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

ICCV 2025

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao,

Le Yang, Shuai Liu, Minghui Yang, Cong Wang, Chao Shen

AI-Video Detection: Background

➢ **AI-generated Videos: from GANs to Diffusions**

- More diverse scenes
- Advanced authenticity
- Unseen generators

V.S.

- Unclear detection principle
- Biased datasets
- Limited computing resources

2017

DEEPFACELAB

sora

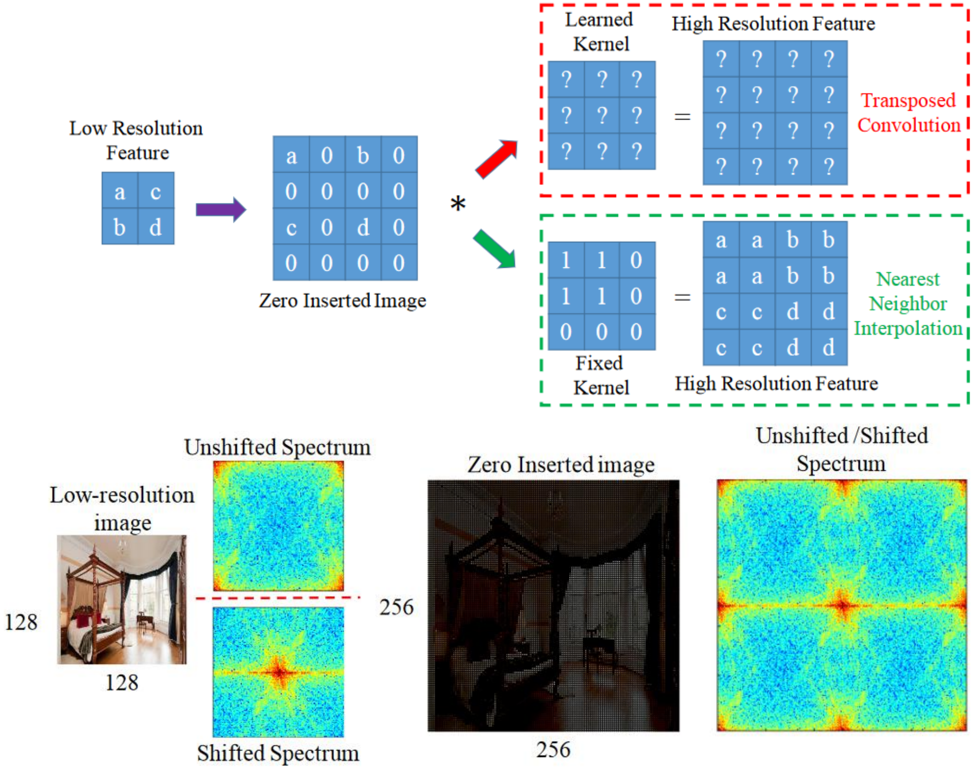
2025

AI-Video Detection: Motivation

➢ Existing Limitations - Temporal Artifact Analysis Gap

- **Low-level Artifacts** (*Pixel domain modeling*)
  - Example: Up-sampling artifacts
- **Statistical Attribution** (*Spectra domain modeling*)
  - Example: Spectral artifacts analysis

How about temporal artifacts?



