al. (2023b) are also explored for video generation. The success of decoder-only architecture in language model has also motivated research in generating long videos using autoregressive models Kondratyuk et al. (2024); Yu et al. (2023); Yin et al. (2025). Commercial video generation products Brooks et al. (2024); Kuaishou (2024); Jimeng AI (2024); Pika Labs (2023); MiniMax (2024), employ complex and proprietary pipelines and produces hyper-realistic videos. However, the lack of transparency surrounding these systems limits detailed analysis of their methodologies.

In this paper, we propose a generative video dataset that encompasses most of the aforementioned architectures, including Diffusion U-Net Chen et al. (2024a); Guo et al. (2024), DiT Brooks et al. (2024); Team (2025); Ju et al. (2024); Zheng et al. (2024); Polyak et al. (2024); Lin et al. (2024); Ma et al. (2025), MMDiT Kong et al. (2024); Si et al. (2025), and auto-regressive Yin et al. (2025) models. The diversity of generative models included in our dataset ensures broad coverage and supports the generalizability of the proposed method.

2.2 AI-Generated Image and Video Detection

Generated Image Detection. As generative technologies rapidly advance, a growing number of forged images are now entirely synthesized by GANs Goodfellow et al. (2014) and Diffusion models Rombach et al. (2022), moving beyond traditional limited manipulation techniques. Consequently, substantial research efforts have focused on developing generalizable synthetic image detection methods Tan et al. (2023); Ojha et al. (2023); Yan et al. (2024); Liu et al. (2024b), including approaches based on reconstruction error Wang et al. (2023c); Luo et al. (2024); Guillaro et al. (2025), pixel-level features Wang et al. (2020); Tan et al. (2024); Cheng et al. (2025), or adapting visual backbones Koutlis & Papadopoulos (2024); Yan et al. (2025); Liu et al. (2024a). These methods are typically trained on images generated by specific models Karras et al. (2018); Song et al. (2022) and aim to achieve cross-architecture generalization.

Generated Video Detection. More recently, research has expanded to the detection of fully AI-generated videos Ma et al. (2024a); Ni et al. (2024); Ji et al. (2024); Chang et al. (2025). VLM-based

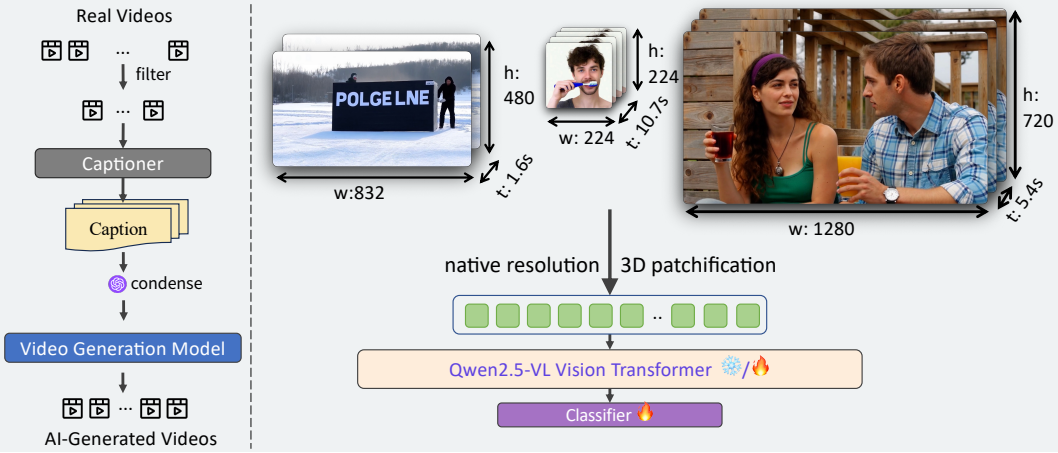


Figure 2: **An overview of our data generation and detection methodology. Left**: High-quality captions are curated from real videos and refined before being fed into state-of-the-art video generation models, which produce highly realistic synthetic videos for evaluation. **Right**: Our proposed detection framework operates at native resolution. Unlike conventional methods that resize frames to inputs with a fixed resolution and temporal duration, our framework processes videos into 3D patches at their native spatial resolution and temporal duration. Built on the Qwen2.5-VL ViT, our approach preserves subtle forgery artifacts that are essential for robust detection.

methods Song et al. (2024); Wen et al. (2025) prompts large vision language models to identify unnatural AI-like clue, while ViT-based methods Chen et al. (2024b); Corvi et al. (2025) introduce forgery-posed generated datasets and design modules to detect spatial-temporal inconsistencies. However, existing methods for detecting AI-generated images and videos commonly suffer from a reliance on fixed resizing operations. Such preprocessing can lead to the loss of fine-grained details and spatial distortions, ultimately compromising model robustness across diverse inputs. In this work, we address this issue by training on native spatial resolution and temporal duration, without resizing or temporal padding. This design fundamentally avoids the pitfalls of conventional preprocessing and significantly enhances the model's generalization capability.

3 Methodology

3.1 Data Curation

To address the lack of comprehensive datasets for AI-generated video detection, we collect and construct a large-scale dataset comprising over 140,000 videos. This dataset includes more than 70,000 AI-generated videos from 18 distinct generative models, balanced with an equal number of real videos sampled from five high-quality sources: MSVD Chen & Dolan (2011), Kinetics Kay et al. (2017), Panda-70M Chen et al. (2024d), Mixkit and Pexels. The AI-generated portion covers 18 mainstream generative models developed between September 2023 and June 2025, including early Diffusion U-Net models (AnimateDiff Guo et al. (2024), VideoCrafter2 Chen et al. (2024a), mira Ju et al. (2024)), recent DiT-based models (Open-Sora Zheng et al. (2024), Sora Brooks et al. (2024), MovieGen Polyak et al. (2024), HunyuanVideo Kong et al. (2024), Apple-STIV Lin et al. (2024), CausVid Yin et al. (2025), Step-Video Ma et al. (2025), Wan2.1 Team (2025), and leading commercial models Pika Labs (2023); Jimeng AI (2024); Kuaishou (2024); Lumalabs (2024); Germanidis (2024); MiniMax (2024) that are accessible only via web interfaces or APIs. The generated videos vary in duration from 1 to 12 seconds and in resolution from 240p to 1080p, ensuring a diverse and representative collection of AI-generated content. A detailed breakdown of the dataset's composition is provided in the Appendix.

Selection of Data Sources. The AI-generated videos are curated from multiple sources: (1) VBench Huang et al. (2023; 2024), which provides generated videos from various text-to-video models using a predefined suite of diverse prompts; (2) Movie Gen Polyak et al. (2024), which contributes videos generated by its proprietary model; and (3) A collection of highly realistic videos

Data	Source	Number	Resolution	Duration
Training Data (15 models)	Vbench	140K	240p-768p	1-10s
Movie Gen	MovieGenBench	2003	1920x1088	10.7s
Wan2.1	Open Source	450	1280x720	5.4s
Wan-1.3B	Open Source	584	832x480	5s
Hailuo	API (T2V-01)	450	1280x720	5.6s
Seaweed	API (Jimeng-S2.0)	450	1472x832	5s
Seedance	API (Jimeng-S3.0)	450	1248x704	5s
StepVideo	Open Source	450	950x540	8.2s

Table 1: **Statistics of our proposed Magic Videos Testset**. Please refer to Appendix for details of our training data.

synthesized using various cutting-edge open-source and commercial models, guided by our custom-designed prompt library.

Realistic Video Generation Pipeline. To evaluate the capability of generative content detectors in real-world scenarios, we design a pipeline for constructing synthetic videos that closely resemble authentic content. We prioritize scenarios that pose significant risks to information security, such as realistic landscapes, architectural scenes, and human interactions, as these categories are particularly susceptible to misuse and misinformation due to their inherent plausibility. Leveraging ShareGPT4Video Chen et al. (2024c) repository of detailed and high-quality captions, we curate content specifically within these realism-oriented themes. To accommodate the capabilities of state-of-the-art architectures, we filter videos by duration (3-12 seconds) and caption length (fewer than 1000 characters). The curated prompts are further optimized using GPT-4o to condense the description to under 500 characters. Table 1 summarizes videos that are synthesized by six distinct video generators using our comprehensive prompt library. These videos represent the current frontier of photorealistic synthetic content, enabling a rigorous assessment of detection models under practical and high-risk conditions.

3.2 QWEN2.5-VL VIT

Contemporary AI-generated content detectors primarily operate by identifying two categories of features: local artifacts and global semantic inconsistencies Cheng et al. (2025). However, a common practice in existing methodologies is to resize input images to a low, fixed resolution, typically 224x224 pixels. This downscaling operation adversely affects the features crucial for detection: it degrades subtle local artifacts and distorts global semantic structures. In this paper, we introduce a unified framework that processes images and videos at native resolution, thereby preserving the original forgery artifacts. The framework begins by tokenizing input videos into 3D patches at the native scale and adopts Qwen2.5-VL ViT Bai et al. (2025) as a novel visual backbone for general video forgery detection.

