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

**Chende Zheng**[1] **Ruiqi Suo**[1] **Chenhao Lin**[2*] **Zhengyu Zhao**[1] **Le Yang**[1]
**Shuai Liu**[1*] **Minghui Yang**[2] **Cong Wang**[3] **Chao Shen**[1]
[1]Xi'an Jiaotong University [2]Guangdong OPPO Mobile Communications Co., Ltd.
[3]City University of Hong Kong

## Abstract

*The evolution of video generation techniques, such as Sora, has made it increasingly easy to produce high-fidelity AI-generated videos, raising public concern over the dissemination of synthetic content. However, existing detection methodologies remain limited by their insufficient exploration of temporal artifacts in synthetic videos. To bridge this gap, we establish a theoretical framework through second-order dynamical analysis under Newtonian mechanics, subsequently extending the Second-order Central Difference features tailored for temporal artifact detection. Building on this theoretical foundation, we reveal a fundamental divergence in second-order feature distributions between real and AI-generated videos. Concretely, we propose Detection by Difference of Differences (D3), a novel training-free detection method that leverages the above second-order temporal discrepancies. We validate the superiority of our D3 on 4 open-source datasets (GenVideo, VideoPhy, EvalCrafter, VidProM), 40 subsets in total. For example, on GenVideo, D3 outperforms the previous state-of-the-art method by 10.39% (absolute) mean Average Precision. Additional experiments on time cost and post-processing operations demonstrate D3's exceptional computational efficiency and strong robust performance. Our code is available at https://github.com/Zig-HS/D3.*

## 1. Introduction

With the development of AI [23, 37, 53], generative models have evolved into versatile tools for high-fidelity content creation. However, their pervasive deployment across digital ecosystems has precipitated critical societal security risks—systematically undermining information integrity and eroding public trust [12, 38, 57] - evidenced by cases like the Taylor Swift deepfakes [2]. Consequently, there is an increasing and urgent demand to develop a detector of AI-generated videos. Traditional research on deep forgery detection focuses on facial forgery (e.g., Deepfakes) but is unable to generalize to universal videos. For this, there has been some recent aiming at the generalization detection of AI-generated videos, e.g., DuB3D [27], DeCoF [33], DeMamba [16], etc. These methods commonly learn the differences between real and AI-generated videos from the training data by using deep learning frameworks.

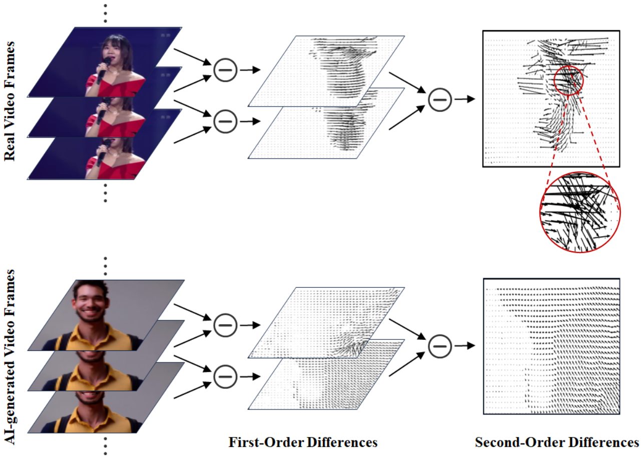


Figure 1. Real and AI-generated videos can be differentiated by second-order (temporal) features in our **D**etection by **D**ifference of **D**ifferences (D3) method. The optical flow vectors are obtained by RAFT [42]. First-order features are less powerful (see Table 8.)

However, despite these efforts, there is still a lack of interpretability research on AI-generated video detection. To address this, we introduce the second-order position control system under Newtonian mechanics theory. We extend the *Second-order Central Difference* features from this system to AI-generated video detection and validate significant differences in second-order features between real and AI-generated videos by conducting visualization experiments on differential optical flow. More concretely, we propose Detection by Difference of Differences (D3), a training-free AI-generated video detection framework based on second-order features.

Technically, we use a pre-trained visual encoder to extract zero-order features from the input video by frames. Then, we employ L2 distance (or cosine similarity) to assess the inter-frame differences as first-order features.

[*] Corresponding Authors

We further compute the second-order features according to the *Second-order Central Difference* formula. Our empirical findings show that the second-order features of real videos exhibit more significant volatility, while AI-generated videos tend to follow a flatter pattern (Figure 1). Therefore, we quantify the volatility by calculating the standard deviation of the second-order features and realizing the generalized detection of AI-generated videos. Besides, to comprehensively evaluate the generalization ability of D3, we conduct an open-world evaluation, which validate D3's superior detection performance.

Our main contributions can be summarized as follows:

- By rethinking Newtonian mechanics, we innovatively introduce the second-order central difference features of videos. Our experiments reveal the differences in second-order features between existing AI-generated videos and real videos.
- We propose D3, a novel training-free AI-generated video detection method. By extracting second-order features, our detector is capable of generalizing across various generators.
- The extensive experimental results on 4 different open-source datasets, including 40 test subsets, demonstrate the state-of-the-art (SOTA) generalization performance as well as strong robustness to the post-processing operations of our method.

