Training-free Detection of Text-to-video Generations via Over-coherence

Jonathan Brokman[1] Oren Rachmil[1] Omer Hofman[1] Roy Betser[1] Amit Giloni[1]
Roman Vainshtein[1] Hisashi Kojima[2]
[1]Fujitsu Research of Europe (FRE) [2]Fujitsu Research
{jonathan.brokman, oren.rachmil, omer.hofman, roy.betser, amit.giloni,
roman.vainshtein, hisashi.kojima}@fujitsu.com

Abstract

Text-to-video generative models have emerged as powerful tools in content creation, capable of synthesizing highly realistic videos from textual prompts. However, the rapid advancement of these models introduces significant security and trust concerns - adversaries can now fabricate convincing videos, while existing detection methods struggle to generalize to unseen generative techniques. This is due to existing approaches' reliance on supervised learning, requiring continuous dataset curation and retraining, which is impractical given the fast-paced evolution of generative models. In this work, we introduce the first training-free detection method for text-to-video generations, eliminating the need for labeled training data or prior exposure to generation techniques. Our approach exploits a fundamental weakness in text-to-video models: Unnatural temporal over-coherence in frame transitions. By leveraging a novel time-coherence detection criterion, our method identifies subtle inconsistencies in video embeddings, capturing temporal artifact patterns that reliably distinguish AI-generated content from real videos. We extensively evaluate our approach on an extended combination of benchmarks, amounting to 23 video generative models and 54.4K diverse videos, demonstrating that it outperforms the existing baselines, and shown to be effective on unseen generative models. This work establishes a new direction for training-free detection of text-to-video generated content, providing a scalable and time resilient solution.

1. Introduction

Text-to-video (T2V) generative models have recently gained significant attention, enabling the synthesis of compelling video content from textual description (prompts) [5, 11, 14, 16, 26, 34, 42, 54, 57, 64, 65]. However, this also introduces new security and trust challenges, as highly realistic generated videos become increasingly difficult to distinguish from authentic content. Major platforms are responding to this challenge - YouTube, for instance, is rolling out policies requiring creators to label any "altered or synthetic" AI content, and will remove or suspend content that isn't properly flagged [12]. Governmental bodies employ similar tactics - The EU's AI Act, for example, requires that any AI-generated or manipulated video resembling real people or events be clearly labeled as artificial in origin [43]. However, currently, these challenges remain unsolved - law enforcement, such as the United States' FBI, warns that criminals are now using synthetic AI-generated footage to create fictitious individuals, fabricate events, and stage fraudulent scenarios in real-time scams [32].

Detecting fully generated video content, such as T2V generations, poses a new challenge, as existing detection technologies primarily focus on deepfake detection, i.e., determining whether a real video has been manipulated [47]. The existing deepfake detection techniques mainly focus altered parts of the frame, distinguishing altered and real pixels [29], including training-free approaches [67]. However, in T2V generation, all pixels are generated, and this distinction does not take place, making the deepfake detection and T2V generation detection tasks inherently different.

Currently, T2V detection strategies rely on supervised approaches that assume the existence of labeled examples of both real and AI-generated videos. However, a significant challenge remains in blind scenarios, where unseen generative techniques must be identified at test time without prior exposure during training. This limitation was recently highlighted in [52], which reported a substantial performance gap when detectors encountered unseen video generative techniques. Such blind detection is critical given the rapid evolution of generative models; without blind detection capabilities, detection teams resort to a high-resource procedure to collect, label and train on new synthetic data coupled with the continual training of supervised methods [23, 44].

In contrast to the current state of T2V generation de-

Official code: https://github.com/FujitsuResearch/training-free-detection-of-text-to-video-generations-via-over-coherence/tree/main

tection, recent AI-generated image detection research has emphasized methods that significantly reduce reliance on large datasets, addressing both the resource-intensive nature of dataset curation and biases toward previously seen generative techniques. These methods include low-data approaches, which aim to minimize the No. of training examples, and training-free approaches, which eliminate the use of labeled datasets entirely [10, 20, 21, 48, 61]. Training-free detectors forgo task-specific training altogether, instead they typically repurpose models pre-trained solely on real data and thereby sidestepping the generator-specific bias that can cause supervised and even low-data methods to fail in blind scenarios. Despite their promise in the image domain, analogous approaches remain unexplored for detecting T2V generated content (Fig. 1).

In this paper, we introduce the first training-free method for text-to-video detection, eliminating the need for costly dataset curation and maintenance. We leverage the inherent additional information in videos compared to images by focusing on the temporal aspect - how frames change over time, thus utilizing cues that static images cannot provide. Our method relies on an inherent property of video generative models: They actively enforce high temporal coherence via architectural design and specialized loss functions [16, 30, 38, 49, 60]. Thus, we hypothesize that unnaturally high temporal coherence is a key indicator of generated text-to-video content and propose to use this as an indicator for the video detection task: We embed frames with a frozen visual encoder (CLIP), compute similarity between consecutive frames denoted $\{s_i\}_i$ and use $\Gamma = \max_i s_{i,i+1}$ to detect "bursts" of over-coheerence and $\gamma = \min_i s_{i,i+1}$ to characterize overall over-coherence of the whole video. Finally - these are combined in an adaptive manner to produce our criterion - S_T. We extensively evaluate our method on a diverse multi-source set of 54.4K videos from the VideoFeedback (including VidProm) and GenVideo datasets [15, 28, 55], consisting of videos generated by various generation techniques, such as Pika, Sora [36, 45]. Additionally, to keep the dataset timely we sourced *all available* .mp4 samples for Hunyuan-Video, Wan-2.1, and LTX-Video [27, 35, 53] from the *Civitai* website - resulting in a diverse and up-to-date collection of community-generated videos.

Contributions.