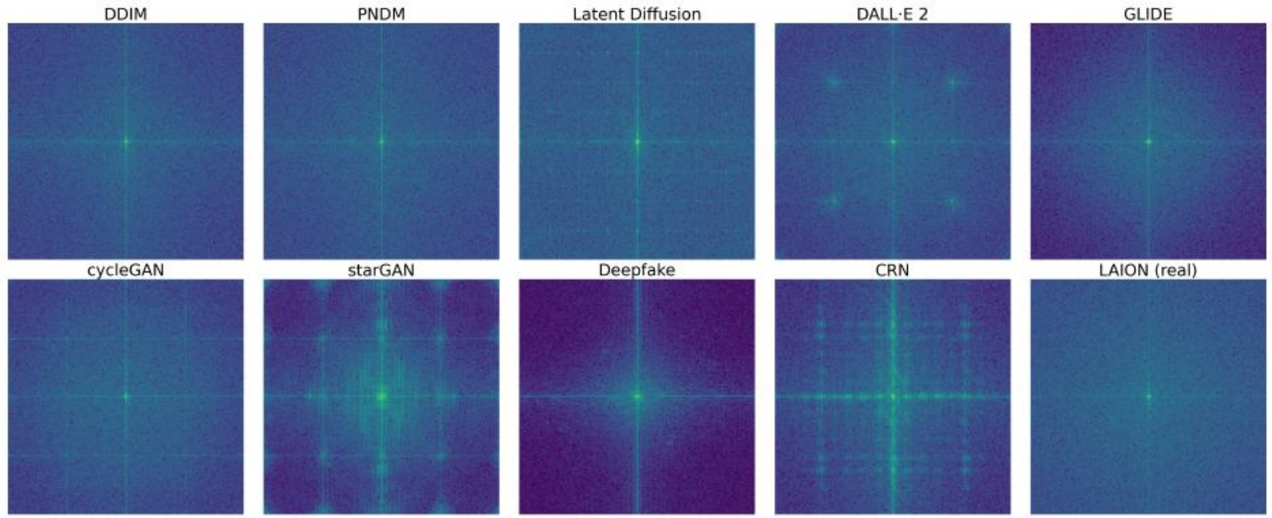


Figure 1: **Generator artifacts:** noise residuals power spectrum of images from 9 generative models and 1 real dataset. Top row: 5 Diffusion Models. Bottom row: 2 GANs, cycleGAN and starGAN, 2 CNN-based generators, Deepfake and CRN, and 1 real dataset, LAION.

Image Sources - Detecting and Simulating Artifacts in GAN Fake Images - Breaking Semantic Artifacts for Generalized AI-generated Image Detection

AI-Video Detection: Analysis

➢ **Temporal artifacts based on Newtonian mechanics**

Second-order system (Newtonian mechanics)

$$A_2 \frac{d^2x(t)}{dt^2} + A_1 \frac{dx(t)}{dt} + A_0 x(t) = u(t)$$

Second-order Central Difference **to approximate the acceleration**

$$f''(x) = \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = \frac{f'(x) - f'(x-h)}{h}$$

2nd-order flow

$$X_{diff} = \frac{OF(x_{t+1}, x_{t+2}) - OF(x_t, x_{t+1})}{\Delta t^2}$$

2nd-order semantics

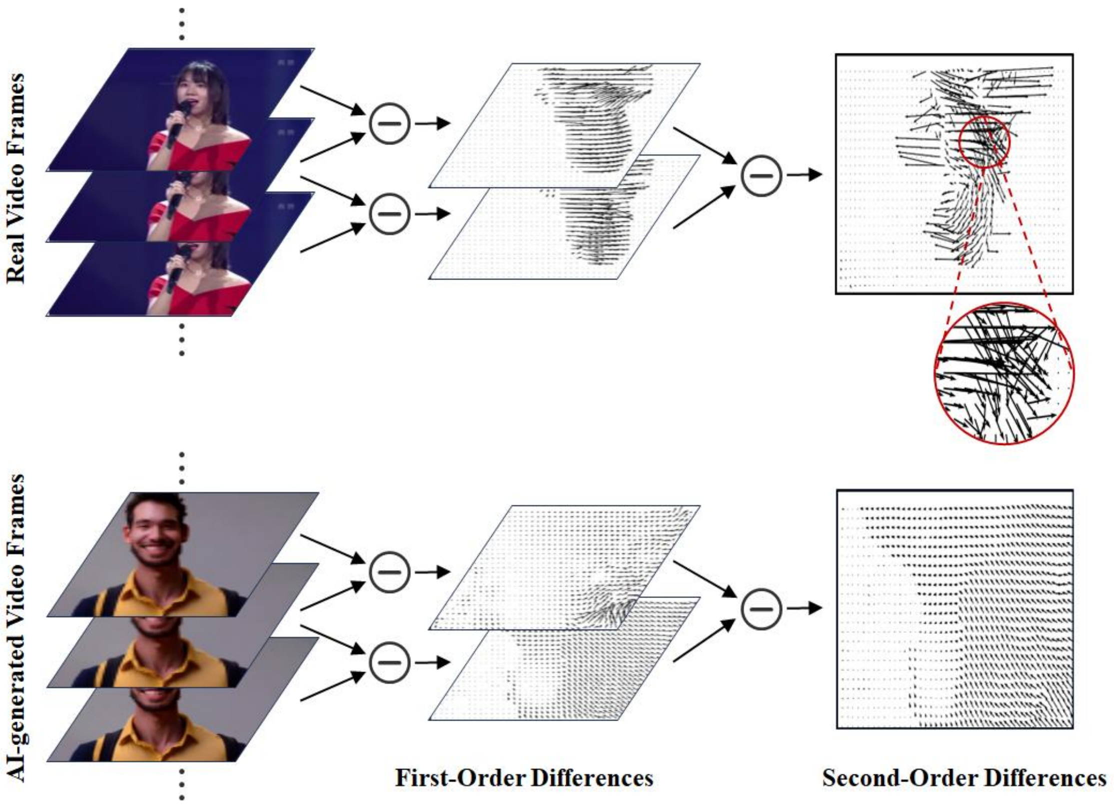
$$F_2(k) = \frac{F_1(k) - F_1(k-1)}{\Delta t}, \quad k = 2, ..., T-1$$