## 2. Related Work

### 2.1. Video Generation Methods

Recent advances in video synthesis, predominantly built on diffusion models, focus on text-to-video, image-to-video, and combined text-image approaches. A core challenge remains ensuring logical coherence and smooth temporal continuity. To address this, Text2Video-Zero [28] enriches latent codes with motion dynamics, while Zhang et al. [55] leverage visual guidance in I2VGen-XL for coherent high-resolution generation. Similarly, Xing et al. [48] and Chen et al. [15] incorporate motion cues—via video diffusion priors and generation-stage cues, respectively—to simulate realistic motion and enhance frame transitions. For long videos, SEINE [18] automates smooth transitions using scene images with text control.

Regarding datasets, SVD [13] demonstrates that fine-tuning on small, curated datasets yields higher-quality, stable models compared to non-curated alternatives.

While these video generation methods have yielded promising results, we have found that they still produce videos that do not fully comply with the physical laws of the real world. Therefore, there is still significant room for improvement in video generation.

### 2.2. AI-Generated Video Detection

As video synthesis quality advances, effective AIGC detection for videos becomes increasingly critical. Traditional deepfake detection focuses on facial artifacts (e.g., distortions in landmarks [51] or head pose inconsistencies [30]). These methods struggle with complex scenes. Recent approaches shift toward global characteristics: NPR [41] analyzes neighboring pixel relationships, while others leverage pre-trained models [9, 36, 40] or data augmentation [20, 26] to improve diffusion-image detection.

For videos generated by Diffsion Models, recent advances now extend to universal detection, exemplified by DeMamba [16], which introduced a dedicated Mamba module for video detection and developed the Gen-video dataset, which was specifically designed for AI-generated video detection tasks. In a similar vein, Liu et al.[31] proposed a CNN+LSTM architecture that utilizes DIRE[46] residuals to classify videos as real or generated, while DeCoF [33] presented a detector that focuses on temporal artifacts, aiming to identify AI-generated videos by analyzing these specific features. These approaches contribute to the growing field of AI-generated video detection, each addressing unique aspects of the challenge.

However, although these methods strive to distinguish between real and AI-generated videos, they still lack a deep analysis of temporal artifacts, resulting in a missing enlightening motivation for the interpretable detection of AI-generated videos.

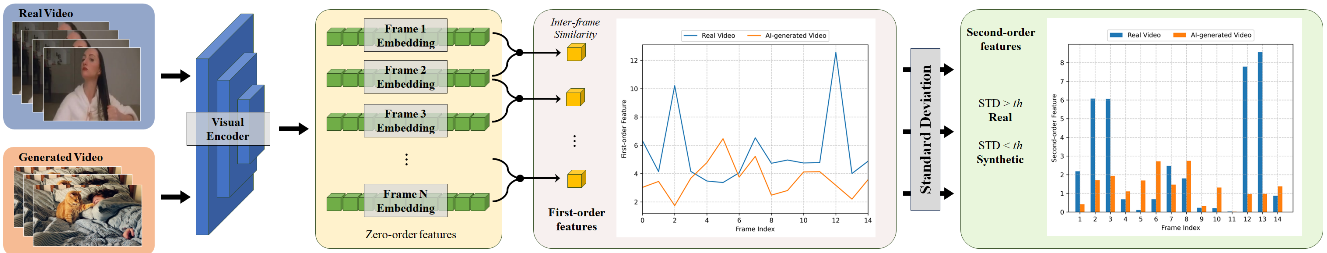
## 3. Methodology

### 3.1. Motivation

The key to detecting AI-generated videos lies in identifying the artifacts that differ from real videos. Existing detection methods focus on pixel-level temporal artifacts of specific regions (e.g., lips and facial edges). However, as the quality of the generated video improves, the generalization performance of these methods continually decreases. An effective solution is to analyze the artifacts from a theoretical perspective. For this, we propose an analysis method based on second-order features to investigate the differences in second-order features between real and AI-generated videos. Specifically, we start by modeling a second-order position control system under Newtonian mechanics in the real world, which can be represented by the following equation:

$$A_2 \frac{d^2 x(t)}{dt^2} + A_1 \frac{dx(t)}{dt} + A_0 x(t) = u(t) \tag{1}$$

where $A_2$ is the inertia coefficient (i.e. second-order coefficient), $A_1$ is the damping coefficient, $A_0$ is the elasticity coefficient, and $u(t)$ represents the input force. Note that real-world systems are typically higher-order systems,



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

Figure 2. Framework of our training-free detection method D3. For a given video, (a) zero-order, (b) first-order, and (c) second-order features are subsequently extracted.

but they can be reduced to second-order systems using the *Dominant Pole Approximation* [35]).

According to *the Principles of Automatic Control*, when solving a second-order control system, we use the second-order central difference method to approximate the derivative of the differential equation. In other words, the second-order ordinary differential equation can be discretized to obtain a numerical solution. Therefore, we can approximate the acceleration $f''(x)$ (i.e. second-order feature) using the *Second-order Central Difference* [35] as follows:

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

$$= \frac{f'(x) - f'(x-h)}{h} \tag{3}$$

where $x$ is the time point, $h$ is the sampling interval, and $f'(x)$ represents the first-order difference feature.