- **Empirical observations of temporal over-coherence:** By analyzing frame-embedding similarities, we discover that generated videos contain temporal over-coherence, visible as heavy upper tails of consecutive frame similarity distributions (Fig. 5).
- **Training-free detector exploiting over-coherence:** Building on our empirical observations, we devise *the first training-free criterion*, capitalizing on both "highly localized bursts" as well as "global elevation" of temporal over-coherence, providing an effective scalar score for distinguishing real from generated videos.
- **Results:** On a combined 54.4K-video dataset containing data from 22 video generative models, our training-free method improves average AUC by 10.6% over the leading training-free approaches (Table 1), while remaining robust to common video artifacts, namely varying FPS and locally unordered frames.

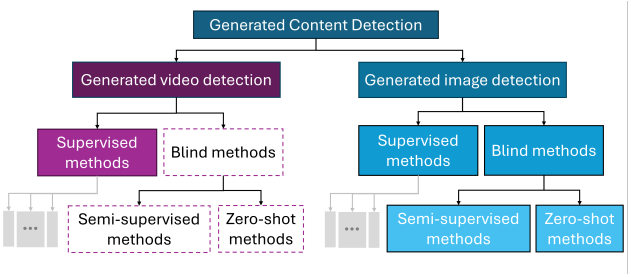

Figure 1. Taxonomy overview for computer vision generated content detection methods, emphasizing the data-usage aspect. Supervised methods are trained on real and generated images/videos. Low-data and training-free regimes emphasize low-dependence on generated datasets, with performance on "unseen" generation techniques in mind (lower reliance for lower bias to "seen" techniques). Despite their merits and reported importance in the image domain, both low-data and training-free detection regimes remain unexplored for videos.

2. Related Work

Detection of T2V generations has become critical with the advent of generative models capable of producing highly realistic synthetic video content. Supervised methods have been proposed, training classifiers on datasets of real and synthetic videos. [40] combines spatial and temporal architectures to fuse both aspects of AI-generated videos. [15] applies a Mamba state-space layer to video embeddings and a small MLP. MLLMs have also been proposed for supervised video detection [25, 59], often with adapters/LoRA. [33] trains a video detector on the geometric trajectory of a signal. However, as emphasized in [52], a gap remains in generalization to the "blind regime", i.e., when encountering videos from unseen generators.

Parallel to detection, a safety line studies policy compliance and evaluation [17, 41, 58, 63], and adversarial methods probes automated video annotation and jailbreak risks [31, 39]. Most of these are supervised or benchmark-oriented (with training-free variants such as [63]); in contrast, training-free *video detection* remains nascent.

Similarly to today's video detection landscape, early text-to-image detectors primarily utilized supervised learning [2, 56], however, here too the effectiveness is signifi-

cantly reduced in the blind regime [23]. In response to this limitation, Ojha et al.[44] proposed a method that captures frequency-domain artifacts from early CNN layers, training on a single generator and evaluating on unseen ones. Their approach showed improved robustness and inspired follow-up methods that exploit model-agnostic features to enhance generalization[13, 18, 37]. While promising, these approaches still rely on large volumes of synthetic data during training—data that is often costly to generate and may require access to proprietary models.

Followingly, low-data methods emerged in the blind settings - with the aim of not only generalizing to unseen techniques, but also reducing the reliance on data from seen techniques. Apart from the obvious resource-efficient advantage, reducing the amount of data may also reduce the bias towards the seen techniques, enhancing generalization to the unseen scenario. In this setting, [21] proposed to combine CLIP embeddings with SVM, setting a new SOTA with a fraction of the data. Several additional works have emerged in this field [46, 61, 62].

Training-free methods have been developed for image detection, pushing low-data detection further by relying only on models pre-trained on real images. AEROBLADE [48] uses the auto-encoder component of latent diffusion models, computing reconstruction errors to distinguish between real and generated images without requiring training. ZED [20] measures image compressibility as their criterion using a compression model trained on real data. The Manifold Induced Biases method [10] analyzes statistical and geometric biases of diffusion manifolds, including a curvature-based criterion later extended for memorization detection [8, 9] and rooted in classical TV-based geometric analysis [6, 7, 22, 24, 50]. A domain-adaptive variant was recently proposed in [4], based on CLIP-whitening [3]. Concurrent to our work, D3 [66] introduced a training-free video detector building on physical intuition, where real videos' second-order time dynamics obey Newtonian laws.

Fourier analysis has a longstanding presence in the field, where it was shown to produce discriminative frequency patterns that highlight differences between generated and real images [1, 19]. In [52], they compared generated images and videos via a Fourier transform of the spatial (frame/image) domain, showing clear differences.

To conclude, there remains a critical gap in both training-free and general low-data methods for video detection, both are especially important for blind scenarios. In this paper, we first elaborate on the low-data gap and then introduce the first training-free method for AI-generated video detection.

3. Motivation for Training-free Detection: Low-data Regimes' Blind Detection Gap

In the field of AI-generated video detection, one would naturally ask: *Could a simple, low-data supervised baseline already provide robust blind detection of AI-generated videos?* If so, the need for a training-free alternative would be weaker. Inspired by the blind-performance analyses in [23, 52] and by the low-data CLIP+SVM study of [21], we construct an AI-generated video detection analogue and measure its performance in the blind settings. In this motivational section, we first clarify what we mean by *blind detection*, then present an experiment designed to probe the limitations of low-data detectors in this setting.

Blind detection. We define *blind detection* as the evaluation of a detection method on generated videos from models that were entirely unseen during training—that is, no synthetic samples from these models were used during training. Real videos are naturally included in both training and evaluation. Training-free methods, which do not rely on any generated data, are blind by construction. While this evaluation setting is sometimes referred to as *zero-shot* [52], the term is inconsistently used across the literature—at, at times referring to training-free approaches [10, 20]. To avoid this ambiguity, we adopt the more specific term *blind*.

Setup. We extract VideoMAE embeddings [51] from real and generated clips. A linear SVM is trained on progressively expanding training sets that include videos from an increasing number of generation models, added in chronological order of public release (T2VZ $\rightarrow \ldots \rightarrow$ Sora). At each step k, we evaluate on a balanced held-out test set drawn from all *unseen* generators $> k$ plus real videos; we also report performance on the *seen* generators that were included in training at step $\leq k$.