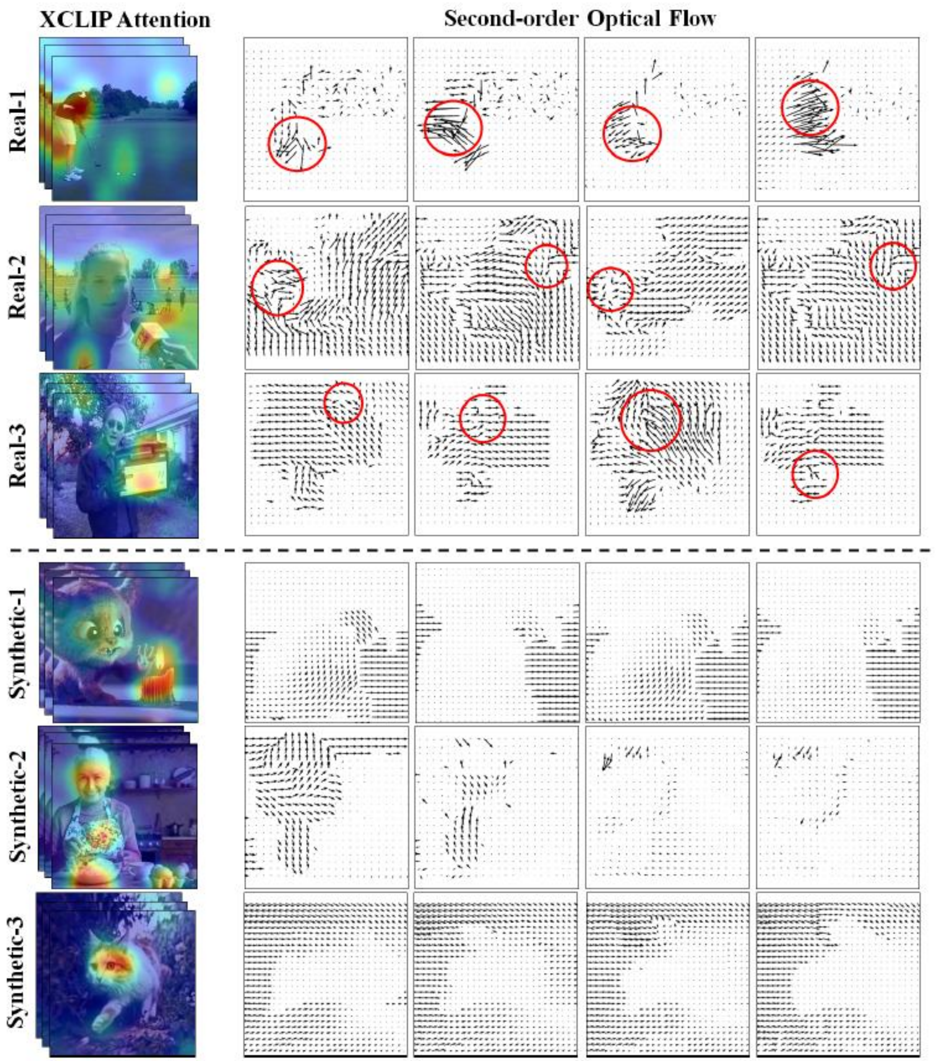
- Synthetic videos exhibit **hyper-smooth transitions** violating Newtonian dynamics (e.g., unnatural acceleration patterns)
- Realistic scenarios can be simulated using **second-order systems**.