3D Video Patchifying at Native Scale. We follow the video processing steps of Bai et al. (2025), which introduces a 3D patch partitioning strategy that enables native-resolution inputs. For static images, it employ a standard spatial patch extraction method (e.g., 14x14 pixels). Unlike conventional ViTs that operate on static frames independently, our method extends patchification into the temporal dimension for video data. Given an input video tensor $V \in \mathbb{R}^{T\times H\times W\times C}$, it partitions V to non-overlapping 3D patches of size $(P_t, P_h, P_w) = (2, 14, 14)$ and computes patch embedding via linear projection matrix E. This design eliminates the need for conventional resizing and padding operations, allowing the Transformer model to operate natively on both the spatial and temporal scales. The initial transformation of a raw video tensor V into a sequence of feature embeddings $X^{(0)}$ is described in Equation (1). The 3D patchification is particularly effective in detecting subtle texture artifact and minimal temporal consistencies at the patch level. By preserving the original resolution during preprocessing, our method ensures potential features critical for forgery detection remain intact and undistorted.

$$X^{(0)} = \text{Unfold}(V; P_t, P_h, P_w)^T \cdot E \tag{1}$$

Transformer Layer Structure. Qwen2.5-VL ViT consists of 32 Transformer layers, each adopting a pre-normalization structure, in which RMSNorm is applied before both the self-attention and feed-forward network (FFN). The FFN component employs the SwiGLU activation function. To effectively encode the spatial relationships between patches, 2D Rotary Positional Embedding (RoPE) Su et al. (2023) is applied to the queries and keys in self attention, enhancing the model's extrapolation capability across input resolutions. The computations performed within each Transformer layer are described as

$$\begin{aligned} \hat{X}^{(l)} &= X^{(l-1)} + \text{Attention}(\text{RMSNorm}(X^{(l-1)})), \\ X^{(l)} &= \hat{X}^{(l)} + \text{FFN}_{\text{SwiGLU}}(\text{RMSNorm}(\hat{X}^{(l)})). \end{aligned} \tag{2}$$

In Equation (2), $X^{(l-1)}$ and $X^{(l)}$ denote the input and output hidden states of the l-th Transformer layer, respectively.

Infrastructure Optimization for Efficiency. To address the computational challenges associated with high-resolution inputs, which typically lead to quadratic complexity, several optimizations are integrated. A batch packing strategy from NaViT Dehghani et al. (2023) is adopted to allow the model to handle variable-length sequences without padding or attention masks. This is combined with Flash Attention Dao (2023), enabling GPU awareness of sequence boundaries and significantly improving both computational efficiency and memory usage through optimized CUDA kernels. In addition, a hybrid attention strategy is adopted where the majority of Transformer layers utilize 114×114 windowed attention, ensuring that the computational cost scales linearly with the number of input patches.

Classifier and Tuning Methods. For the final binary classification task of distinguishing between authentic and AI-generated content, we append a simple yet effective classification head to the Qwen2.5-VL ViT backbone. The output tokens from the final Transformer layer is first aggregated into a single, fixed-size feature vector using global average pooling. This vector is then passed through a single fully connected (FC) linear layer that outputs the logits corresponding to the "real" and "generated" classes. To adapt the pre-trained model to this task, we explore three fine-tuning strategies: (1) Full Finetuning: Both the visual backbone and classification head are jointly optimized during training. (2) Linear-Probing: Serves as a baseline, where the entire vision backbone is frozen and only the classification head is trained. (3) Parameter-Efficient Fine-Tuning (PEFT): Specifically. we adopt Low-Rank Adaptation (LoRA Hu et al. (2021)), which introduces small, trainable low-rank matrices into the frozen backbone, allowing only a subset of parameters to be updated.

4 EXPERIMENTS

4.1 DATASETS

Traning Set. We construct a training set of 70K AI-generated videos and 70K real videos. The synthetic videos are generated by VBench Huang et al. (2023) using their prompt set, while the real

Model	Training Data	Movie Gen	Wan 2.1	Wan-1.3B	Hailuo	Seaweed	Seedance	StepVideo	mACC
RINE†	ldm	52.97	49.07	45.03	50.70	48.37	48.60	48.37	48.36
FatFormer†	ProGAN	50.02	50.23	45.55	50.00	50.00	50.00	50.23	49.34
B-Free†	SD 2.1	64.30	56.74	72.14	60.93	28.60	36.51	51.86	51.13
Effort†	GenImage(SD1.4)	70.74	81.40	29.97	70.93	85.58	76.74	50.23	65.81
F3Net		92.51	67.44	66.95	67.67	69.53	70.00	59.53	66.85
TALL		91.71	52.33	58.05	56.98	56.98	58.14	53.26	55.96
NPR		92.66	70.47	50.68	70.23	73.02	71.63	66.51	67.09
TimeSformer		91.41	66.74	64.55	61.86	68.84	66.28	67.91	66.03
SAFE	15Model-140K	91.76	72.33	52.91	71.40	56.51	80.70	74.42	68.41
CLIP ViT-L/14	(Ours)	99.20	76.98	60.62	76.05	66.51	77.67	77.21	<u>72.51</u>
X-CLIP-B/16		98.55	70.23	60.10	71.40	72.79	74.42	62.33	68.55
X-CLIP-L/14		98.85	74.65	59.93	74.42	64.65	74.42	73.26	70.22
Moon-ViT		98.25	75.35	59.76	75.12	74.88	74.42	73.49	72.17
Qwen2.5-ViT (Ours)		97.20	**85.12**	**72.26**	**83.26**	**84.65**	**84.19**	**77.67**	**81.19**

Table 2: **Benchmarking Results in terms of ACC Performance on the Movie Gen valid set and Magic Videos Testset**. † indicates that the results are obtained by using the official pre-trained model.

Model	Training Data	Movie Gen	Wan 2.1	Wan-1.3B	Hailuo	Seaweed	Seedance	StepVideo	mAP
RINE†	ldm	71.11	38.70	35.62	54.43	33.69	35.53	34.14	38.69
FatFormer†	ProGAN	58.84	48.04	34.01	54.50	46.52	38.93	57.73	46.62
B-Free†	SD 2.1	70.38	63.76	83.58	68.21	35.24	39.77	55.99	57.76
Effort†	GenImage(SD1.4)	80.60	89.86	34.09	78.71	92.65	86.83	57.04	73.20
F3Net		96.20	75.93	74.81	72.48	82.88	79.61	62.90	74.77
TALL		96.07	82.36	70.23	74.97	88.63	81.32	82.36	79.98
NPR		97.10	78.74	50.82	76.93	87.33	79.40	73.71	74.49
TimeSformer		96.91	86.62	74.29	75.92	87.11	82.15	90.04	82.69
SAFE	15Model-140K	96.55	76.77	85.27	71.63	58.22	86.93	80.56	76.56
CLIP ViT-L/14	(Ours)	99.95	92.02	63.45	92.74	90.84	95.35	**94.81**	88.20
X-CLIP-B/16		99.87	81.08	87.88	87.37	86.24	95.25	64.96	83.80
X-CLIP-L/14		99.94	**96.91**	81.30	**93.83**	69.26	**96.95**	93.66	<u>88.65</u>
Moon-ViT		99.24	92.18	74.17	89.70	88.12	90.46	86.19	86.80
Qwen2.5-ViT (Ours)		99.46	96.11	**91.51**	90.83	**92.38**	94.59	77.31	**90.46**

Table 3: **Benchmarking Results in terms of AP Performance on the Movie Gen valid set and Magic Videos Testset**. † indicates that the results are obtained by using the official pre-trained model.

ones are sampled from MSVD Chen & Dolan (2011) and Kinetics Kay et al. (2017). For validation, we use 1,003 fake videos from MovieGenVideoBench Polyak et al. (2024) and 1,000 real videos from Panda-70M Chen et al. (2024d). To comprehensively evaluate detection performance, we introduce three test sets spanning different generations of video models.

Test Sets. To benchmark robustness against state-of-the-art synthetic content, we curate the **Magic Videos** (Table 1), containing high-quality, hyper-realistic videos generated by six cutting-edge video generators with carefully filtered prompts. Each subset is paired with corresponding real videos, and performance is reported using Accuracy (ACC) and Average Precision (AP). **GenVideo-Val and DVF-Test:** We also evaluate on the test sets of two external datasets: DVF Song et al. (2024) and GenVideo Chen et al. (2024b). These datasets include videos from models released before September 2024, offering insights into the detector performance on earlier synthetic content. These sets provide a complementary, historically grounded perspective on detector performance.

Baselines. We benchmark four categories of methods: (1) AI-generated video detection methods (MM-Det Song et al. (2024), DeMamba Chen et al. (2024b), UNITE Kundu et al. (2025b), and TruthLens Kundu et al. (2025a)); (2) visual foundation backbones (X-CLIP-B/16 Ni et al. (2022), X-CLIP-L/14 Ni et al. (2022), TimeSformer Bertasius et al. (2021), and Moon-ViT Team et al. (2025)); (3) facial forgery detection methods (TALL Xu et al. (2023) and F3Net Qian et al. (2020)); and (4) general AI-generated image detection methods (NPR Tan et al. (2024), FatFormer Liu et al. (2024a), RINE Koutlis & Papadopoulos (2024), B-Free Guillaro et al. (2025), and Effort Yan et al. (2025)). For image-based methods, we average the logits across T frames to obtain video-level predictions.

Implementation Details. We train our model for 3 epochs using binary cross-entropy loss and the AdamW optimizer. The learning rate is set to 1e-5 for full fine-tuning and 1e-4 for PEFT tuning. To balance performance and computational cost, we follow the preprocessing steps described in Bai et al. (2025); Team et al. (2025), which define the minimum and maximum token budgets for images. Input frames are resized to the highest possible resolution within the range `(min_pixels, max_pixels)`, while preserving the original aspect ratio. In our experiments, we set the resolution range to $(224 \times 224, 720 \times 720)$. For temporal sampling, videos are sampled at 2 fps. We select either $T = 8$ consecutive frames randomly during training or centrally during testing. Additional implementation details for other baseline models are provided in the Appendix.