Results. Fig. 2 shows a persistent blind gap: accuracy remains high (typically >0.9) on seen generators but degrades sharply on unseen ones (down to ≈ 0.5–0.8 depending on the generator). Adding more generators improves performance *on those generators* without reliably closing the unseen gap. The aggregate seen-unseen deltas (Fig. 2b) remain substantial across training steps.

Overall, we see that even low-data detectors remain biased toward the generators seen during training, where blind accuracy on unseen generators drops sharply. To our knowledge, this is the first systematic demonstration of this low-data blind gap for *video* detection. These results motivate minimizing reliance on generated training data, possibly eliminating it entirely. We therefore propose a *training-free* temporal over-coherence criterion (Sec. 5).

4. Preliminaries

We model a video as a temporal sequence of D-dimensional frame embeddings $\{\mathbf{e}_n\}_{n=0}^{N-1}$ from a frozen encoder Φ, with coordinates $x_n(d) = [\mathbf{e}_n]_d$. The length-$N$ DFT along time:

$$X_k(d) = \sum_{n=0}^{N-1} x_n(d)\, e^{-2\pi i k n/N}, \qquad k = 0, \ldots, N-1. \tag{1}$$

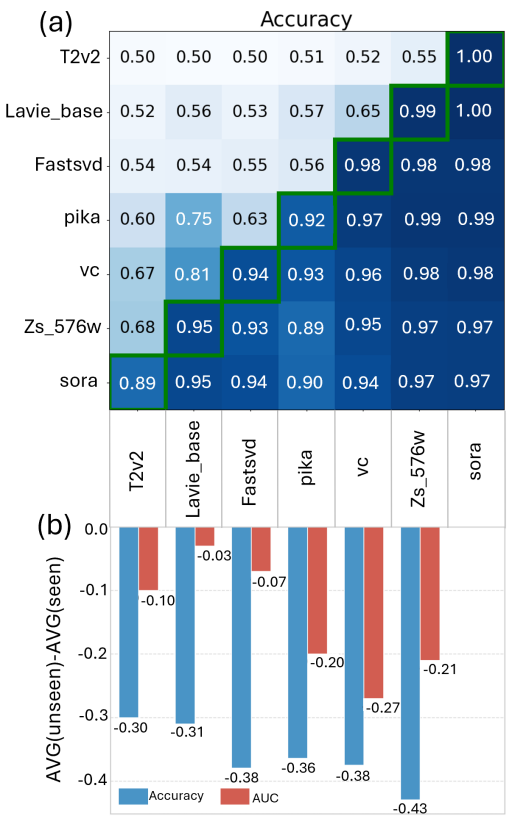

Figure 2. Online SVM training on VideoMAE embeddings in a progressive manner (explained in Sec. 3 and [23]), scheduled by publication time. We report the accuracy heatmap (a) for distinguishing real from generated videos and bar plot (b) of the difference in accuracy and AUC means between seen and unseen cases for each training step. Both axes in the heatmap are chronologically ordered based on the official release dates of video generation techniques. The green counter diagonal separates seen (bottom-right) from unseen cases (top-left), highlighting performance drop on unseen data and the challenges of blind detection.

Key idea. Local "freeze" plateaus in the consecutive-frame similarity signal create sharp entry/exit edges whose first differences act like impulses, injecting broadband (high-frequency) energy in (1). Thus short-lag measurements retain discriminative energy; wide gaps tend to suppress it. Details and proofs are in the Supplement.

4.1. Plateau Fourier Considerations

A localized plateau can be viewed as a value-continuous signal with jumps in the first difference at the plateau boundaries; these behave as two impulses whose DFT spreads energy across frequencies. Sparse ("spacious") pairing of samples filters out much of that energy, while short-lag pairing preserves it. See the Supplement for the complete discrete derivation and extensions.

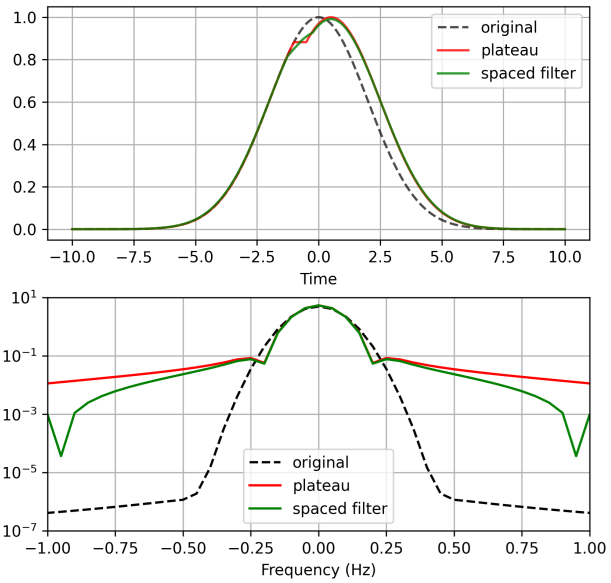

Figure 3. **Toy 1D illustration.** A local plateau is value-smooth but has non-differentiable edges, which spread energy across frequencies. Probing only distant samples blurs these edges; short-lag comparisons retain the signal. Full derivation in the Supplement.

5. Method

Our detector measures *temporal over-coherence* in a training-free manner. Given a video, we embed each frame using a frozen pretrained vision model. We compute cosine similarity between adjacent frame embeddings and use the *maximum* consecutive-pair similarity as a scalar score (see pipeline illustration in Fig 4). As evident by our results, generated videos, whose synthesis pipelines often enforce strong short-range temporal consistency, exhibit unusually high local similarity spikes compared to real footage.

Over-coherence in Generative Models Text-to-video systems enforce temporal smoothness via architectural coupling and explicit regularizers [16, 30, 38, 49, 60]. These mechanisms can overshoot natural dynamics, producing short stretches of near-identical consecutive frames ("temporal over-coherence").