A correct mechanical simulation model should adhere to the physical laws of the real world, which means that the simulated acceleration (i.e., the second-order feature) should follow the same paradigm as real-world objects. Therefore, a video generation model is expected to ensure the second-order features of the videos to exhibit similar patterns in real videos.

To verify whether existing video generators can accurately fit the second-order features of real videos, we conduct a visualization experiment on both real and AI-generated videos by extracting optical flow, using RAFT [42]. In this experiment, we extract two optical flow features from the video $X = x_1, x_2, \ldots, x_N$. The first optical flow is calculated between frames $x_t$ and $x_{t+1}$, while the second optical flow is calculated between frames $x_{t+1}$ and $x_{t+2}$. These optical flows represent the speed of pixel change over time, i.e., the first-order difference feature in Formula (2). Therefore, the difference between these two optical flows, $X_{diff}$, reflects the change in optical flow speed from frame $x_t$ to frame $x_{t+2}$ (i.e., optical flow acceleration), and is expressed by the following formula:

$$X_{diff} = \frac{OF(x_{t+1}, x_{t+2}) - OF(x_t, x_{t+1})}{\Delta t^2} \tag{4}$$

where $OF(x, y)$ represents the optical flow between frames $x$ and $y$, and $\Delta t$ is the sampling interval.

Figure 1 displays the visualized optical flow and the corresponding vector diagrams. The results reveal a clear distinction in the second-order features between real videos and AI-generated videos. Real videos exhibit more chaotic speed variations, as shown in Figure 1 (1)-(d) & (2)-(d), reflecting the higher complexity and diversity driven by various influencing factors in real-world scenes. In contrast, generated videos tend to show very flat patterns, which could be due to existing generators' difficulty in simulating second-order dynamics, resulting in ”smoother” video outputs constrained within the distribution range of their training data. These observations support our hypothesis that current generators fail to accurately replicate real videos' second-order features, providing motivation and opportunity for AI-generated video detection through second-order feature analysis.

## 3.2. Detection by Difference of Differences

We have confirmed the distortion of second-order features in existing video generators on pixel level and optical flow level, however, these features are challenging to compute directly. A feasible approach is to use a visual encoder to perform feature dimensionality reduction on video frames, transforming pixel-level features into deep representations. Based on this, we introduce a training-free mathematical detection method based on second-order features, aiming to realize the Detection on Difference of Differences (D3).

Given an input video $X \in \mathbb{R}^{T\times 3\times H\times W}$, which is sampled into a sequence of frames $X = X_1, X_2, \ldots, X_T$ at a regular intervals of $\Delta t$. We utilize visual encoders (e.g., DINOv2, XCLIP or pre-trained ResNet-18, etc.) to encode the input frames into a sequence of features $F_0 = F_0^1, \ldots, F_0^T, F_0 \in \mathbb{R}^{T\times N}$. Within the feature space, the first-order features are first extracted. Specifically, we use *L2 Distance* or *Cosine Similarity* to calculate inter-frame similarity, as the first-order difference features, formulated as follows:

$$F_1^{L2}(k) = \frac{dis(F_0^k, F_0^{k+1})}{\Delta t}, \quad k = 1, 2, ..., T-1 \quad (5)$$

$$F_1^{Cos}(k) = \frac{sim(F_0^k, F_0^{k+1})}{\Delta t}, \quad k = 1, 2, ..., T-1 \quad (6)$$

where $k$ is the frame index. Then, we further compute the second-order central difference feature according to Formula (2), as follows:

$$F_2(k) = \frac{F_1(k) - F_1(k-1)}{\Delta t}, \quad k = 2, ..., T-1 \quad (7)$$

We present the first-order and second-order features (absolute values) extracted from real and AI-generated videos in Figure 2. The results show that, compared to generated videos, the temporal second-order features of real videos exhibit more pronounced fluctuations, which is consistent with the conclusion drawn in Figure 1.

To measure this volatility, we calculate the standard deviation of the second-order features, as the following formula:

$$\sigma(F_2) = \sqrt{\frac{1}{T-3}\sum_{i=2}^{T-1}(F_2(i) - \frac{1}{T-3}\sum_{i=2}^{T-1}F_2(i))^2} \quad (8)$$

We use this standard deviation as the final output to classify real and generated videos. The overall pipeline of D3 is shown in Figure 2. Compared to existing image or video detection methods, our approach has significant advantages: A) D3 is training-free, consisting solely of an inference process, and does not require generated videos. B) D3 demonstrates exceptional computational efficiency, with the primary computational cost stemming from visual feature extraction. Furthermore, our experimental results (see Section 5.1) indicate that D3 remains effective even when using lightweight feature extractors.

# 4. Experiments

## 4.1. Experimental Setup

**Training Datasets.** Considering that D3 operates solely during inference, the experiments require no training datasets. However, for a comprehensive comparison, we still set up a training dataset for baselines. Specifically, following the settings from DeMamba [16], we use real videos from Youku-mPLUG [49] and AI-generated videos from Pika [8] to train baselines.