4.2 AI-Generated Video Detection

Evaluation on Magic Videos. The experimental results presented in Table 2 and Table 3 offer a comprehensive evaluation of our model against several distinct classes of detection methods. A notable observation is the underwhelming performance of models originally developed for AI image detection, including RINE, FatFormer, B-Free, and Effort. These models exhibit relatively poor performance on video-based benchmarks, even compared to image-based methods that are trained on our video datasets. This suggests a fundamental difference between forgery patterns present in static

Method	Video-Crafter	Zero-scope	Open-Sora	Sora	Pika	Stable Diff.	Stable Video	AVG
CNNDet*	87.4	88.2	78.0	63.8	77.3	73.5	78.9	78.2
DIRE*	55.9	61.8	53.8	60.5	65.8	62.7	69.9	62.1
MM-Det	93.5	94.0	88.8	86.2	95.9	**95.7**	89.9	92.0
NPR	86.6	85.6	96.0	81.0	94.6	71.1	97.0	87.4
TALL	**95.4**	91.8	97.2	94.9	97.5	83.6	98.2	92.6
F3Net	90.4	90.2	95.9	90.1	97.8	93.1	98.5	93.7
TimeSformer	94.5	92.7	98.0	92.5	98.4	92.4	99.5	95.4
Ours	93.5	**99.8**	**98.6**	**96.4**	**99.1**	95.6	**99.7**	**97.6**

Table 4: **Benchmarking Results in terms of AUC Performance on DVF-Test Song et al. (2024)**. Results with * are derived from Song et al. (2024).

Model	Averaged			Overall	
	Recall	F1	AP	ACC	Recall
UNITE	89.60	-	92.76	-	-
TruthLens	-	-	-	90.49	-
DeMamba-CLIP	**91.58**	89.19	93.45	96.14	92.29
NPR	83.01	47.99	63.66	86.75	92.40
F3Net	83.48	56.78	71.57	88.26	93.06
TimeSformer	86.42	65.38	77.67	87.51	91.55
TALL	89.44	61.51	76.67	90.05	91.76
XCLIP-B	89.79	53.76	72.04	92.60	90.90
Qwen2.5-ViT (Ours)	91.16	**90.64**	**96.13**	**96.64**	**93.18**

Table 5: **Benchmarking Results in terms of averaged Recall, F1, AP per subset and overall Recall and ACC Performance on Genvideo-Val Chen et al. (2024b)**. Detailed results are in Appendix.

Method	ACC	Fake ACC	Real ACC
GPT-5	90.7	84.6	98.8
GPT-4.1	92.9	89.1	97.9
Gemini 2.5 Pro	84.3	75.7	95.8
Qwen2.5-VL 7B	51.7	20.2	93.4
Qwen2.5-VL 72B	50.0	16.6	94.3
DeepTraceReward (w/ Qwen2.5 VL 7B)	74.7	55.7	100.0
Qwen2.5-ViT (Ours)	97.2	96.3	98.2

Table 6: **Benchmarking Results in terms of ACC on DeepTraceReward Fu et al. (2025)**. Results of baseline methods are reported in Fu et al. (2025).

images and those in dynamic video sequences. Features learned for detecting image artifacts do not generalize well to the spatio-temporal domain required for video-level analysis. Similarly, methods designed specifically for deepfake detection, such as F3Net and TALL show limited effectiveness. While these models excel at identifying at facial manipulations, their specialization becomes a constraint when faced with the broader challenge of detecting fully synthesized videos. In contrast, large-scale visual backbones like TimeSformer, CLIP-ViT and X-CLIP demonstrate competitive performance, leveraging extensive pre-training on diverse visual data. However, their effectiveness is ultimately constrained by architectural limitations. A key issue is the conventional practice of resizing input frames to a fixed resolution of 224×224 pixels. This downsampling may eliminate subtle pixel-level forgery artifacts and disrupt global semantic features that are crucial for detecting sophisticated generative content. Moon-ViT Team et al. (2025), which applies a similar processing pipeline based on NaViT, also suffers from this limitation as it operates on static images and cannot capture temporal inconsistencies. Our proposed method achieves the highest average scores in both ACC and AP, establishing a new state-of-the-art on these benchmarks. While our model does not achieve the best AP on every individual generator, it consistently delivers strong performance across all generator types, highlighting its robust generalizability. This superior performance is attributed to its advanced architecture. By leveraging the Qwen2.5-ViT backbone, our model integrates native-resolution modeling with dynamic temporal duration modeling, avoids destructive downsampling and preserves the fidelity of forgery cues present in the original content. By effectively capturing both fine-grained artifacts and high-level semantic inconsistencies, our model delivers a more robust and accurate solution for detecting AI-generated videos.

Evaluation on DVF-test. We train our model and four competing methods on our training set, excluding any data originating from the Sora and Pika generators, and evaluate them on the DVF test set. The results are presented in Table 4. Notably, despite the imbalance between real and generated samples in certain subsets of DVF-test, we report the Area Under the ROC Curve (AUC) for consistent and direct comparison. Our model achieves the highest average AUC of 97.6, demonstrating

Archs.	Variants	Magic	Genvideo	Avg.
spatial resolution	random crop to 224p	52.41	93.50	72.96
	random resize to 224p	67.90	95.52	81.71
	dynamic [224p, 448p]	77.01	96.01	86.51
	dynamic [224p, 720p]	**81.19**	**96.64**	**88.92**
temporal resolution	T=2	71.20	94.70	82.95
	T=4	73.46	94.40	83.93
	T=8	**77.01**	**96.01**	**86.51**
tuning mode	LP	62.04	91.91	76.98
	LoRA(r=16)	73.14	94.95	84.05
	full	**77.01**	**96.01**	**86.51**

Table 7: **Ablation studies regarding spatial-temporal resolution and tuning mode.** We report averaged ACC(%) on Magic Testset and Genvideo. For temporal and tuning experiments, the spatial resolution is set to dynamic[224p, 448p].

the high quality of our training dataset and the strong generalizability of our model in detecting AI-generated videos across diverse generation techniques.

Evaluation on GenVideo-Val. We train our method and five baseline models on our curated training set, excluding generators that appear in the GenVideo-Val set. Due to the substantial class imbalance between real and generated samples in the GenVideo evaluation subsets, we report both overall Recall and Accuracy (ACC) to enable a more comprehensive comparison. As shown in Table 5, our method outperforms all baselines, including larger MLLM-based models such as TruthLens Kundu et al. (2025a) and other baseline methods trained on the same data as ours. Despite using only 1/20th of the training data scale employed by DeMamba Chen et al. (2024b), our model showcases superior effectiveness and generalizability to earlier generated content.

Evaluation on DeepTraceReward. To further demonstrate robustness against unseen generators, we evaluated our method on the DeepTraceReward Fu et al. (2025), which contains 4,335 videos from 7 recent generators (including Pika-1.5, Kling-1.5, etc). Table 6 compares our Qwen2.5-ViT against leading multimodal LLMs. Our model achieves 97.2% accuracy, significantly outperforming massive foundation models (e.g., GPT-5, Gemini 2.5 Pro) on the binary classification task. Moreover, while general-purpose VLMs often struggle with detecting fakes (showing lower Fake ACC), our model demonstrates balanced performance (96.3% Fake ACC vs. 98.2% Real ACC), proving its effectiveness in identifying artifacts from the latest generation engines without overfitting to specific training generators.

4.3 Ablation Study and Analysis

We conduct a series of ablation studies, as detailed in Table 7, to systematically investigate the impact of spatial resolution, temporal resolution, and different fine-tuning strategies on our model's performance. **Dataset Difference.** The performance gains from using higher-fidelity inputs are substantially larger on our Magic dataset. This is because GenVideo, with its lower native resolution and shorter clips, is less susceptible to the performance degradation caused by downsampling preprocessing. **Ablation Study on Spatial Resolution.** Our analysis reveals critical performance differences. The conventional random crop to 224p method yields the lowest average accuracy on high-resolution content (52.41). Switching to random resize to 224p boosts performance to 67.90, but this approach can still cause degradation of subtle artifacts. In contrast, our dynamic resolution strategy, which preserves the original aspect ratio, demonstrates markedly superior performance, with the average accuracy peaking at 88.92 when using resolutions up to 720p. This confirms our hypothesis that maintaining aspect ratio and processing at higher resolutions are critical for capturing subtle, pixel-level forgery artifacts. **Ablation Study on Temporal Resolution.** For all candidates, we sample the original videos at 2 fps and select random or center-aligned T frames during training and testing, respectively. We observe that incorporating more temporal context is beneficial. Increasing the number of sampled frames (T) from 1 to 8 improves the average performance from 82.95 to 85.91. This suggests that longer sequences enhance the model's ability to detect temporal

inconsistencies common in AI-generated videos. **Ablation Study on Tuning method.** Regarding tuning strategies, full fine-tuning achieves the best average performance (85.91). Although the parameter-efficient LoRA approach (81.10) significantly outperforms linear probing (76.98), full fine-tuning is justified for maximizing detection accuracy.