Pairwise Frame Similarity in Embedding Space Given frames $f_1, \ldots, f_T$, we compute embeddings $\mathbf{e}_i = \Phi(f_i)$ and cosine similarities

$$s_{ij} = \frac{\mathbf{e}_i^\top \mathbf{e}_j}{\|\mathbf{e}_i\| \, \|\mathbf{e}_j\|}. \tag{2}$$

We operate on short-lag entries $s_{i,i+1}$ by default (Sec. 5).

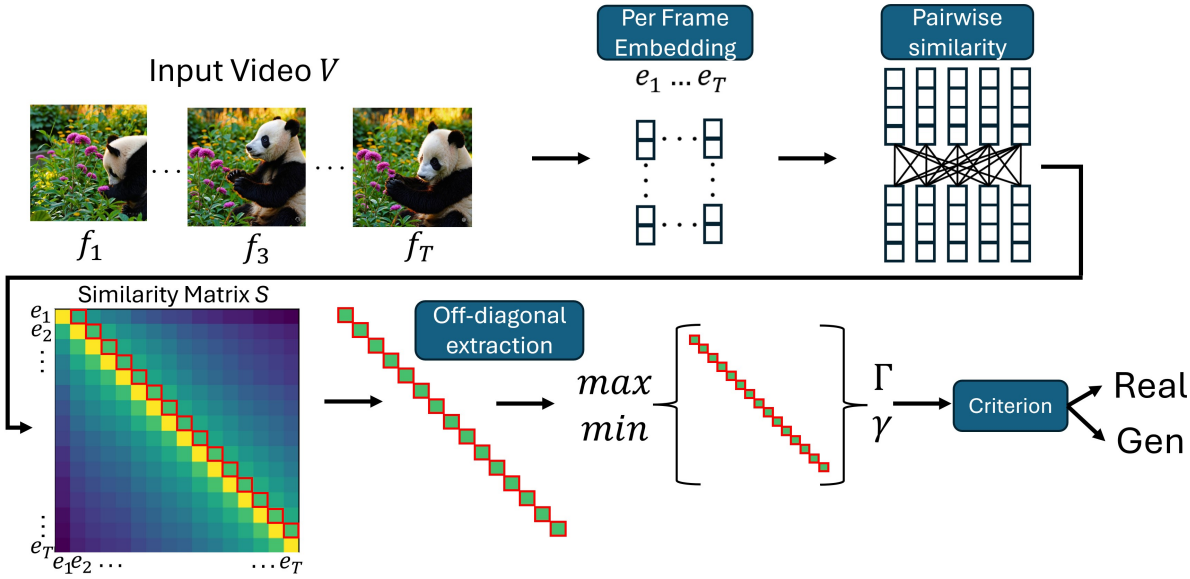

Figure 4. **Pipeline overview.** Frames $f_1, \dots, f_T$ are extracted from a video V and embedded via a frozen pretrained encoder Φ into $\mathbf{e}_i$. Pairwise cosine similarities form a matrix $\mathbf{S}$ (shown in full for intuition; in practice, we only need the first off-diagonal, k=1). Localized high-similarity spikes (temporal over-coherence) are more common in generated than real videos and drive our scalar detection score.

Sub-Diagonals as Sparse ("Spacious") Temporal Pairing Let the gap be $k = j - i$. The k-th sub-diagonal collects

$$\mathcal{D}_k = \{\, s_{i,i+k} \mid 1 \le i \le T - k \,\}. \tag{3}$$

Short-lag k=1 retains plateau-edge energy; larger k increasingly attenuates it. We therefore use k=1 by default and report ablations (Table 2).

Localized and Global Consecutive Over-coherence For k=1, let $\mathcal{D}_1 = \{s_{i,i+1}\}_{i=1}^{T-1}$. We define

$$\Gamma(v) = \max_{1 \le i < T} s_{i,i+1}, \qquad \gamma(v) = \min_{1 \le i < T} s_{i,i+1}. \tag{4}$$

High Γ captures localized spikes; high γ indicates globally elevated short-lag similarity.

Adaptive Criterion S_T We choose Γ when a clear spike exists, otherwise back off to γ:

$$S_T(v) = \begin{cases} \Gamma(v), & \Gamma(v) > T, \\ \gamma(v), & \text{otherwise}, \end{cases} \qquad T{>}0.99. \tag{5}$$

This preserves sensitivity to spikes while remaining robust when motion is uniformly smooth.

Statistical Analyses and Intuition Aggregate similarity matrices decay with temporal gap for both classes, but generated clips decay more slowly and exhibit heavy upper tails at small gaps (Fig. 5). This supports using the per-video extrema Γ (bursts) and γ (global elevation) rather than modeling full distributions; full statistics are in the Supplement.

5.1. DFT-informed Best Practices

Frame rate. Short plateaus span a few frames; 8 fps preserves enough edge energy; lower fps weakens Γ. **Gap.** Comparing k-apart frames acts like a two-tap filter with response $|H_k(\omega)| = 2|\cos(\omega k/2)|$, creating denser notches as k grows. We therefore default to k=1; ablations in Table 2. Derivations in the Supplement.

6. Experiments and Results

6.1. Experimental settings

Datasets and Evaluation Settings. We conduct our experiments using three main sources: 1) VIDEOFEEDBACK dataset [28], which incorporates VIDPROM [55] and extends it with data that they contribute of additional generative techniques - $37.6K$ videos both real and generated; 2) GENVIDEO [15], which incorporates 10 generative models in its test set providing $10K$ videos in total (real and generated). Lastly, we have sourced 1.7K videos from three timely models: LTX-Video, hunyuanvideo and WanVideo-2.1[27, 35, 53] - notably, these are transformer-based. For convenience - we refer these NG (NEW GENERATIONS). In total, we have a versatile and extensive benchmark of 54.4K videos, containing generations from 22 generative models.

For consistency across varying lengths, we restrict each video to 16 frames, equal to the shortest video in the dataset, aside from a negligible amount of 8-frame videos; this ensures that temporal metrics remain comparable even when videos originally differ in duration.