**Test Datasets.** To assess the generalization ability of our approach to real-world scenarios, we adopt the 4 out-of-distribution datasets, including 40 test sets:

- **GenVideo** [16]: ModelScope (MSE) [43], MorphStudio (MSO) [7], MoonValley (MV) [6], HotShot [5], Show_1 [54], Gen2 [21], Crafter [15], LaVie [45], Sora [14], and WildScrape (WS) [47].
- **EvalCrafter** [32]: MoonValley (MV), Floor33 [1], Gen2, Gen2-December (Gen2-Dec), HotShot, LaVie-Base (LaVie-B), LaVie-Internet (LaVie-I), Mix-SR, ModelScope, Pika, Pika_v1, Show_1, VideoCrafter (VC) [15], and ZeroScope (ZS) [4].
- **VideoPhy** [11]: LaVie, OpenSora[59], CogVideoX[52], CogVideoX-5B, Dream-Machine[3], Gen2, Pika, SVD[13], VideoCrafter2 (VC2)[17], and ZeroScope.
- **VidProM** [44]: ModelScope (MSE), OpenSora (OS), Pika, StreamingT2V (ST2V) [25], Text2video-zero (T2VZ)[29], and VideoCrafter2 (VC2).

**Baselines.** We perform comparisons of our approach with existing popular and state-of-the-art detectors, including 3 image-level detectors, FID (NeurIPS'24) [56], NPR (CVPR'24) [39] and STIL (MM'21) [24], 6 video-level detectors, FTCN (ICCV'21) [58], MINITIME (TIFS'24) [19], TALL (ICCV'23) [50], XCLIP (ECCV'22) [34], AIGVDet [10], and DeMamba [16].

We re-implement baselines [10, 16, 39, 56] with the official codes using our training set. We report the results of the baselines [16, 19, 24, 34, 50, 58] on GenVideo dataset from [16].

**Implementation Details.** We adopt pre-trained XCLIP-B/16 [34] as the visual encoders to extract zero-order features and L2 Distance to calculate first-order features. During inference, we extract a segment from the input video (up to 2 seconds) and frames are sampled at equal intervals of 8 frames per second. All frames are set to JPEG format. For pre-processing, we crop 10% of the longer edge of all frames and then resize the frames to 224 × 224 pixels. We adopt the Average Precision (AP) and the Area Under the Receiver Operating Characteristic curve (AUROC, AUC) as the evaluation metric, which is widely used in baselines [16, 19, 24, 34, 39, 50, 58]. (AUC results are provided in the Supplementary Materials.) All of our experiments are conducted using PyTorch on AMD EPYC 7763 64-Core CPU and NVIDIA GeForce RTX 4090 Tensor Core GPU.

## 4.2. Detection on GenVideo

We conduct a comprehensive comparative analysis of various AI-generated video detectors, evaluating their generalization performance on the GenVideo dataset. The AP results are presented in Table 1. As can be seen, our D3 achieves the base overall results. In addition, except for FID [56], image-level detectors suffer from performance degradation on generated videos, which is attributed to

| Detection | Detection | Datasets (AP↑) | | | | | | | | | | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | Level | Crafter | Gen2 | HotShot | Lavie | MSE | MV | MSO | Show-1 | Sora | WS | |
| FID [56] | Image | 92.41 | 93.27 | 86.10 | 83.68 | 91.50 | 93.67 | 92.24 | 90.61 | 74.95 | 82.24 | 88.07 |
| NPR [39] | Image | 97.02 | 96.35 | 40.17 | 22.37 | 84.67 | 96.79 | 96.53 | 21.61 | 90.55 | 66.51 | 71.26 |
| STIL [24] | Image | 85.82 | 93.19 | 40.61 | 53.24 | 58.99 | 94.94 | 71.62 | 47.73 | 22.35 | 61.91 | 63.04 |
| MINITIME [19] | Video | 88.62 | 60.66 | 39.03 | 82.29 | 23.85 | 74.79 | 74.33 | 41.08 | 16.92 | 72.25 | 57.38 |
| FTCN [58] | Video | 95.41 | 97.18 | 37.47 | 44.90 | 79.71 | 99.75 | 97.05 | 17.33 | 83.69 | 66.86 | 71.94 |
| TALL [50] | Video | 87.85 | 93.47 | 44.00 | 59.07 | 51.11 | 92.09 | 63.63 | 51.06 | 15.82 | 64.43 | 62.25 |
| XCLIP [34] | Video | 97.32 | **99.44** | 44.68 | 72.69 | 88.00 | **99.96** | 97.53 | 38.37 | 71.08 | 74.00 | 78.31 |
| AIGVDet [10] | Video | 75.87 | 89.98 | 51.81 | 88.62 | 70.91 | 56.22 | 67.93 | 72.59 | 65.70 | 64.96 | 70.46 |
| Demamba [16] | Video | 97.91 | 99.16 | 52.97 | 76.72 | 82.83 | 99.80 | 98.42 | 56.24 | 77.75 | 74.81 | 81.66 |
| Our D3 | Video | **98.53** | 99.39 | **98.52** | **97.22** | **97.12** | 99.52 | **98.68** | **99.18** | **99.91** | **96.49** | **98.46** |

Table 1. Detection results on GenVideo 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 [16]. **Bold** represents the best and underline represents the second best.

| 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.

| Detection | Datasets (AP↑) | | | | | | mAP |
|---|---|---|---|---|---|---|---|
| Method | MSE | OS | Pika | ST2V | T2VZ | VC2 | |
| FID | 91.35 | 87.68 | 99.59 | **97.87** | 68.51 | 85.92 | **88.49** |
| NPR | 87.04 | 89.85 | **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 | 79.62 | 80.28 | 81.72 |
| Our D3 | **96.85** | **97.85** | 99.14 | 93.13 | 45.11 | **98.70** | 88.46 |

Table 4. Detection results on 6 VidProM datasets.

their inability on video-level temporal artifacts. Meanwhile, MINITIME, FTCN, and TALL also perform poorly on GenVideo, because they are designed for detecting forged facial videos.