AI-Video Detection: Analysis

2nd-order Flow $X_{diff} = \dfrac{OF(x_{t+1}, x_{t+2}) - OF(x_t, x_{t+1})}{\Delta t^2}$

- **Real: Perturbations in local regions**
- **Generated: Smoother regions overall**



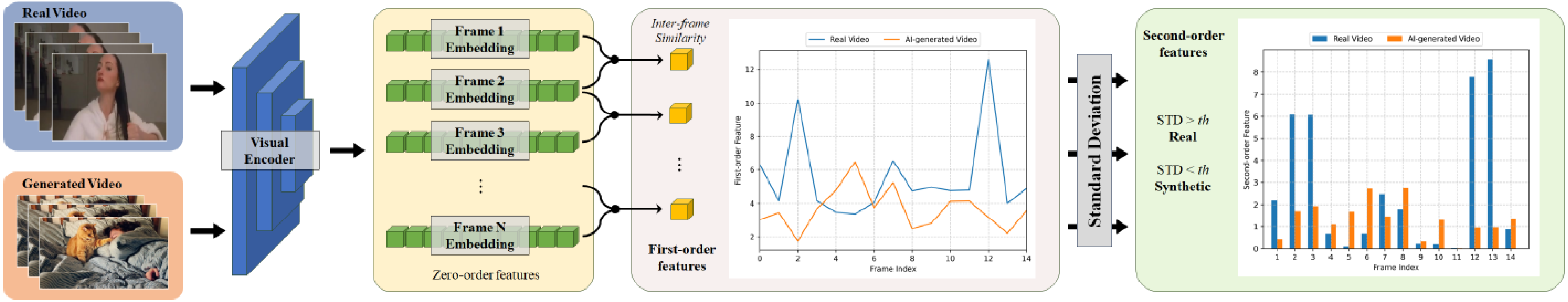


AI-Video Detection: Method

- Real videos contain **high-order** features
- AI videos contains **unusual high-order** features

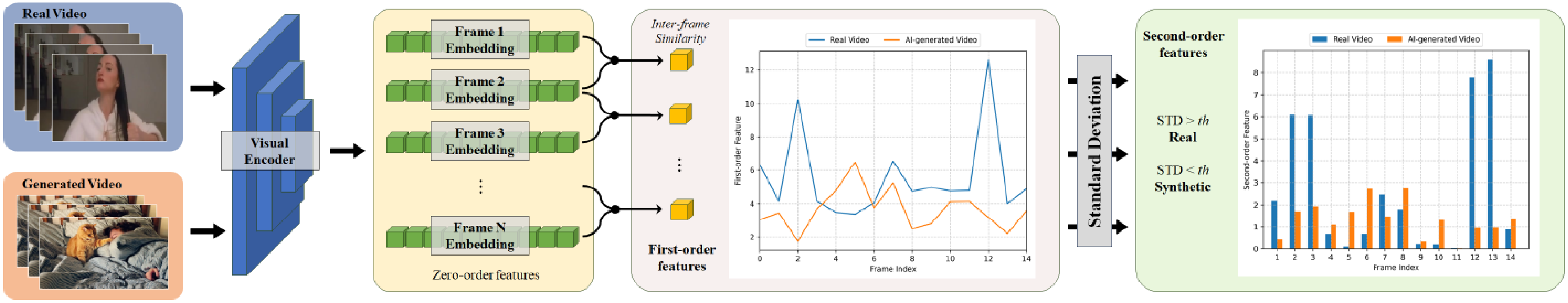
⇒

- Using **pretrained visual encoders** to extract features by frames.
- Using **second-order differential feature** to realize general detection.



(a) Zero-order feature extraction (b) First-order feature extraction (c) Second-order feature extraction

AI-Video Detection: Method



(a) Zero-order feature extraction (b) First-order feature extraction (c) Second-order feature extraction