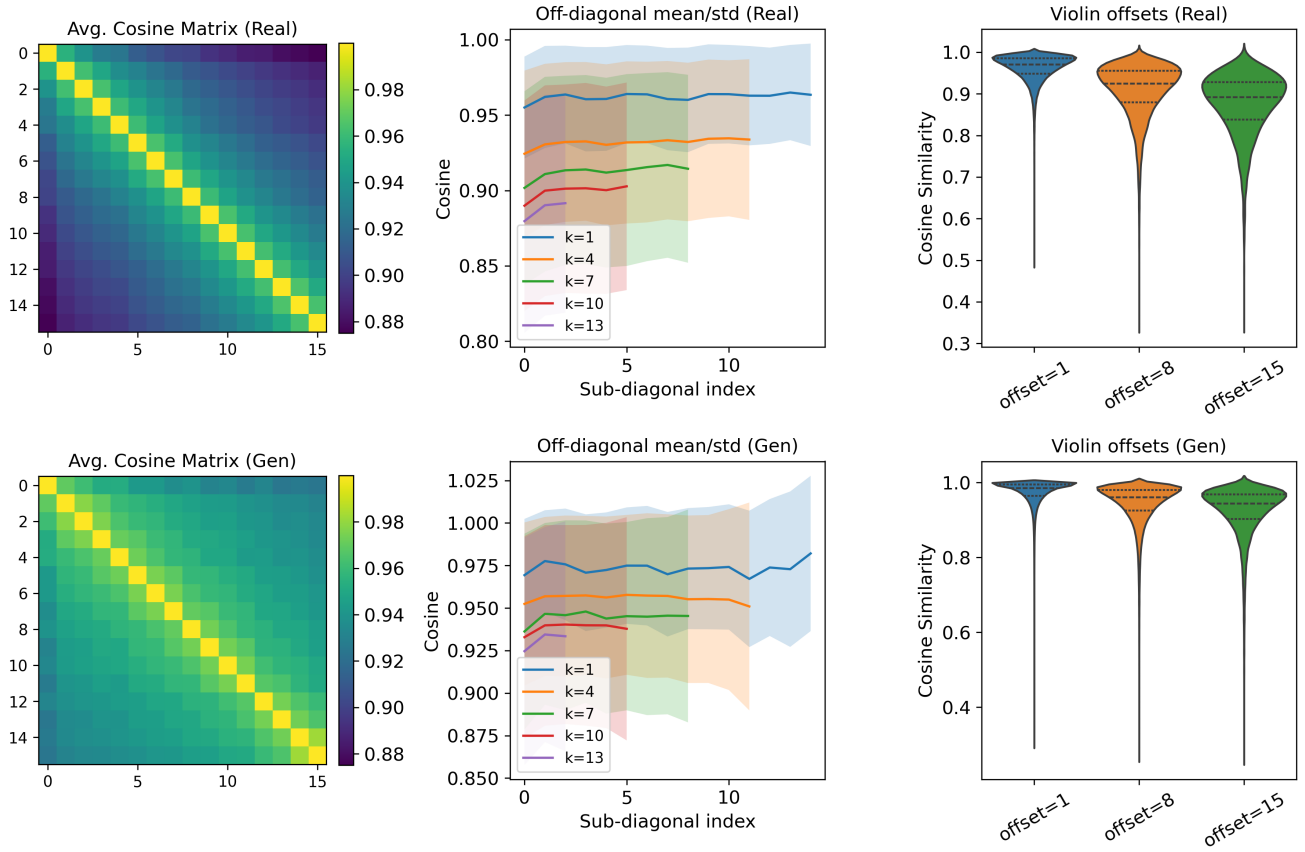

Figure 5. **Statistics of Pairwise frame similarity for real (top row) vs. generated (bottom row) videos.** (Left) The *mean* similarity matrices for real vs. generated videos (subset of our dataset) exhibit a roughly Toeplitz structure: values decline with temporal distance from the main diagonal. Notably, the decline is steeper in real videos, whereas generated videos retain higher similarity further from the diagonal (shallower decay), consistent with excess temporal coherence. However, mean heatmaps alone can hide critical variability. (Middle) For a fine-grained look we examine here the *sub-diagonals* indexed by temporal gap k (Sec. 5): each collects similarities $s_{i,i+k}$ between frames k steps apart. Each plot shows mean±std for a given k; real videos drop off rapidly, while generated ones decay more slowly and with higher variance—evidence of persistent short- to mid-range similarity. (Right) Here we further examine the distributional structure: Violin plots (selected k) reveal heavy upper tails in generated videos, especially at small k, indicating that *some* frame pairs are nearly similar, in other words, we observe over-coherence. Assuming that over-coherence is distributed across the video samples, we adopt a simple video-level score-combining the the maximum consecutive-frame similarity $\Gamma(v) = \max_i s_{i,i+1}$ - which captures high bursts of over-coherence, and $\gamma(v) = \min_i s_{i,i+1}$, that detects whether the whole video overall experiences over-coherence (details in Sec. 5).

Baselines. We benchmark against the three state-of-the-art training-free image detectors: AEROBLADE, ZED, and Manifold Bias. They are run frame-wise and the mean score is taken over all frames of each video. For ZED we use an unofficial implementation due to lack of release code.

6.2. Results

In Table 1 we report AUC for distinguishing fully generated videos from real ones across GenVideo (GV), our timely *New Generations* (NG), and VideoFeedback (VF), each evaluated with two real-video sources (GV real vs. VF real). Our adaptive training-free criterion S_T is the top performer in **7/8** settings: GV (0.851 / 0.796), NG (0.863 / 0.826), and VF with GV-real (0.801). The sole exception is VF with VF-real, where Manifold Bias [10] slightly leads (0.755 vs. 0.727). When aggregated, S_T attains **0.838** (GV total) and **0.783** (VF total), exceeding the next-best baseline by +0.136 and +0.108 AUC, respectively. These results indicate that temporal over-coherence is a robust, training-free cue that holds across datasets and real-source choices, while the existing training-free baselines (AEROBLADE, ZED, Manifold Bias) are less consistent on videos.