Interestingly, FID demonstrates strong generalization, which can be attributed to its unique generalization design. FID focuses on the local feature information of images, which enables it to remain unaffected by specific semantic scenes. Nevertheless, compared to the latest video detection methods or fine-tuned large-scale visual models, D3 demonstrates superior performance on GenVideo. Specifically, the mean AP of the D3 method reached 98.46%, outperforming state-of-the-art FID by 10.39% (absolute) mean AP.

Note that D3 is entirely training-free and does not require additional generated videos. The results validate our hypothesis that existing video generators **cannot** accurately model the second-order features of real videos. Furthermore, based on this hypothesis, we can realize accurate detection by calculating the second-order features using mathematical methods.

## 4.3. Detection on More Challenging Datasets

To further assess the generalizability of D3, we extend the evaluation to three more challenging open-source datasets (EvalCrafter, VideoPhy, VidProM), with results presented in Table 2, 3, & 4. These results demonstrate the consistent

| Detection | Time (s,↓) | | | mAP↑ |
|---|---|---|---|---|
| Method | Preprocess | Train | Inference | 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 | 56 | **98.46** |
| D3 (MobileNet-v3) | Free | Free | **40** | 95.47 |

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.

superiority of D3.

In these additional experiments, we can observe that image-level detectors (FID and NPR) exhibit stronger generalization than video-level baselines. This is interesting, as it contradicts conclusions from recent research (e.g., DeCoF [33] and DeMamba [16]). We attribute this to two factors: 1) FID and NPR are designed for generalization in cross-scene and cross-generator settings; 2) Due to changes in generative models, AI-generated videos on these datasets exhibit more diverse video artifact features. However, universal image-level artifacts still persist (e.g., upsampling artifacts [22]).

Besides, we find that D3 performs poorly on the T2VZ dataset. We attribute this to the low generation quality of T2VZ, which leads to poor semantic consistency in the generated videos, making them resemble chaotic images rather than dynamic videos. A detailed discussion is provided in the Supplementary Materials.

Despite this, D3 achieves impressive mean APs, outperforming the best-performing baselines. In sum, the results across 40 test subsets underscore the remarkable generalization capability of D3, which can be attributed to D3's successful identification of the substantial differences in second-order features between real and AI-generated videos, further highlighting the limitations of current video generators in fitting second-order features.

## 4.4. Efficiency Comparisons

In Table 5, we present the time costs per 1,000 video samples for baselines and D3 across different stages. The results demonstrate D3's superior computational efficiency. Unlike deep learning-based classifiers, D3 is training-free, eliminating substantial preprocessing and training overheads. For instance, on our training set (192,000 samples), Demamba requires 10.45 hours per epoch (batch size 1) and 3.25 hours per epoch (batch size 32). When the training set is set to a larger open-source dataset (e.g., GenVideo training set with a total of 2,262,086 samples), the advantages of D3 become even more pronounced.

In addition, D3 maintains optimal efficiency during inference (56s per 1,000 samples using XCLIP-B/16). D3 further supports lightweight networks (e.g., 40s per 1,000 samples via MobileNet-v3), enabling enhanced computational efficiency and localized deployment.

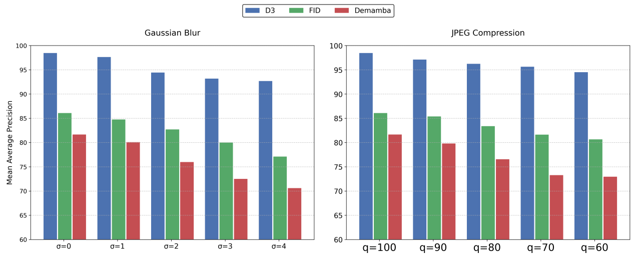


Figure 3. Detection results (mAP) of baselines and D3 against post-processing operations on Genvideo.