➢ Zero-order Feature

use visual encoders to extract features by frames

$$F^0 = \{F_1^0, \ldots, F_T^0\}$$

➢ First-order Feature

Calculate inter-frame differences via *L2 Distance*

$$F_1^L(k) = \frac{dis(F_k^0, F_{k+1}^0)}{\Delta t}$$

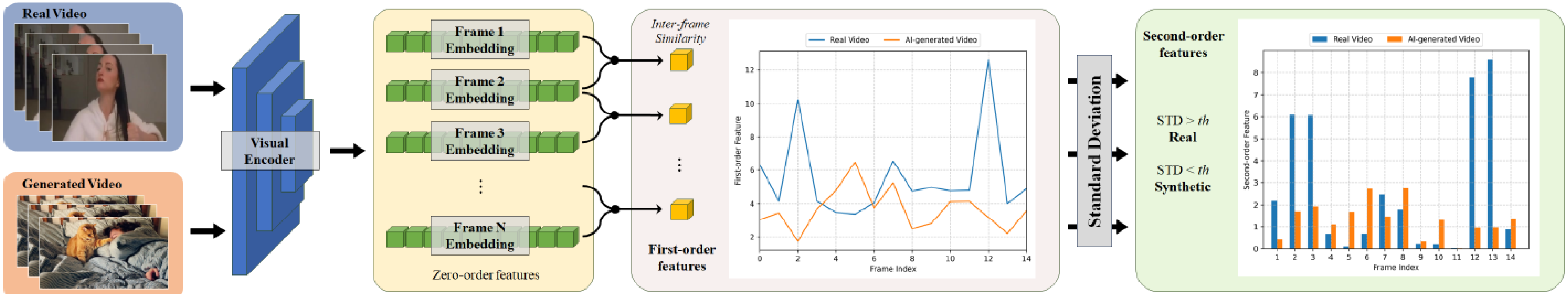
➢ Second-order Feature

Detection metric: Standard deviation of second-order features

$$F_2(k) = \frac{F_1(k) - F_1(k-1)}{\Delta t}$$

$$\sigma(F_2) = \sqrt{\frac{1}{T-3}\sum_{i=2}^{T-1}(F_2(i) - \mu)^2}$$

AI-Video Detection: Method



(a) Zero-order feature extraction (b) First-order feature extraction (c) Second-order feature extraction

- **Use visual encoders to extract features by frames**

$$F^0 = \{F_1^0, \ldots, F_T^0\}$$

- **Calculate inter-frame differences via *L2 Distance***

$$F_1^L(k) = \frac{dis(F_k^0, F_{k+1}^0)}{\Delta t}$$

- **Evaluate standard deviation of second-order features**

$$F_2(k) = \frac{F_1(k) - F_1(k-1)}{\Delta t}$$

$$\sigma(F_2) = \sqrt{\frac{1}{T-3}\sum_{i=2}^{T-1}(F_2(i) - \mu)^2}$$

AI-Video Detection: Results

- We perform the detection experiments on baselines and our **training-free** method across 4 different datasets: *GenVideo*, *EvalCrafter*, *VideoPhy*, and *VidProM*.

Detection Method	Detection Level	Datasets (AP↑) Crafter	Gen2	HotShot	Lavie	MSE	MV	MSO	Show-1	Sora	WS	mAP
FID	Image	92.41	93.27	<u>86.10</u>	83.68	<u>91.50</u>	93.67	92.24	<u>90.61</u>	74.95	<u>82.24</u>	<u>88.07</u>
NPR	Image	97.02	96.35	40.17	22.37	84.67	96.79	96.53	21.61	<u>90.55</u>	66.51	71.26
STIL	Image	85.82	93.19	40.61	53.24	58.99	94.94	71.62	47.73	22.35	61.91	63.04
MINITIME	Video	88.62	60.66	39.03	82.29	23.85	74.79	74.33	41.08	16.92	72.25	57.38
FTCN	Video	95.41	97.18	37.47	44.90	79.71	99.75	97.05	17.33	83.69	66.86	71.94
TALL	Video	87.85	93.47	44.00	59.07	51.11	92.09	63.63	51.06	15.82	64.43	62.25
XCLIP	Video	97.32	**99.44**	44.68	72.69	88.00	**99.96**	97.53	38.37	71.08	74.00	78.31
AIGVDet	Video	75.87	89.98	51.81	<u>88.62</u>	70.91	56.22	67.93	72.59	65.70	64.96	70.46
Demamba	Video	<u>97.91</u>	99.16	52.97	76.72	82.83	<u>99.80</u>	<u>98.42</u>	56.24	77.75	74.81	81.66
Our D3	Video	**98.53**	<u>99.39</u>	**98.52**	**97.22**	**97.12**	99.52	**98.68**	**99.18**	**99.91**	**96.49**	**98.46**

Table 1. Detection results on Video datasets. Our D3 is training-free, while the baselines are trained on real videos from Youku-mPLUG [49] and AI-generated videos from Pika [8], following the setting in Demamba [17]. **Bold** represents the best and <u>underline</u> represents the second best.