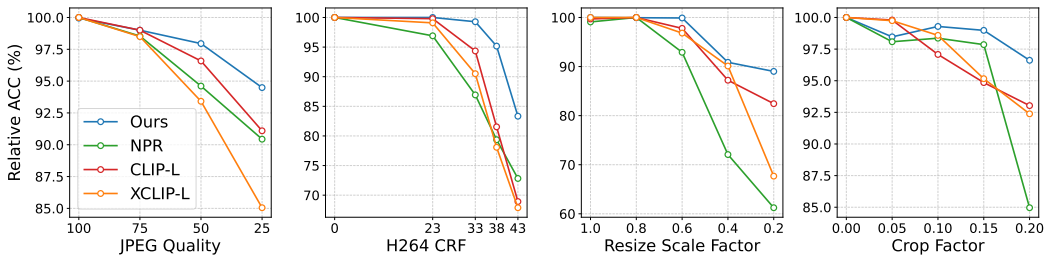


Figure 3: **Robustness Comparison of Relative ACC on MovieGen Under Diverse Perturbation.**

Figure 4: **Saliency Analysis.** Saliency Maps of Our Model on AI-Generated Video Samples.

Robustness Analysis. We evaluate our model's robustness under common video perturbations, including compression, downscaling, and cropping, as shown in Figure 3. The model remains highly accurate under mild degradations such as moderate JPEG and H.264 compression. Performance drops become more pronounced with severe spatial changes. Notably, our model outperforms baselines under aggressive downscaling (scale ≤ 0.4) and cropping (crop factor ≥ 0.15), though all methods are affected by extreme spatial loss. These results highlight strong robustness to moderate noise and sensitivity to substantial spatial degradation.

Saliency Analysis. We examine the model's attention responses to better understand its discriminative behavior, as illustrated in Figure 4. The results confirm that our native-resolution framework effectively captures two key types of features crucial for AIGC detection. (1) Low-level Artifacts: In billboard scenes, the model focuses on fine details such as distorted text rendering and unnatural edge transitions that are often lost during resolution downsampling. These high-frequency artifacts are indicative of generation errors and are critical for reliable detection. (2) High-level Semantics: In the fruit-cutting examples, the model attends to global inconsistencies, including object deformations and unrealistic lighting, suggesting it captures holistic content-level anomalies. This dual focus demonstrates that our approach leverages both spatial fidelity and semantic context, validating the design choice of preserving native resolution.

5 Conclusion

In this work, we address two critical weaknesses in AGIC detection: the reliance on outdated training datasets and the destructive practice of resizing inputs to a fixed, low resolution. Our primary contributions are two-fold: the construction of a comprehensive and up-to-date dataset comprising AI-generated videos from 18 diverse generators, and the development of a novel detection framework that operates directly on videos at their native resolution and temporal length. Experimental results demonstrate that our method, built upon Qwen2.5-VL ViT backbone, establishes a new state-of-the-art across three established benchmarks. Crucially, by avoiding downsampling, our model preserves both fine-grained artifacts and high-level semantic inconsistencies, resulting in significantly improved robustness and generalization to recent advances in generative video content.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, pp. 813–824, 2021.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. URL `https://openai.com/index/sora/`.

Chirui Chang, Jiahui Liu, Zhengzhe Liu, Xiaoyang Lyu, Yi-Hua Huang, Xin Tao, Pengfei Wan, Di Zhang, and Xiaojuan Qi. How far are ai-generated videos from simulating the 3d visual world: A learned 3d evaluation approach. In *ICCV*, 2025.

David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *ACL*, pp. 190–200, 2011.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pp. 7310–7320, 2024a.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. Demamba: Ai-generated video detection on million-scale genvideo benchmark, 2024b.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024c.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, 2024d.

Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai. In *CVPR*, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. Simple and controllable music generation. In *NeurIPS*, volume 36, pp. 47704–47720, 2023.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable ai-generated video detection with forensic-oriented augmentation, 2025.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023.

DeepSeek-AI. Deepseek-v3 technical report, 2024. URL `https://arxiv.org/abs/2412.19437`.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim Alabdulmohsin, Avital Oliver, Piotr Padlewski, Alexey Gritsenko, Mario Lučić, and Neil Houlsby. Patch n' pack: Navit, a vision transformer for any aspect ratio and resolution. In *NeurIPS*, 2023.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, Keyush Shah, Chung Un Lee, Yejin Choi, James Zou, Dan Roth, and Chris Callison-Burch. Learning human-perceived fakeness in ai-generated videos via multimodal llms, 2025.

Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, Xunsong Li, Yifu Li, Shanchuan Lin, Zhijie Lin, Jiawei Liu, Shu Liu, Xiaonan Nie, Zhiwu Qing, Yuxi Ren, Li Sun, Zhi Tian, Rui Wang, Sen Wang, Guoqiang Wei, Guohong Wu, Jie Wu, Ruiqi Xia, Fei Xiao, Xuefeng Xiao, Jiangqiao Yan, Ceyuan Yang, Jianchao Yang, Runkai Yang, Tao Yang, Yihang Yang, Zilyu Ye, Xuejiao Zeng, Yan Zeng, Heng Zhang, Yang Zhao, Xiaozheng Zheng, Peihao Zhu, Jiaxin Zou, and Feilong Zuo. Seedance 1.0: Exploring the boundaries of video generation models, 2025.

Anastasis Germanidis. Runway gen-3, 2024. URL https://runwayml.com/research/introducing-gen-3-alpha.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014.

Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection. In *CVPR*, 2025.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench++: Comprehensive and versatile benchmark suite for video generative models, 2024.

Lichuan Ji, Yingqi Lin, Zhenhua Huang, Yan Han, Xiaogang Xu, Jiafei Wu, Chong Wang, and Zhe Liu. Distinguish any fake videos: Unleashing the power of large-scale data and motion features, 2024.

Jimeng AI. https://jimeng.jianying.com/ai-tool/home, 2024. URL https://jimeng.jianying.com/ai-tool/home.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *NeurIPS*, 2024.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv*, pp. 1–22, 2017.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation. In *ICML*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2024.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *ECCV*, 2024.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. In *ICLR*, 2023.

Joseph B Kruskal. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29 (2):115–129, 1964.

Kuaishou. https://klingai.kuaishou.com, 2024. URL `https://klingai.kuaishou.com`.

Rohit Kundu, Athula Balachandran, and Amit K. Roy-Chowdhury. Truthlens: Explainable deepfake detection for face manipulated and fully synthetic data, 2025a.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K. Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *CVPR*, 2025b.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *KDD*, 2025.

Zongyu Lin, Wei Liu, Chen Chen, Jiasen Lu, Wenze Hu, Tsu-Jui Fu, Jesse Allardice, Zhengfeng Lai, Liangchen Song, Bowen Zhang, Cha Chen, Yiran Fei, Yifan Jiang, Lezhi Li, Yizhou Sun, Kai-Wei Chang, and Yinfei Yang. Stiv: Scalable text and image conditioned video generation, 2024.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pp. 10770–10780, 2024a.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *CVPR*, pp. 10770–10780, 2024b.

Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, 2024c.

Lumalabs. https://lumalabs.ai/dream-machine, 2024. URL `https://lumalabs.ai/dream-machine`.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ2: Latent reconstruction error based method for diffusion-generated image detection. In *CVPR*, pp. 17006–17015, 2024.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Haonan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jiahao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025.

Long Ma, Jiajia Zhang, Hongping Deng, Ningyu Zhang, Qinglang Guo, Haiyang Yu, Yong Liao, and Pengyuan Zhou. Decof: Generated video detection via frame consistency: The first benchmark dataset, 2024a.

Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation, 2024b.

MiniMax. Minimax officially releases the video-01 video generation model , https://hailuoai.com/video, 2024. URL https://www.minimax.io/news/video-01.

mixkit. mixkit. https://mixkit.com/videos/, 2024.

Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *CVPR*, pp. 17395–17405, 2024.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, 2022.

Zhen-Liang Ni, Qiangyu Yan, Tianning Yuan, Mouxiao Huang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A challenging benchmark for detecting ai-generated video, 2024.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *CVPR*, pp. 24480–24489, 2023.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *CVPR*, pp. 27102–27112, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

pexels. pexels. https://www.pexels.com/videos/, 2024.

Pika Labs. https://pika.art/, 2023. URL https://pika.art/.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean

Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *ICLR*, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, 2022.

Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, Feng Cheng, Feilong Zuo Xuejiao Zeng, Ziyan Yang, Fangyuan Kong, Zhiwu Qing, Fei Xiao, Meng Wei, Tuyen Hoang, Siyu Zhang, Peihao Zhu, Qi Zhao, Jiangqiao Yan, Liangke Gui, Sheng Bi, Jiashi Li, Yuxi Ren, Rui Wang, Huixia Li, Xuefeng Xiao, Shu Liu, Feng Ling, Heng Zhang, Houmin Wei, Huafeng Kuang, Jerry Duncan, Junda Zhang, Junru Zheng, Li Sun, Manlin Zhang, Renfei Sun, Xiaobin Zhuang, Xiaojie Li, Xin Xia, Xuyan Chi, Yanghua Peng, Yuping Wang, Yuxuan Wang, Zhongkai Zhao, Zhuo Chen, Zuquan Song, Zhenheng Yang, Jiashi Feng, Jianchao Yang, and Lu Jiang. Seaweed-7b: Cost-effective training of video generation foundation model, 2025.

Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Mostgan-v: Video generation with temporal motion styles. In *CVPR*, pp. 5652–5661, 2023.

Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. Repvideo: Rethinking cross-layer representation for video generation, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2022.

Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *NeurIPS*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *CVPR*, pp. 12105–12114, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, pp. 28130–28139, 2024.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Haotian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie,

Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihan Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-vl technical report, 2025.

Wan Team. Wan: Open and advanced large-scale video generative models, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *CVPR*, pp. 8695–8704, 2020.

Tao Wang, Yushu Zhang, Shuren Qi, Ruoyu Zhao, Zhihua Xia, and Jian Weng. Security and privacy on generative data in aigc: A survey. *ACM Computing Surveys*, 57(4):1–34, 2024.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models. *IJCV*, 2023a.

Yuhan Wang, Liming Jiang, and Chen Change Loy. Styleinv: A temporal style modulated inversion network for unconditional video generation. In *ICCV*, pp. 22851–22861, 2023b.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *ICCV*, pp. 22445–22455, 2023c.

Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, 2024.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation, 2025.

Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. In *ICML*, 2025.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *ICCV*, pp. 22658–22668, 2023.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. In *CVPR*. arXiv, 2024. doi: 10.48550/arXiv.2311.16498. URL `http://arxiv.org/abs/2311.16498`. arXiv:2311.16498 [cs].

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *NeurIPS*, 2023.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *CVPR*, pp. 8984–8994, 2024.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. In *ICML*, 2025.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Tianyi Yang, Zixuan Huang, Junjie Cao, et al. Deepfake network architecture attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 4662–4670, 2022.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models, 2025.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, and Lu Jiang. Magvit: Masked generative video transformer. In *CVPR*, pp. 10459–10469, 2023.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024. URL `https://github.com/hpcaitech/Open-Sora`.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image, 2023.

APPENDIX

This appendix provides a detailed analysis of our dataset, implementation details, additional experimental results, and visualizations:

- Section A: Data distribution and analysis of our dataset.
- Section B: Cross-validation experiment.
- Section C: Additional implementation details for both our method and the baseline methods.
- Section D: Additional experimental results and ablation studies.
- Section E: Visualizations and discussion.

A DATASET COMPOSITION

Model / Video Source	Ver.	Availability	Videos	Resolution	FPS	Frame	Duration
Kinetics-400 Kay et al. (2017)	17.05	real videos	68K	720p	-	-	5-10s
MSVD Chen & Dolan (2011)	11.06	real videos	1970K	240-1080p	6-60		1-60s
Overall Real	-	-	70,543	240-1080p	6-60	-	1-60s
RepVideo Si et al. (2025)	25.01	open-source	4720	720x480	8	49	6.1s
Wan2.1 Team (2025)	25.01	open-source	4725	1280x720	16	81	5.0s
CausVid (5s) Yin et al. (2025)	25.01	open-source	4720	640x352	24	120	5.0s
Apple-STIV Lin et al. (2024)	24.12	open-report	4715	512x512	60	60	1.0s
Sora Brooks et al. (2024)	24.12	private	4720	854x480	30	150	5.0s
HunyuanVideo Kong et al. (2024)	24.12	open-source	4725	1280x720	24	129	5.4s
Gen-3 Germanidis (2024)	24.06	private	4707	1280x768	24	256	10.7s
Luma Lumalabs (2024)	24.06	private	4680	1360x752	24	121	5.0s
Kling Kuaishou (2024)	24.06	private	4679	1280x720	30	153	5.1s
Jimeng Jimeng AI (2024)	24.05	private	6214	1280x720	8	96	12.0s
OpenSora V1.1 Zheng et al. (2024)	24.04	open-source	4720	424x240	8	64	8.0s
Mira Ju et al. (2024)	24.04	open-source	4721	384x240	6	60	10.0s
VideoCrafter-2.0 Chen et al. (2024a)	24.01	open-source	4720	320x512	10	16	1.6s
Pika 1.0 Pika Labs (2023)	23.11	private	4715	1280x720	24	72	3.0s
AnimateDiff-V2 Guo et al. (2024)	23.09	open-source	4715	512x512	8	16	2.0s
Overall Fake	-	-	70,692	240-720p	6-60	16-256	1-12s

Table 8: **Statistics of real and synthetic videos in the proposed training set.**

Model / Video	Split	Videos	Resolution	FPS	Frame	Duration
Movie Gen Polyak et al. (2024)	validation (fake)	1003	1920x1088	24	256	10.7s
Panda-70M Chen et al. (2024d)	validation (real)	1000	720p	6-30	-	10-50s
Mixkit mixkit (2024)	test (real)	215	720p	15-60	-	10-17s
Pexels pexels (2024)		292	720p	24-60	-	6-39s
Wan2.1 Team (2025)	test (fake)	215	1280x720	30	161	5.4s
Wan-1.3B Team (2025)		292	832x480	16	81	5.0s
Hailuo MiniMax (2024)		215	1280x720	25	141	5.6s
Seaweed Seawead et al. (2025)		215	1472x832	24	121	5.0s
Seedance Gao et al. (2025)		215	1248x704	24	121	5.0s
StepVideo Ma et al. (2025)		215	960×540	25	204	8.2s

Table 9: **Statistics of real and synthetic videos in the proposed validation and Magic Videos Testset.**

A.1 TRAINING SET.

Table 8 provides a comprehensive summary of the training dataset used in our work. Previous research has emphasized the critical importance of dataset quality and diversity in training robust detectors Rajan et al. (2025), especially given the variety of artifacts produced by different generative

models Wu et al. (2025). To advance the field of AI-generated video detection, we curated a large-scale dataset comprising outputs from 15 distinct video generation models. The majority of these synthetic videos are sourced from VBench Huang et al. (2023), a benchmark selected for its high-quality prompt library and extensive evaluation of state-of-the-art models. This choice allowed us to avoid the costly and time-consuming processes of large-scale video filtering, quality control, and generation while ensuring high quality and consistency of generated video data.

Our dataset reflects the diverse and evolving landscape of video generation, featuring models developed between 2023 and 2025. It includes a wide range of model types in terms of availability (i.e., open-source, open-report, and private) and architecture (e.g., Diffusion U-Net, DiT-based, auto-regressive models, and others with undisclosed architectures). The models differ significantly in training methodology, data scale, output resolution, and video duration, contributing to a richly diverse training set.

To complement the synthetic videos, we sampled an equal number of real videos from two authentic sources, MSVD Chen & Dolan (2011) and Kinetics-400 Kay et al. (2017). These were carefully selected to match the resolution, duration, and encoder distribution of the generated videos. This matching is essential for reducing potential biases and ensuring that the learned features are genuinely discriminative between real and fake content.

A key feature of our dataset is that all generative models were conditioned on the same prompt library, ensuring a shared semantic distribution across the generated videos. This unique setup enables controlled cross-validation experiments, allowing us to investigate inter-model relationships and identify key factors that influence detector performance, as discussed in Section B.

A.2 Validation and Test Set

Table 9 presents the composition of our validation set and introduces a novel, high-quality test set, which we name the Magic Videos Testset.

Validation Set. Rather than adopting the common practice of partitioning a subset of the training data, we constructed the validation set from videos generated by Movie Gen Polyak et al. (2024), a model that is architecturally and semantically similar but not identical to the models used in training. These synthetic videos are paired with 1,000 of real videos sampled from the Panda-70M Chen et al. (2024d) dataset. During training, we apply early stopping based on the validation loss computed on this set. This strategy helps mitigate overfitting to the specific models and scenarios encountered during training, promoting the selection of a model checkpoint with stronger generalization capabilities.

Test Set. We identified a critical gap in existing benchmarks: they often lack coverage of the latest generative models and may exhibit evaluation biases. To address this, we constructed the Magic Videos Testset using a high-quality video generation pipeline, as introduced in Section 3 of the main paper. This test set includes real videos from two premium platforms—Mixkitmixkit (2024) and Pexelspexels (2024)—covering a diverse range of common scenes such as landscapes, architecture, human subjects, and news footage. These videos are provided at resolutions up to 1080p to ensure both high fidelity and content diversity. For evaluation, real and generated videos are matched into balanced subsets, allowing for the computation of accuracy and other performance metrics.

To generate the synthetic counterparts, we first applied ShareGPT4Video Chen et al. (2024c) to produce high-quality captions for the real videos. These captions were then refined through a rigorous process of filtering, rewriting, and final prompt polishing. The resulting prompts were input to six advanced text-to-video models, comprising both open-source and commercial systems. Below, we detail the generative models used to construct the Magic Videos Testset:

- Wan2.1 Team (2025): We used the Wanxiang platform API with the ”professional” model, default settings, and prompt optimization disabled. Prompts were derived from the Mixkit collection. This model may apply post-processing, resulting in a higher frame rate than Wan-14B.
- Wan-1.3B Team (2025): Videos were generated using the official open-source implementation and pre-trained model, with prompts from the Pexels collection.

- Hailuo MiniMax (2024): Accessed via the MiniMax-T2V-01 commercial API, this model was configured to generate 5-second videos using prompts from the Mixkit collection. Prompt optimization was not applied.
- SeaweedSeawead et al. (2025): As official model weights are not publicly available, we used the commercial model Jimeng-S2.0Jimeng AI (2024), which is based on the Seaweed-alpha model. Prompts were sourced from the Mixkit collection. Generation was performed using prompts from the Mixkit collection.
- SeedanceGao et al. (2025): In place of unavailable official weights, we used the commercial model Jimeng-S3.0Jimeng AI (2024), corresponding to the Seedance 1.0 Mini model. Prompts were sourced from the Mixkit collection.
- StepVideo Ma et al. (2025): Videos were generated using the official API with the Step-Video-T2V endpoint (544px × 992px × 204f), using prompts from the Mixkit collection.