Notably, the strongest competitor is Manifold Bias, performs well on the VideoFeedback data, a natural outcome given its reliance on access to the diffusion model architecture. However, as text-to-video methods evolve beyond

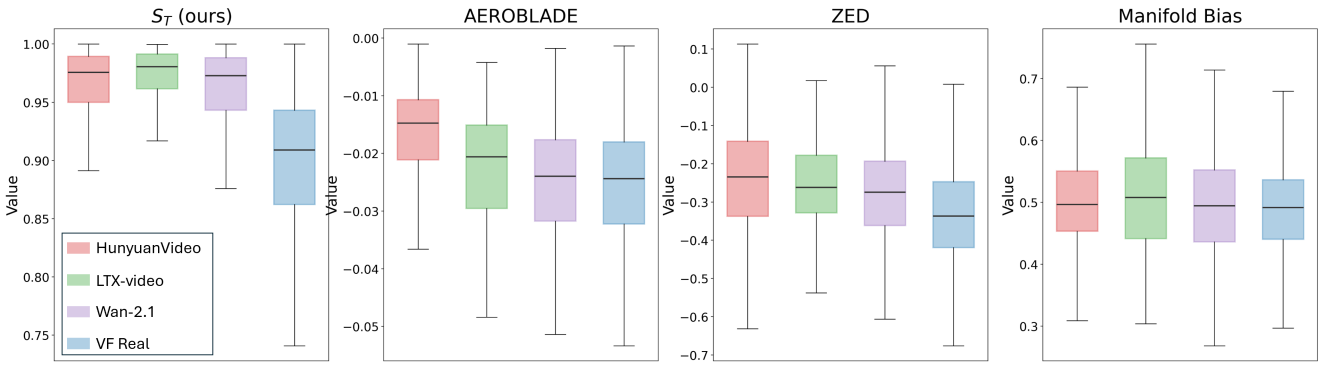

Figure 6. Criterion statistics (boxplots) for real vs. generated videos across up-to-date generators (LTX-Video, hunyuanvideo and WanVideo-2.1[27, 35, 53]) - real data is combined from under both choices (GV-real, VF-real). Boxes show the middle 50% (Q1–Q3; IQR) with the median (Q2) inside; whiskers extend to 1.5×IQR; points denote per-model aggregates. Our S_T yields clear separation between real and generated distributions - observed as medians split with minimal Q1–Q3 overlap. Competing baselines (AEROBLADE, ZED, Manifold Bias) exhibit overlapping quartiles, indicating poor separability on these modern models.

Method	GenVideo (GV)		New Generations (NG)		VideoFeedback (VF)		Totals	
	GV (real)	VF (real)	GV (real)	VF (real)	GV (real)	VF (real)	GV (real)	VF (real)
S_T (ours)	**0.851**	**0.796**	**0.863**	**0.826**	**0.801**	0.727	**0.838**	**0.783**
AEROBLADE	0.630	0.655	0.632	0.619	0.761	0.650	0.674	0.641
ZED	0.698	0.632	0.736	0.661	0.673	0.635	0.702	0.643
Manifold Bias	0.742	0.721	0.575	0.548	0.774	**0.755**	0.697	0.675

Table 1. AUCs for detecting fully generated videos from real videos, In total this covers 22 generative models - see per-model (big table) breakdown in the supplementary. Datasets: **GenVideo (GV)**, **VideoFeedback (VF)** (including sourced videos from VidProm), and *new generations* (NG) - a timely addition we provide, covering HunyuanVideo, WanVideo 2.1, and LTX-Video. Parentheses indicate the source of the real data (GV or VF). Bold indicates the best method per column. S_T is the overall best performer across datasets.

diffusion paradigms, its generalizability may diminish. In contrast, our approach is more model-agnostic - intuitively this is due to reliance on CLIP which is not a generative model. Consequently, our approach is better positioned to remain effective on future generation techniques.

For a per-model extensive breakdown please see the big table in the supplementary - which breaks down performance on all generative models as well as real data sources. Here we further analyze Fig. 6 showing per-model AUC distributions for S_T versus the image-domain training-free baselines. Our method S_T consistently achieves high values (median $\approx$ 0.95–0.99 across HunyuanVideo, LTX-Video, and Wan-2.1), clearly separating generated from real videos. By contrast, AEROBLADE and ZED hover around zero or negative median values, reflecting poor discriminative power when extended to videos. Importantly, the stability of S_T across all three new Transformer-based generators (HunyuanVideo, LTX-Video, Wan-2.1) and VF real clips confirms that temporal over-coherence is a persistent artifact of modern text-to-video systems, whereas image-domain training-free detectors fail to generalize reliably.

Ablation Study. For the ablation, we have a two-dimensional test: First we identify our solution as composed from two criteria - Γ and γ, where there combination is controlled by a hyper-parameter $t = T$. We test the two extremes $t = 0, t = \infty$ corresponding to Γ-only and γ-only use respectively. A second dimension would be the temporal sensitivity of these criteria: We evaluate detection performance across different sub-diagonal offsets k, where similarity is measured between frames k steps apart. As shown in Table 2, for Γ - accuracy peaks at k=1, supporting our hypothesis that over-coherence manifests most strongly between consecutive frames. Higher k values lead to reduced AUC, suggesting that distant frame comparisons dilute the local artifacts characteristic of AI-generated videos. On the other hand - γ is shown to be robust in this regard.