AI-Video Detection: Results

- We perform the detection experiments on baselines and our **training-free** method across 4 different datasets: *GenVideo*, *EvalCrafter*, *VideoPhy*, and *VidProM*.

Detection	Datasets (AP↑)														mAP
Method	MV	Floor32	Gen2	Gen2-D	HotShot	LaVie-V	LaVie-I	Mix-SR	MSE	Pika	Pika-v1	Show-1	VC	ZS	
FID	98.29	96.4	97.36	98.68	89.9	92.92	84.19	98.51	95.74	99.49	99.17	96.77	95.71	95.18	95.59
NPR	**99.96**	**99.77**	99.34	**99.95**	47.39	76.45	72.23	**99.67**	**98.54**	**99.97**	**99.93**	69.82	**99.68**	98.21	90.07
AIGVDet	56.50	67.84	71.86	74.24	51.46	73.81	70.72	57.64	71.00	94.95	92.92	72.41	64.58	67.00	70.50
Demamba	99.49	91.76	96.98	99.27	34.60	56.89	37.85	97.49	71.33	98.69	99.33	26.83	94.30	64.39	76.37
Our D3	99.52	98.68	**99.46**	99.74	**98.52**	**97.79**	**98.48**	99.16	97.13	99.43	99.55	**99.18**	98.77	**98.83**	**98.87**

Table 2. Detection results on 14 EvalCrafter datasets.

Detection	Datasets (AP↑)										mAP
Method	LaVie	OpenSora	CogVideoX-5B	CogVideoX	Dream-Machine	Gen-2	Pika	SVD	VC2	ZeroScope	
FID	96.51	87.9	91.41	93.34	97.5	98.35	99.55	95.66	96.03	90.6	94.69
NPR	63.72	88.78	81.99	81.37	**99.86**	**99.90**	**99.91**	**99.54**	60.21	78.23	85.35
AIGVDet	61.06	59.07	58.95	63.15	59.27	61.55	92.96	53.73	58.22	63.11	63.11
Demamba	28.80	16.00	24.35	22.97	94.03	97.52	96.75	87.28	23.86	23.17	51.47
Our D3	**98.49**	**98.55**	**99.03**	**98.87**	99.54	99.87	99.70	98.75	**99.46**	**99.38**	**99.16**

Table 3. Detection results on 10 VideoPhy datasets.

AI-Video Detection: Results

- We perform the detection experiments on baselines and our **training-free** method across 4 different datasets: *GenVideo*, *EvalCrafter*, *VideoPhy*, and *VidProM*.

Detection Method	Datasets (AP↑) MSE	OS	Pika	ST2V	T2VZ	VC2	mAP
FID	<u>91.35</u>	87.68	<u>99.59</u>	**97.87**	68.51	<u>85.92</u>	**88.49**
NPR	87.04	<u>89.85</u>	**99.98**	89.88	**88.93**	70.79	87.75
AIGVDet	63.33	62.12	66.07	55.46	63.49	52.15	60.44
Demamba	58.73	85.87	99.34	86.48	<u>79.62</u>	80.28	81.72
Our D3	**96.85**	**97.85**	99.14	<u>93.13</u>	45.11	**98.70**	<u>88.46</u>

Table 4. Detection results on 6 VidProM datasets.

- Existing video generators **cannot accurately model the second-order features** of real videos.
- We can **realize accurate detection by calculating the second-order features** using mathematical methods.

AI-Video Detection: Results

- **We conduct an ablation study to see how the choice of visual encoder and first-order calculation method affects D3's performance.**

Visual Encoder	GenVideo		EvalCrafter		VideoPhy		VidProM	
	L2	Cos	L2	Cos	L2	Cos	L2	Cos
DINOv2-B	95.84	87.17	96.76	89.31	93.98	82.14	82.17	73.23
DINOv2-L	94.92	85.33	95.84	87.31	92.49	79.12	80.90	70.83
CLIP-B/16	<u>97.00</u>	87.82	97.63	89.82	97.01	86.24	84.79	75.77
XCLIP-B/16	**97.72**	<u>91.30</u>	**98.24**	<u>92.81</u>	<u>97.14</u>	<u>89.10</u>	**87.08**	**79.87**
CLIP-B/32	96.73	87.87	97.26	89.53	96.61	87.04	83.97	75.52
XCLIP-B/32	96.99	90.43	<u>97.72</u>	92.31	96.35	88.74	<u>85.57</u>	<u>79.62</u>
ResNet-18	96.39	89.73	97.26	91.64	95.67	86.83	81.59	75.68
VGG-16	96.97	**92.63**	<u>97.84</u>	**94.16**	**97.50**	**91.21**	81.54	77.02
EfficientNet-B4	94.28	85.51	95.49	88.08	92.46	82.40	80.73	73.00
MobileNet-V3	95.47	87.14	96.48	89.50	94.70	84.71	80.76	73.74