B Cross-Validation Experiment

B.1 Experiment Setup

Cross-Validation Setup. This experiment focuses on in-domain, cross-model validation of detectors. The benchmark utilizes data generated by 15 models from VBench Huang et al. (2023), which evaluates various generative models using a shared set of predefined prompts. Because all models generate videos from the same prompt library, we consider their outputs to belong to the same semantic domain. Let F_i denote the subset of videos generated by model i, and let R_0 represent a fixed set of real videos, sampled to contain the same number of examples as each F_i. For each model i, we train a deepfake detector on the dataset F_i, R_0 and evaluate its performance on all other generated subsets F_j (for $j \neq i$). This setup allows us to rigorously assess the generalization ability of detectors across different generative architectures while keeping the semantic domain fixed. It also provides a controlled environment for analyzing the relationships between generative model architectures and detection performance. This Cross-Validation Benchmark produces an $n \times n$ matrix $\mathbf{M}$, where $\mathbf{M}[i, j]$ represents the recall of a detection model trained on subset i and evaluated on subset j. Based on preliminary observations, we propose the following two hypotheses, which will be validated in subsequent experiments.

Similarity Between Generative Models. The matrix entry $M[i, j]$ reflects the output similarity between generative models i and j, influenced by factors such as model architecture, sampling strategies, and training data. We observe that models with more similar architectures tend to exhibit higher cross-validation accuracy between them. To quantify this relationship, we define a non-directional distance metric, $d(i, j) = 1 - 0.5 \times (M[i, j] + M[j, i])$. Using this metric, we apply Non-metric Multidimensional Scaling (MDS) Kruskal (1964) to produce a 2D spatial representation of the generative models. This visualization aids in understanding the architectural relationships and clustering patterns among the models, offering insights into how architectural similarity correlates with cross-detection performance.

Impact of Generation Quality. In addition to architecture, $M[i, j]$ is also influenced by the generation quality of model i. We hypothesize that higher-quality synthetic videos provide more realistic and informative supervision signals, enabling the classifier to learn more effective forgery-discriminative features. Since ground-truth quality labels are unavailable, we adopt scores from recent T2V benchmarks Huang et al. (2023); Liu et al. (2024c); Huang et al. (2024) as a proxy for generation quality. To assess the relationship between generation quality and detection effectiveness, we compute Pearson correlation coefficients (ρ) between the benchmark quality scores and corresponding detection accuracies.

B.2 Cross-Validation Results

Cross Validation. As discussed above, we use the cross-validation matrix $\mathbf{M}$ to evaluate the similarity between generative models. Four detection models—F3NetQian et al. (2020), X-CLIP-B/32Ni et al. (2022), TALLXu et al. (2023), and NPRTan et al. (2024)—are trained on 5K real videos from MSR-VTT and 5K generated videos from each specific model subset. These detectors are then

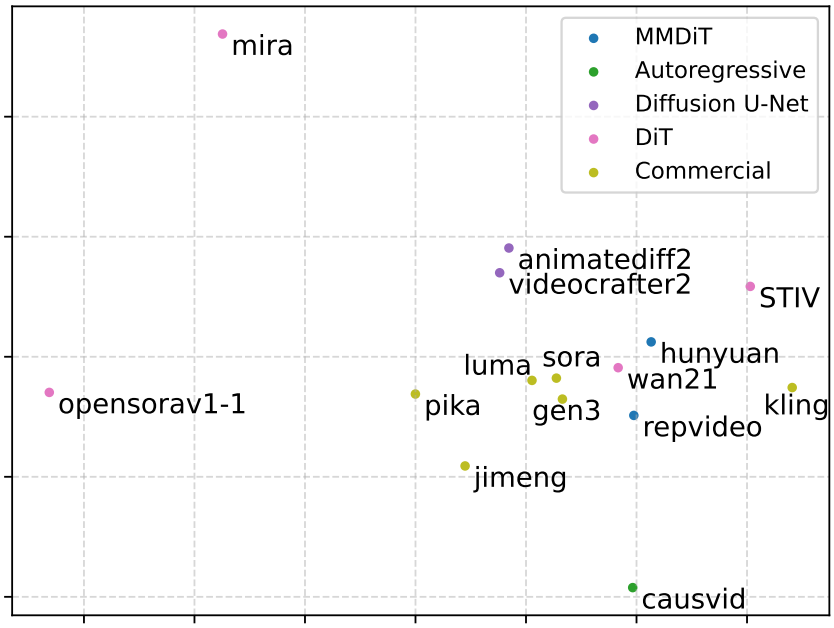


Figure 5: **MDS Kruskal (1964) Visualization of video generation models**. Model similarity is based on pairwise detection accuracy.

Model	Wan21	Hun-yuan	Kling	Sora	Gen-3	Rep-Video	Jimeng	Luma	Mira	Pika	Open Sora	STIV	Caus Vid	VCraf ter-2	ADiff -V2	AVG
Wan21	**99.7**	97.4	93.1	97.1	98.4	88.7	96.8	97.5	33.8	91.3	44.7	88.3	97.7	92.1	94.2	87.38
Hunyuan	93.9	**99.4**	81.3	92.9	92.0	69.7	89.7	91.4	38.5	84.4	34.2	74.2	89.8	86.3	90.6	80.55
Kling	85.9	82.0	**98.5**	82.2	84.4	66.3	76.7	84.8	26.2	65.9	35.2	69.1	92.2	71.0	76.8	73.15
Sora	90.3	84.9	75.6	**99.5**	93.7	78.7	99.2	96.1	35.2	87.1	34.1	70.7	98.0	88.3	90.4	81.44
Gen-3	92.6	84.9	81.7	93.0	**99.6**	83.8	94.4	95.3	23.7	88.1	50.6	75.1	96.5	80.4	84.5	81.60
RepVideo	97.0	91.5	92.8	96.9	98.3	**98.8**	98.6	97.8	29.4	94.8	56.6	86.9	99.8	91.7	91.1	**88.13**
Jimeng*	61.1	47.9	39.9	78.4	81.6	54.5	**99.7**	81.3	14.8	70.9	26.4	39.6	91.7	55.0	60.8	60.23
Luma	90.8	84.3	80.2	95.7	94.2	79.1	98.4	**98.6**	41.4	89.4	47.7	72.5	99.3	88.0	92.0	83.44
Mira	20.6	33.6	21.4	33.3	24.2	12.4	32.6	42.0	**98.7**	25.3	24.4	16.6	29.5	59.7	79.7	36.93
Pika	77.2	64.3	59.1	81.0	87.1	68.2	91.1	82.6	21.9	**97.2**	43.9	65.6	80.3	83.9	72.1	71.69
Opensora V1.1	58.1	55.2	57.4	64.1	79.1	49.3	70.3	73.7	50.4	73.8	**90.4**	48.2	55.0	65.2	64.1	63.62
Apple-STIV	87.3	74.8	77.5	74.4	88.7	78.6	73.3	74.9	29.7	78.0	36.4	**96.3**	67.6	91.1	86.2	74.31
CausVid	28.7	21.8	19.6	39.6	34.5	38.3	50.7	44.8	7.3	16.0	7.4	16.7	**99.6**	18.4	28.8	31.47
VideoCrafter-2	76.8	67.8	54.2	79.7	74.3	60.4	81.8	77.6	64.1	78.8	28.9	70.4	77.0	**99.2**	94.6	72.39
AnimateDiff-V2	68.4	60.5	49.4	75.6	67.7	51.1	83.7	73.9	60.4	58.1	20.0	58.9	76.2	89.4	**99.1**	66.16

Table 10: **Cross-Validation Results**. Each cell in the table represents the average recall (%) of four detection models (NPR Tan et al. (2024), TALL Xu et al. (2023), X-CLIP-B/32 Ni et al. (2022), F3Net Qian et al. (2020)). The model is trained on generated videos of each subset and 5k real videos from MSR-VTT dataset.

tested on all other generative subsets. The average cross-validation accuracy across the four detectors is reported in Table 10. Each element in the table represents the mean detection accuracy across the four models. Diagonal entries correspond to in-subset evaluations, where the detector is tested on the same generative model used for training As shown in Figure 5, we interpret the matrix $\mathbf{M}$ as a distance metric between generative models and apply Multidimensional Scaling (MDS) to project their relationships into a 2D space. This visualization reveals clusters of architecturally similar models, such as AnimateDiff2Xu et al. (2024) and VideoCrafterV2Chen et al. (2024a), while autoregressive-based models, such as Yin et al. (2025), appear more distant from the rest. This mapping also informs a diverse training set selection of generative models, we could combine the cross validation accuracy and similarity to construct a high-quality and diverse dataset for data-efficient training.