Sensitivity analysis: FPS and Frame-shuffling tests. To probe sensitivity to local frame disorder, we partition each video timeline into non-overlapping windows - Fig. 7. We re-partition for increasing window sizes - from 64_{ms} to 1024_{ms}, and *shuffle frames within every window*. This forces the detector (both Γ and γ) to encounter unordered

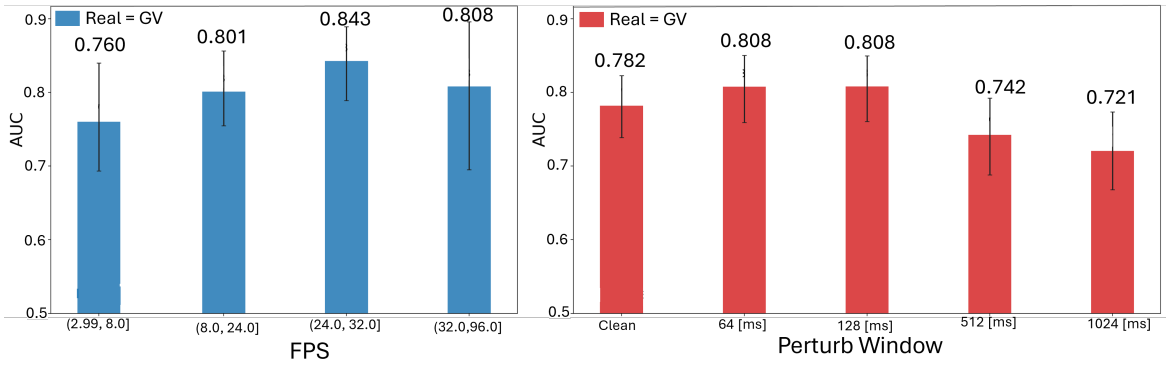

Figure 7. **Sensitivity of the adaptive criterion S_T to FPS and temporal shuffling.** *Left (AUC vs FPS).* Videos are grouped into FPS bins (x-axis). Performance is stable across bins and peaks at higher FPS, where short-lag coherence is better preserved (enabling the Γ component information); robustness at low FPS is retained via the γ component that relies on low-frequency structure. *Right (AUC vs perturb window).* We stress-test S_T upon temporal instability by shuffling frames within fixed windows (clean, 64–1024 ms). Each bin encompasses the whole generative evaluation set. Shuffling is applied in *every* window along time (even harder than realistic scenario which is usually sparse shuffles). S_T degrades only mildly even at 1024 ms, indicating strong robustness. Bars show mean AUC with error bars denoting across-model variability. In both panels, the task is real (VF) vs. all generators.

	Ablation: T Extremes and Sub-diagonal Offset k $S^k_{t,\Delta} := \mathrm{AUC}(S^k_t) - \mathrm{AUC}(S_T)$							
	k=1	3	5	7	9	11	13	15
$S^k_{t=0,\Delta}$	-.06	-.15	-.16	-.14	-.13	-.12	-.12	-.10
$S^k_{t=\infty,\Delta}$	-.01	-.05	-.07	-.06	-.06	-.06	-.06	-.05

Table 2. Ablation over decision threshold T and sub-diagonal gap k. Entries are ΔAUC relative to the full S_T at k=1. By construction, t=0 corresponds to the Γ-only case (localized spikes), and $t\to\infty$ corresponds to the γ-only case (global elevation). k=1): Here we transition to standard Γ, γ. *Horizontal (increasing k):* Γ becomes less robust (larger negative deltas) as short-lag spikes are filtered by wider gaps, while γ remains comparatively stable.

frames. To obtain FPS sensitivity results, we simply slice the data by FPS bins. We observe robustness in both FPS and frame-shuffling tests - see Fig. 7. This is due to the complementary nature of Γ and γ - where the former needs high FPS to be captured, but if not captured - then γ can serve as an effective detector that operates in very low FPS: Since $gamma$ bounds over-coherence from below, this means we can model $gamma$-coherence as a constant added to a non-negative function - and such constants are captured by the DC frequency aligning with the observed robustness.

7. Limitations and Future Directions

We rely on short-range temporal over-coherence, which introduces several caveats. It can yield false positives on naturally static or slide-show–style videos, and on clips recorded at very high frame rates where genuine motion between consecutive frames is minimal. One straightforward mitigation is to fuse our temporal score with a frame-level image detector that targets spatial artifacts: Such hybrid scoring would remain effective when temporal cues are unreliable and could jointly exploit spatial and temporal inconsistencies. Designing principled fusion strategies, remains an open, promising research direction for future research.

8. Conclusions

We have proposed the first training-free method for detecting text-to-video videos. This regime has the advantage of not being overly fitted to certain generative techniques - this is useful when a method is expected to perform well on unseen techniques. By focusing on temporal dynamics—specifically, the phenomenon of local and global over-coherence in time—we exploit a crucial distinction between real and generated videos: The excessive frame-to-frame similarity found in synthetic content. This is rooted in the intuition that generative video techniques tend to over-shoot enforcement of temporal coherence. Our findings highlight that, leveraging temporal cues, namely over-coherence, is highly effective for robust detection across diverse generative models. Experimental results show that our approach substantially outperforms the baselines, improving overall AUC by 10.6%. Moreover, our method is easy to implement, and requires no training data - enhancing its applicability in rapidly evolving generative landscapes where collecting synthetic samples for every new model is infeasible. In future work, we aim to combine spatial–temporal cues to better detect advanced generative models. We hope our efforts will encourage greater focus on training-free video detection and its important role in blind detection.