- *L2 Distance* contains more inter frame features.
- **Large-scale**, pre-trained encoders (e.g. CLIP or XCLIP) perform better.
- Nonetheless, lightweight visual encoders **still possessing excellent performance**.

AI-Video Detection: Results

- We compared the **robustness to post-processing operations** and **real-time efficiency** of the baselines and D3.

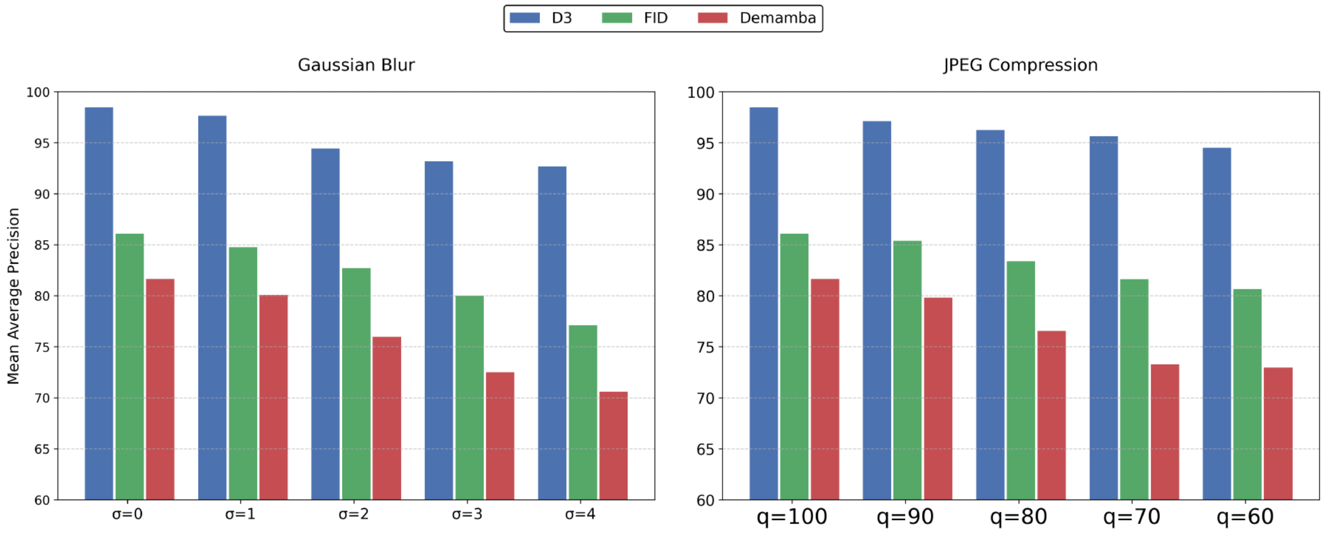


Figure 3. Detection results (mAP) of baselines and D3 against post-processing operations on Genvideo.

Detection Method	Time (s,↓) Preprocess	Train	Inference	mAP↑ on GenVideo
FID	Free	415	213	88.07
NPR	Free	256	188	71.26
AIGVDet	500	642	74	70.46
Demamba	Free	196	91	81.66
D3 (XCLIP-B/16)	Free	Free	<u>56</u>	**98.46**
D3 (MobileNet-v3)	Free	Free	**40**	<u>95.47</u>

Table 5. Efficiency results on GenVideo with 1000 video samples and batch size of 1. The preprocessing overhead of AIGVDet comes from the optical flow extraction using RAFT. For image-level methods (FID, NPR), 8 images form a video.

- D3 demonstrates **strong robustness** and **computational efficiency**
- Attributed to **second-order feature hypothesis** and **training-free** framework.

Thank you for listening!

Wechat

arXiv

github