Better Generation, Better Detection. In our cross-validation experiment, we observed that detection models trained on higher-quality generated videos exhibit stronger detection performance. To validate this observation, we retrieved the overall VBench scores Huang et al. (2023) for each

generative model and conducted a correlation analysis between these scores and the average detection accuracies reported in Table 10. The results are visualized in Fig.1 of our main paper. Since the cross-validation data is directly sampled from VBench's evaluation set, the VBench scores provide an accurate proxy for the generation quality of each subset. Across 14 models (excluding CausVid, which features a fundamentally different model structure and training paradigm), we compute a Pearson correlation coefficient of $\rho = 0.86$ between average detection accuracy and VBench scores, indicating a strong positive correlation. Furthermore, when restricting the analysis to the six DiT-based models, the correlation increases to $\rho = 0.92$. These results strongly support our hypothesis: among models with similar architectures, higher-quality generation leads to better supervision signals, enabling detection models to learn more effective forgery-discriminative features.

C Implementation Details

This section outlines the configurations and hyper-parameters used for training our proposed method, as well as the baseline models.

Our Method. For our detector and Moon-ViT, all experiments are conducted using PyTorch with Automatic Mixed Precision (AMP) in bfloat16 to enable Flash Attention optimization and accelerate training. The visual backbone is initialized with Vision Transformer (ViT) weights from the officially released Qwen2.5-VL model. We explore multiple fine-tuning strategies with distinct hyperparameter settings: (1) Full fine-tuning: We set the batch size to 4 and train for 3 epochs with a learning rate of 1e-5. (2) Linear Probing (LP) and Parameter-Efficient Fine-Tuning (PEFT): These approaches use a larger batch size of 32 and a learning rate of 1e-4. Training continues for up to 30 epochs, with early stopping based on validation loss (patience = 5 epochs) to prevent overfitting.

Other Baseline Methods. To ensure fair comparison, all baseline models are trained under a unified experimental setup. We used a consistent batch size of 32 and trained for a maximum of 30 epochs, also employing an early stopping strategy with a 5 epochs patience. The learning rate was adjusted based on the model architecture: for baselines utilizing a CLIP ViT backbone, such as X-CLIP and CLIP-based detectors, we set the learning rate to 1e-6; for all other models, a learning rate of 1e-5 was used.

Data Pre-processing for Baseline Methods. A consistent data pre-processing pipeline is applied across all models during both training and testing. During training, each video is first sampled at a rate of 2 frames per second, from which 8 consecutive frames are extracted. If a video contains fewer than 8 frames, it is padded with blank frames to meet the required sequence length. Each frame is resized such that the shorter side is 224 pixels, followed by a random crop to a final resolution of 224×224. To enhance model robustness, we apply two forms of data augmentation: random horizontal flipping and random Gaussian noise. During testing, frames are sampled in the same manner as during training. After resizing the shorter side of each frame to 224 pixels, a center crop to 224×224 is applied instead of a random crop to ensure deterministic evaluation.

D Additional Results and Ablations

Full Results on Genvideo-Val. As shown in Table 4, our proposed method achieves state-of-the-art performance across several key metrics. Notably, it attains an F1 score of 90.64 and an average precision (AP) of 96.13, surpassing all other leading methods—including DeMamba-CLIP, which was trained on the GenVideo dataset comprising 2.2 million samples. In contrast, our model was trained on only 140K samples, over ten times fewer, underscoring both the high quality of our training data and the efficiency of our method in learning robust forgery-discriminative features at native resolution.

In addition, our model achieves a balanced accuracy (bACC) of 95.38, significantly outperforming all competing methods. This result demonstrates not only high overall detection performance but also the model's well-rounded and consistent capabilities across diverse forgery cases.

Efficiency Comparison. As detailed in Table 10, we conduct a comprehensive efficiency analysis comparing our proposed Qwen2.5-VL ViT (QwenViT) with several strong baseline models. For a

Model	Training Data	Metric	Sora	Morph Studio	Gen2	HotShot	Lavie	Show-1	Moon Valley	Crafter	Model Scope	Wild Scrape	Avg.
UNITE	FaceForensics++, SAIL-VOS-3D	Recall	92.11	100.0	94.62	96.93	98.12	99.86	98.69	100.0	96.29	89.89	89.60
		F1	-										
		AP	88.57	100.0	100.0	90.16	89.91	98.34	99.52	100.0	98.96	92.56	92.76
DeMamba-CLIP	GenVideo	Recall	95.71	100.0	98.70	69.14	92.43	93.29	100.0	100.0	83.57	82.94	**91.58**
		F1	64.63	96.15	97.39	78.03	94.14	92.76	95.72	98.04	87.23	87.82	89.19
		AP	85.50	100.0	99.59	76.15	96.78	96.99	99.97	100.0	89.80	89.72	93.45
RINE	ProGAN	bACC	-	84.00	89.10	66.00	96.70	91.80	85.70	98.30	76.60	-	74.10*
DeMamba	PyramidFlow	bACC	-	83.80	92.20	62.00	79.60	72.60	92.40	87.50	68.60	-	78.10*
Corvi et al.	PyramidFlow	bACC	-	97.00	98.80	81.40	95.50	92.10	98.40	98.30	97.10	-	94.30*
Ours	15model-140k	Recall	82.14	97.14	99.49	89.00	98.79	92.29	99.05	99.07	83.00	71.60	91.16
		F1	65.25	95.84	98.35	91.48	98.02	93.29	96.51	98.16	88.03	81.45	**90.64**
		AP	82.49	99.36	99.95	96.55	99.78	97.88	99.87	99.89	94.50	90.98	**96.13**
		bACC	90.87	98.38	99.55	94.31	99.20	95.95	99.33	99.34	91.31	85.61	**95.38**

Table 11: **Benchmarking Evaluation in terms of Recall, F1 score (F1), average precision (AP), and balance accuracy (bACC) on Genvideo-Val.** The results of RINE and DeMamba are reported in Corvi et al. (2025).

Model	Resolution	#Params	FLOPS	Peak GPU Mem	Training Time / Epoch
CLIP-L	[224, 224]	303.2M	622.6G	21.5GB (bs=4) 129.3GB (bs=32)	9.5 A100 hours
X-CLIP-L	[224, 224]	429.2M	650.6G	21.5GB (bs=4) 129.3GB (bs=32)	10.5 A100 hours
Effort	[224, 224]	0.2M/504.6M	623.4G	17.3G (bs=4) 75.1GB(bs=32)	7.5 A100 hours
QwenViT	[224, 224]	668.7M	656G	16.0GB(bs=4)	2.3 A100 hours
QwenViT	dynamic [224p, 448p]	668.7M	-	37.9GB(bs=4)	7 A100 hours
QwenViT-LoRA	dynamic [224p, 448p]	2.6M/671.31M	-	27.4GB(bs=4)	5.5 A100 hours

Table 12: Efficiency comparison results on model parameters, FLOPS, GPU memory utilization and time consumed during training.

standard input resolution of [224, 224], QwenViT exhibits remarkable training efficiency. Despite having more parameters (668.7M) than CLIP-L (303.2M), it achieves a 4.1× reduction in training time (2.3 vs. 9.5 A100 hours) and a 25% decrease in peak GPU memory usage (16.0GB vs. 21.5GB at a batch size of 4). These gains are primarily attributed to efficiency-oriented design choices such as bfloat16 training and Flash Attention, which allow QwenViT to utilize computational resources more effectively. When adopting a dynamic resolution strategy, the training overhead naturally increases, yet QwenViT remains faster and more memory-efficient than the baselines. Moreover, our parameter-efficient fine-tuning variant, QwenViT-LoRA, requires updating only 2.6M parameters. This substantially reduces resource demands compared to full dynamic fine-tuning, lowering GPU memory from 37.9GB to 27.4GB and cutting training time from 7 to 5.5 hours. Overall, these results highlight that the superior efficiency of QwenViT stems from architectural optimizations, making the additional cost of higher dynamic resolutions acceptable in practice.

E Visualization and Discussion

Figures 6 to 9 present a selection of video samples from our dataset, with Figures 3–5 offering detailed visualizations along with their corresponding generative prompts. As illustrated in these figures, the videos in our test set exhibit high visual quality, characterized by aesthetic appeal, rich motion, and diverse themes and visual effects.

By using carefully curated prompts to control the generative themes, we are able to evaluate a model's detection performance without introducing content bias. This methodological design promotes a fairer and more reliable assessment, encouraging the detector to learn generalizable forgery artifacts rather than memorizing specific object- or scene-level patterns.

Acknowledgment of LLM Usage. This manuscript has benefited from the assistance of a large language model, which was employed solely for grammar checking and language polishing. All scientific ideas, experimental designs, analyses, and conclusions are made by the authors.