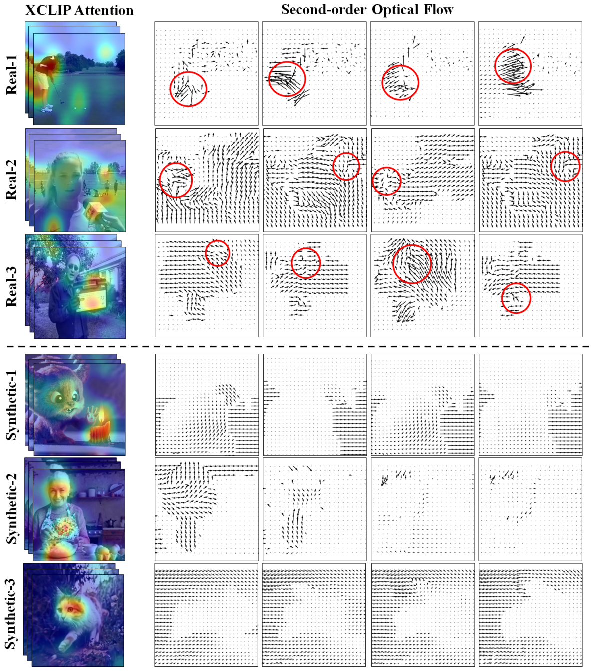


Figure 4. Visualizations of XCLIP attention and 2nd-order flows.

## 4.5. Qualitative Analysis

To further demonstrate the effectiveness of D3, we conduct qualitative analysis by visualizing XCLIP attention and second-order optical flow, as shown in Figure 4.

We can see from Figure 4 that variations in 2nd-order flow appear around moving semantics or objects (highlighted by the XCLIP attention). This is aligned with physical principles (e.g., Newtonian inertia) stating that the object's motion in real-world scenarios follows high-order dynamics.

Besides, the results demonstrate that current video generators fail to learn similar second-order patterns from real videos well (see second-order optical flow in Figure 4),

| Visual Encoder | Gaussian Blur | | | | | JEPG Compression | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0$ | $\sigma = 1$ | $\sigma = 2$ | $\sigma = 3$ | $\sigma = 4$ | $q = 100$ | $q = 90$ | $q = 80$ | $q = 70$ | $q = 60$ |
| DINOv2-B | 96.84 | 96.00 | 93.39 | 90.42 | 88.34 | 96.84 | 95.60 | 94.93 | 94.36 | 93.59 |
| DINOv2-L | 96.23 | 95.28 | 92.00 | 88.85 | 87.10 | 96.23 | 94.65 | 93.93 | 93.16 | 92.53 |
| CLIP-B/16 | 97.82 | 97.16 | 83.70 | 83.20 | 84.89 | 97.82 | 84.86 | 83.31 | 81.59 | 78.81 |
| XCLIP-B/16 | **98.46** | 97.63 | 94.43 | 93.19 | 92.69 | **98.46** | 97.11 | 96.24 | 95.63 | 94.50 |
| CLIP-B/32 | 97.71 | 97.05 | 84.50 | 85.13 | 86.56 | 97.71 | 91.65 | 89.79 | 88.04 | 86.40 |
| XCLIP-B/32 | 98.15 | **97.64** | **95.51** | **94.22** | **93.59** | 98.15 | **97.59** | **97.37** | **97.25** | **96.93** |
| ResNet18 | 97.35 | 96.60 | 92.84 | 90.70 | 90.03 | 97.35 | 96.61 | 95.48 | 94.09 | 92.48 |
| VGG16 | 97.21 | 95.56 | 90.91 | 87.28 | 85.60 | 97.21 | 94.06 | 92.35 | 89.47 | 87.58 |
| EfficientNet-b4 | 96.53 | 95.44 | 90.63 | 88.59 | 88.13 | 96.53 | 94.81 | 93.37 | 91.85 | 90.54 |
| MobileNet-v3 | 96.86 | 95.92 | 87.91 | 84.66 | 84.38 | 96.86 | 94.58 | 91.29 | 88.31 | 85.95 |

Table 6. Detection results of D3 against post-processing operations on GenVideo.

which explains the effective detection performance of D3 using second-order video features.

### 4.6. Robustness to Post-processing Operations

In real-world scenarios, videos are seldom pristine. As videos circulate in cyberspace, they are continuously subjected to compression and interference, potentially leading to a performance degradation of video detectors. To address this, in this section, we evaluate the robustness of D3 against various post-processing operations. Specifically, we consider five different levels of Gaussian blur ($\sigma$ = 0, 1, 2, 3, 4) and JPEG compression with five different quality factors ($q$ = 100, 90, 80, 70, 60). We use the mean AP on GenVideo as the metric.

Figure 3 reveals detection results of baselines and D3 (using XCLIP-B/16 and L2 Distance) against post-processing, with D3 demonstrating the strongest robustness through minimal degradation under Gaussian blur and JPEG compression. As the degree of post-processing operations increases, D3 maintains high detection performance, indicating its excellent robustness against post-processing operations. However, FID and Demamba show pronounced vulnerability, suffering significant performance declines, reflecting their reliance on high-frequency details and susceptibility to spectral artifacts.

Table 6 reports the results of the robustness experiments of different variants of D3. As can be seen, different visual encoders exhibit varying levels of robustness. For example, MobileNet-V3 shows a noticeable performance drop when confronted with ($\sigma$ = 4) Gaussian blur (from 96.86% to 84.38%), whereas XCLIP-B/16 exhibits a smaller decrease (from 98.46% to 92.69%). This is understandable because MobileNet is a lightweight vision network with a smaller parameter scale. These results reveal that the detection robustness using second-order features depends on the stability of the feature space. Therefore when using self-supervised visual models based on ViT (e.g., XCLIP-B/16), D3 demonstrates better robustness.