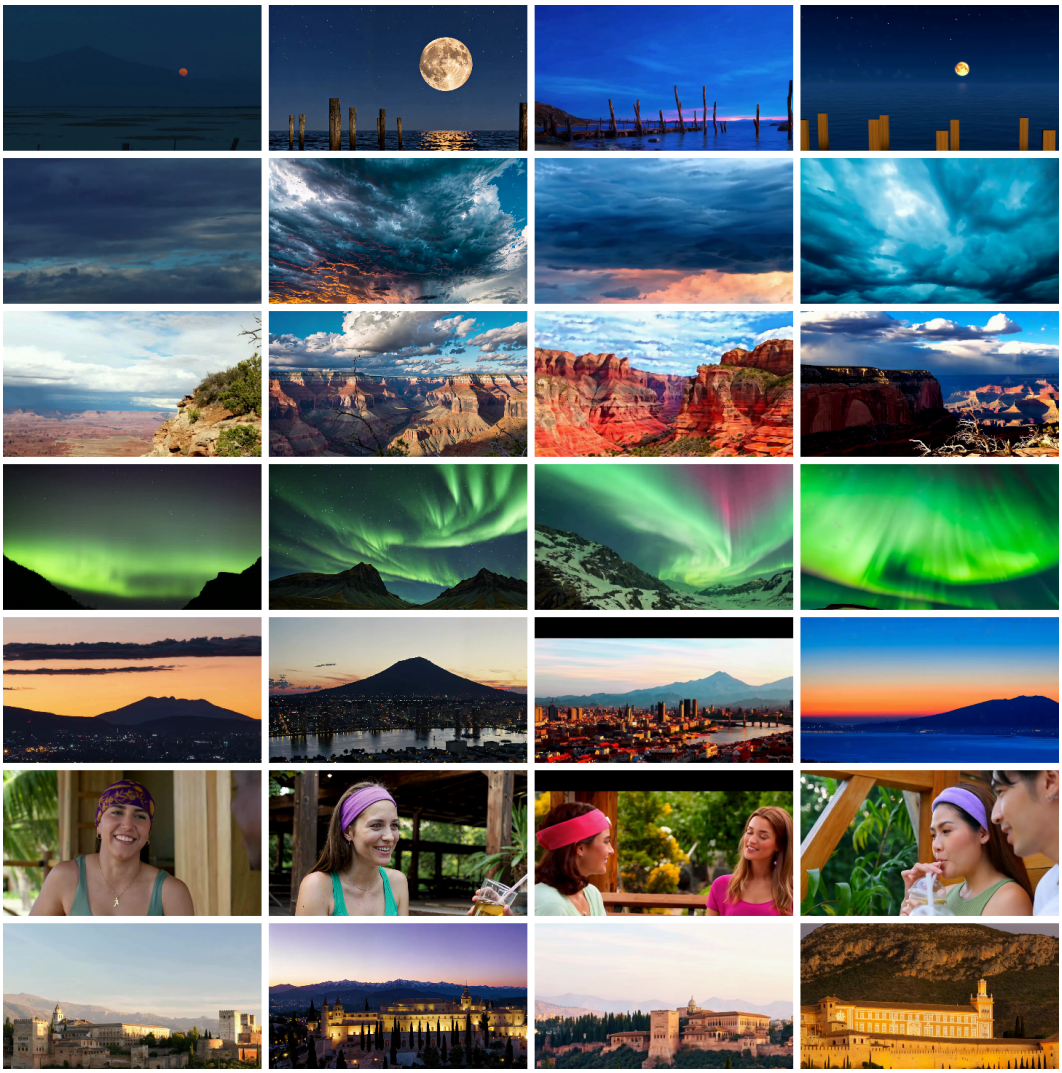
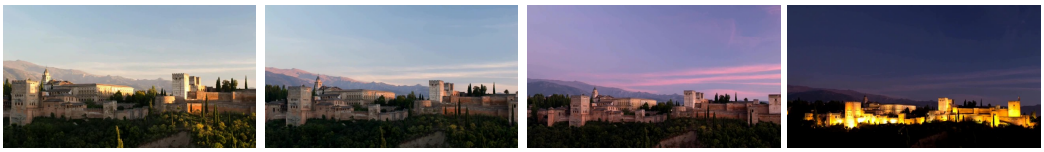


Figure 6: **Video Visualization from Magic Video Testset.** From left to right, each column denotes videos from real sources, seaweed, seedance, and wan2.1.

Limitations. First, despite our best efforts, the field of generative AI is advancing at an exceptional pace. As new generator architectures continue to emerge, the dataset and detection framework will require periodic updates to maintain relevance. Second, although optimized for efficiency, processing videos at their native resolution remains more computationally intensive than traditional methods based on downscaled inputs. This may limit deployment in resource-constrained environments. Finally, further investigation into the model's explainability could yield valuable insights into the specific artifacts it learns to detect, thereby advancing the understanding of generative model fingerprints.

real video



prompt: The video showcases the Alhambra in Granada, Spain, transitioning from warm golden sunset tones to deep violet hues as night falls. The palatial structures, set against the Sierra Nevada mountains and lined with cypress trees, shift from sunlit brilliance to dramatic nighttime illumination. A subtle zoom enhances the view, while the changing light casts a striking contrast between the fortress's golden glow and the darkening sky, creating a captivating visual transformation.

generated video

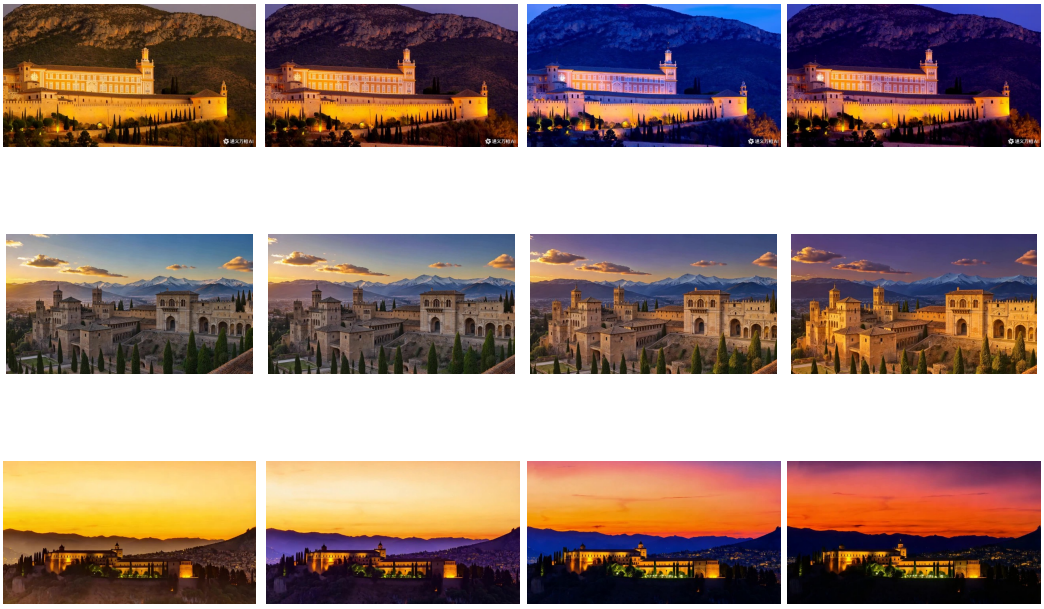


Figure 7: **Video Visualization from Magic Video Testset.**

real video

prompt: A woman and a man engage in a friendly outdoor conversation amid wooden structures and greenery. The woman, wearing a purple headband and green tank top, sips her drink, signaling relaxation. Her expressions shift from savoring to engaging warmly, smiling and making eye contact. The man listens attentively, maintaining a steady demeanor. Both hold beverages, emphasizing the leisurely tone. Their uninterrupted dialogue features moments of humor and enjoyment in a serene setting.

generated video

Figure 8: **Video Visualization from Magic Video Testset.**

real video

prompt: The video showcases billboards for Powerball ($470M) and Mega Millions ($999M) under a sunny sky, with a '3 News Now' banner highlighting a '$1 BILLION MEGA MILLIONS JACKPOT.' Vibrant designs and mentions of 'NEBRASKA POWERBALL POWERPLAY' add local context. A brief error misstates the Mega Millions jackpot as $9M before correcting it. The video ends with a wide shot of the billboards against a residential backdrop, emphasizing their public appeal.

generated video

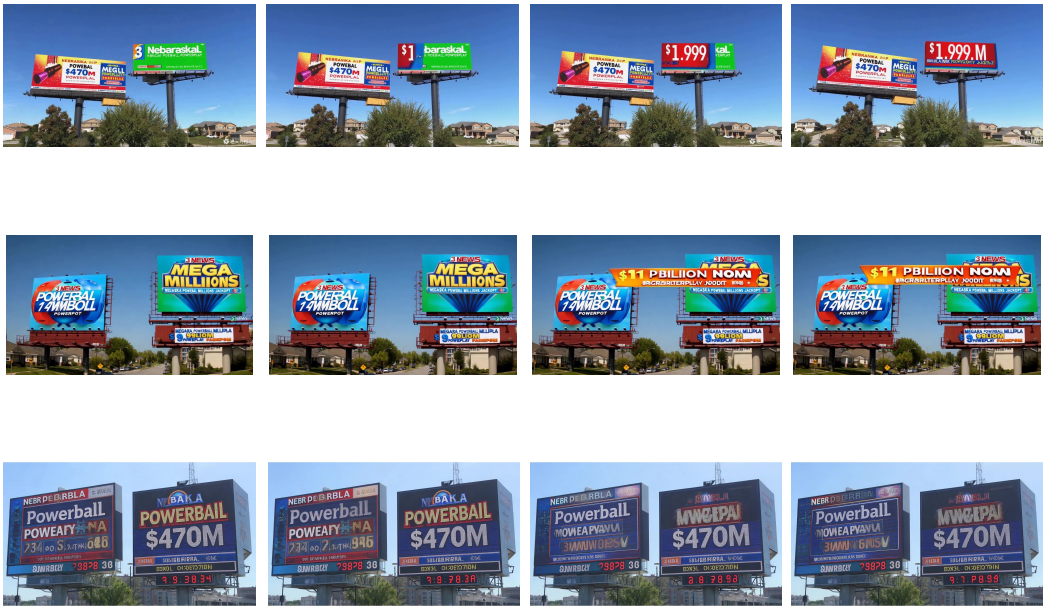


Figure 9: **Video Visualization from Magic Video Testset.**