## 5. Ablation Studies

### 5.1. Impact of Visual Encoders

So far, we have demonstrated the effectiveness of D3 in generalized detection. In the above experiments, D3 utilizes the pre-trained XCLIP-ViT-B/16 model, which raises a new question of whether generated video detection using second-order coefficients relies on large-scale visual encoders. To address this, we conducted several ablation experiments using different visual encoders. We adopt several large-scale self-supervised models based on ViT, including CLIP-ViT, XCLIP-ViT, DINOv2, and their variants with different patch sizes or parameter scales. Additionally, we adopt CNN-based models for classification, e.g., ResNet18, VGG16, EfficientNet-B4, or lightweight networks like MobileNet. These CNN-based models are all pre-trained on ImageNet.

The results of the ablation experiments are shown in Table 7. An intuitive conclusion is that large-scale visual encoders perform best in our experiments, e.g., CLIP-ViT-B/16 and XCLIP-ViT-B/16. Among the ViT-based models, the impact of patch size and parameter scale is negligible, as evidenced by the minimal differences between XCLIP-ViT-B/16 and XCLIP-ViT-B/32.

Nonetheless, despite the significant differences in parameter scale or model architecture among these visual encoders, the differences in detection performance are small. For example, as a lightweight model, MobileNet-V3 still produces an impressive result of 96.31%. This phenomenon is encouraging because it confirms that second-order features remain meaningful across different encoder feature spaces. This insight helps us understand the shortcomings of current video generators in simulating reality.

| 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 | 97.00 | 87.82 | 97.63 | 89.82 | 97.01 | 86.24 | 84.79 | 75.77 |
| XCLIP-B/16 | **97.72** | 91.30 | **98.24** | 92.81 | 97.14 | 89.10 | **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 | 97.72 | 92.31 | 96.35 | 88.74 | 85.57 | 79.62 |
| ResNet-18 | 96.39 | 89.73 | 97.26 | 91.64 | 95.67 | 86.83 | 81.59 | 75.68 |
| VGG-16 | 96.97 | **92.63** | 97.84 | **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 |

Table 7. Ablation studies of visual encoder backbones and the type of first-order features (L2 Distance or Cosine Similarity).

| Detection | GenVideo | | EvalCrafter | |
|---|---|---|---|---|
| Method | mAP↑ | Avg. AUC↑ | mAP↑ | Avg. AUC↑ |
| D3 (1st-Order) | 95.69 | 93.45 | 86.40 | 85.17 |
| D3 (2nd-Order) | **98.46** | **97.72** | **98.87** | **98.24** |
| | VideoPhy | | VidProM | |
| | mAP↑ | Avg. AUC↑ | mAP↑ | Avg. AUC↑ |
| D3 (1st-Order) | 86.06 | 84.22 | 80.61 | 77.31 |
| D3 (2nd-Order) | **99.16** | **97.14** | **88.46** | **87.08** |

Table 8. Ablation studies of feature order on 4 datasets.

### 5.2. Impact of First-order Features

In this section, we delve into the impact of different first-order feature extraction methods. Specifically, we adopt L2 distance and cosine similarity separately as the first-order feature. The experimental results are shown in Table 7.

These results indicate that using L2 distance as the first-order feature yields better performance. The key takeaway is that L2 distance evaluates the absolute distance between inter-frame features, while cosine similarity evaluates the relative distance. Cosine similarity can better mitigate the effects of differing feature dimensions. However, in our experiments, the visual encoder is fixed, and therefore, the output feature dimensions are fixed. Besides, cosine similarity will be influenced by the features of the initial frame, whereas L2 distance can accurately reflect the extent of video change within this fixed feature space. Therefore, it provides more effective information.

### 5.3. Second-Order vs. First-Order Features

In this section, we explore the impact of feature order on the D3 method. Specifically, we replace the second-order feature standard deviation in the original scheme with the first-order feature standard deviation and evaluate it on 4 datasets. Table 8 presents the results of our ablation study. As shown, D3 using first-order features achieves good performance (95.69% mAP and 93.45% Avg. AUC) on GenVideo. However, this performance cannot be generalized to more challenging datasets.

Overall, using second-order features provides stronger detection capability. These results suggest that, although there are some differences in the first-order features between real and AI-generated videos (as shown in Figure 1), such differences are not universal, while second-order differential features are more powerful.

## 6. Conclusion and Outlook

This paper bridges the theory of second-order control systems from Newtonian mechanics with video analysis by extending second-order central difference features for temporal artifact detection. Our systematic analysis reveals that existing AI-generated videos diverge from real videos in their second-order feature, establishing a novel physical perspective for investigating temporal artifacts in synthetic content. Building on these insights, we proposed D3, an innovative, training-free AI-generated video detection framework. By measuring the volatility of second-order features in videos through standard deviation, D3 achieves generalizable detection of AI-generated videos. Through extensive experiments, we demonstrate that D3 achieves state-of-the-art performance in detecting AI-generated videos across various generative models as well as strong robustness against post-processing operations.

Our approach paves a new way for understanding and differentiating between real and AI-generated videos based on the connection between video content and fundamental physical principles. Future investigations can explore this paradigm by examining additional dimensions (e.g., temporal channel relationships, RGB color space distributions, or bitrate characteristics) to deepen our understanding of generation artifacts. We believe this paper will inspire further research into the artifacts of AI-generated videos and contribute to the development of generalized detection.

## References

[1] Floor33 pictures discord server. https://www.morphstudio.com/. [Accessed 05-03-2025]. 4

[2] Taylor Swift deepfakes spark calls in Congress for new legislation — bbc.co.uk. https://www.bbc.co.uk/news/technology-68110476. 1

[3] Ai video generator — lumalabs.ai. https://lumalabs.ai/dream-machine/. [Accessed 05-03-2025]. 4

[4] Zeroscope — huggingface.co. https://huggingface.co/cerspense/zeroscope_v2_576w, . [Accessed 05-03-2025]. 4

[5] hotshotco/Hotshot-XL · Hugging Face — huggingface.co. https://huggingface.co/hotshotco/Hotshot-XL, . 4

[6] Moonvalley — moonvalley.ai. https://moonvalley.ai/. 4

[7] All-in-one AI video creation suite — morphstudio.com. https://www.morphstudio.com/. 4

[8] Pika — pika.art. https://pika.art/. 4, 5

[9] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 467–474, 2023. 2

[10] Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 460–470. Springer, 2024. 4, 5

[11] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 4

[12] Clark Barrett, Brad Boyd, Elie Bursztein, Nicholas Carlini, Brad Chen, Jihye Choi, Amrita Roy Chowdhury, Mihai Christodorescu, Anupam Datta, Soheil Feizi, et al. Identifying and mitigating the security risks of generative ai. *Foundations and Trends® in Privacy and Security*, 6(1):1–52, 2023. 1

[13] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2, 4

[14] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024. 4

[15] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023. 2, 4

[16] Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, et al. Demamba: Ai-generated video detection on million-scale genvideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024. 1, 2, 4, 5, 6

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 4

[18] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2023. 2

[19] Davide Alessandro Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. Mintime: Multi-identity size-invariant video deepfake detection. *IEEE Transactions on Information Forensics and Security*, 2024. 4, 5

[20] David C Epstein, Ishan Jain, Oliver Wang, and Richard Zhang. Online detection of ai-generated images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 382–392, 2023. 2

[21] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 4

[22] Xinyu Gong, Shiyu Chang, Yifan Jiang, and Zhangyang Wang. Autogan: Neural architecture search for generative adversarial networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3224–3234, 2019. 6

[23] Qiuyi Gu, Zhaocheng Ye, Jincheng Yu, Jiahao Tang, Tinghao Yi, Yuhan Dong, Jian Wang, Jinqiang Cui, Xinlei Chen, and Yu Wang. Mr-

cographs: Communication-efficient multi-robot open-vocabulary mapping system via 3d scene graphs. *IEEE Robotics and Automation Letters*, 2025. 1

[24] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021. 4, 5

[25] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024. 4

[26] Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1060–1068, 2022. 2

[27] 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. *arXiv preprint arXiv:2405.15343*, 2024. 1

[28] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023. 2

[29] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 4

[30] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. arxiv 2018. *arXiv preprint arXiv:1811.00656*, 1811. 2

[31] Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Turns out i'm not real: Towards robust detection of ai-generated videos. *arXiv preprint arXiv:2406.09601*, 2024. 2

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22139–22149, 2024. 4

[33] 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. *arXiv preprint arXiv:2402.02085*, 2024. 1, 2, 6

[34] 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 *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 4, 5

[35] Katsuhiko Ogata. Modern control engineering. 2020. 3

[36] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24480–24489, 2023. 2

[37] Xuran Pu, Jianjie Fang, Zhiyuan Deng, Xueqian Wang, Xinlei Chen, et al. A large language model-driven heterogeneous air-ground search swarm. In *ICLR 2025 Workshop on Embodied Intelligence with Large Language Models In Open City Environment*. 1

[38] Dilip Kumar Sharma, Bhuvanesh Singh, Saurabh Agarwal, Lalit Garg, Cheonshik Kim, and Ki-Hyun Jung. A survey of detection and mitigation for fake images on social media platforms. *Applied Sciences*, 13(19):10980, 2023. 1

[39] 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. *arXiv preprint arXiv:2312.10461*, 2023. 4, 5

[40] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 2

[41] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28130–28139, 2024. 2

[42] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 1, 3

[43] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Mod-

elscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 4

[44] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024. 4

[45] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023. 4

[46] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22388–22398, 2023. 2

[47] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6537–6549, 2024. 4

[48] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pages 399–417. Springer, 2025. 2

[49] Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, et al. Youku-mplug: A 10 million large-scale chinese video-language dataset for pre-training and benchmarks. *arXiv preprint arXiv:2306.04362*, 2023. 4, 5

[50] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023. 4, 5

[51] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019. 2

[52] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 4

[53] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025. 1

[54] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 4

[55] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023. 2

[56] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems*, 37:59570–59596, 2025. 4, 5

[57] Junhao Zheng, Chenhao Lin, Jiahao Sun, Zhengyu Zhao, Qian Li, and Chao Shen. Physical 3d adversarial attacks against monocular depth estimation in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24452–24461, 2024. 1

[58] Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15044–15054, 2021. 4, 5

[59] 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, 2024. 4