References

[1] Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023. 3

[2] Samah S Baraheem and Tam V Nguyen. Ai vs. ai: Can ai detect ai-generated images? *Journal of Imaging*, 9(10):199, 2023. 2

[3] Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened clip as a likelihood surrogate of images and captions. In *Forty-second International Conference on Machine Learning*. 3

[4] Roy Betser, Omer Hofman, Roman Vainshtein, and Guy Gilboa. General and domain-specific zero-shot detection of generated images via conditional likelihood, 2025. 3

[5] 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. 1

[6] Jonathan Brokman and Guy Gilboa. Nonlinear spectral processing of shapes via zero-homogeneous flows. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 40–51. Springer, 2021. 3

[7] Jonathan Brokman, Martin Burger, and Guy Gilboa. Spectral total-variation processing of shapes—theory and applications. *ACM transactions on graphics*, 43(2):1–20, 2024. 3

[8] Jonathan Brokman, Itay Gershon, Omer Hofman, Guy Gilboa, and Roman Vainshtein. Tracking memorization geometry throughout the diffusion model generative process. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025. 3

[9] Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Identifying memorization of diffusion models through p-laplace analysis. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 295–307. Springer, 2025. 3

[10] Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 6

[11] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, and Yufei Guo. Video generation models as world simulators. https://openai.com/index/video-generation-models-as-world-simulators/, 2024. Accessed: 2025-03-02. 1

[12] CTV News by The Associated Press. Youtube creators will soon have to disclose use of gen ai in videos or risk suspension. https://www.ctvnews.ca/sci-tech/article/youtube-creators-will-soon-have-to-disclose-use-of-gen-ai-in-videos-or-risk-suspension/, 2023. Accessed: 2025-03-06. 1

[13] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10759–10769, 2024. 3

[14] cerspense. zeroscope: A watermark-free modelscope-based video model. https://huggingface.co/cerspense/zeroscope_v2_576w, 2023. Version 2.0.0, CC BY-NC 4.0 License. 1

[15] 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. 2, 5

[16] 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. 1, 2, 4

[17] Zhaorun Chen, Francesco Pinto, Minzhou Pan, and Bo Li. Safewatch: An efficient safety-policy following video guardrail model with transparent explanations. *arXiv preprint arXiv:2412.06878*, 2024. 2

[18] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12830–12839, 2025. 3

[19] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 3

[20] Davide Cozzolino, Luca Bondi, and Luisa Verdoliva. Zero-shot detection of ai-generated images. In *European Conference on Computer Vision (ECCV)*, 2024. 2, 3

[21] Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4356–4366, 2024. 2, 3

[22] Abderrahim Elmoataz, Olivier Lezoray, and Sébastien Bougleux. Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing. *IEEE transactions on Image Processing*, 17(7):1047–1060, 2008. 3

[23] 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 Workshops*, 2023. 1, 3, 4

[24] Marco Fumero, Michael Möller, and Emanuele Rodolà. Nonlinear spectral geometry processing via the tv transform. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 3

[25] Yifeng Gao, Yifan Ding, Hongyu Su, Juncheng Li, Yunhan Zhao, Lin Luo, Zixing Chen, Li Wang, Xin Wang, Yixu Wang, et al. David-xr1: Detecting ai-generated videos with explainable reasoning. *arXiv preprint arXiv:2506.14827*, 2025. 2

[26] 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. *arXiv preprint arXiv:2307.04725*, 2023. 1

[27] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 2, 5, 7

[28] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2105–2123, 2024. 2, 5

[29] Arash Heidari, Nima Jafari Navimipour, Hasan Dag, and Mehmet Unal. Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520, 2024. 1

[30] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2, 4

[31] Wenbo Hu, Shishen Gu, Youze Wang, and Richang Hong. Videojail: Exploiting video-modality vulnerabilities for jailbreak attacks on multimodal large language models. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025. 2

[32] FBI ic3. Fbi: Criminals use generative artificial intelligence to facilitate financial fraud. https://www.ic3.gov/PSA/2024/PSA241203, 2024. Accessed: 2025-03-06. 1

[33] Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. Ai-generated video detection via perceptual straightening. *arXiv preprint arXiv:2507.00583*, 2025. 2

[34] 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. 1

[35] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2, 5, 7

[36] Pika Labs. Pika: Ai video generation platform, 2024. Accessed: 2025-03-06. 2

[37] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. *arXiv preprint arXiv:2408.06741*, 2024. 3

[38] Jingyun Liang, Yuchen Fan, Kai Zhang, Radu Timofte, Luc Van Gool, and Rakesh Ranjan. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, pages 56–74. Springer, 2024. 2, 4

[39] Haitong Liu, Kuofeng Gao, Yang Bai, Jinmin Li, Jinxiao Shan, Tao Dai, and Shu-Tao Xia. Protecting your video content: Disrupting automated video-based llm annotations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24056–24065, 2025. 2

[40] 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

[41] Yibo Miao, Yifan Zhu, Lijia Yu, Jun Zhu, Xiao-Shan Gao, and Yinpeng Dong. T2vsafetybench: Evaluating the safety of text-to-video generative models. *Advances in Neural Information Processing Systems*, 37:63858–63872, 2024. 2

[42] John Mullan, Duncan Crawbuck, and Aakash Sastry. Hotshot-xl: Personalized gif generation with diffusion models. https://github.com/hotshotco/hotshot-xl, 2023. Version 1.0.0, Apache License 2.0. 1

[43] European AI Office. Eu ai act. https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, 2024. Accessed: 2025-03-06. 1

[44] 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. 1, 3

[45] OpenAI. Sora: Video generation models as world simulators, 2024. Accessed: 2025-03-06. 2

[46] Ziyang Ou. Clip embeddings for ai-generated image detection: A few-shot study with lightweight classifier. *arXiv preprint arXiv:2505.10664*, 2025. 3

[47] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024. 1

[48] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3

[49] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2, 4

[50] Nir Sochen, Ron Kimmel, and Ravi Malladi. A general framework for low level vision. *IEEE transactions on image processing*, 7(3):310–318, 1998. 3

[51] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[52] Sina Vahdati, Amirhossein Habibian, Xinyu Tang, Luigi Celona, M. Alex O. Vasilescu, Wael AbdAlmageed, Yury

Wang, Nicu Sebe, Liming Chen, and Ioannis Patras. Beyond fake images: Detecting ai-generated videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. 1, 2, 3

[53] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 5, 7

[54] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 1

[55] 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. 2, 5

[56] Yuezun Wang, Shengming Wang, Jing Dong, and Tieniu Tan. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8695–8704, 2020. 2

[57] 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. *International Journal of Computer Vision*, pages 1–20, 2024. 1

[58] Yixu Wang, Jiaxin Song, Yifeng Gao, Xin Wang, Yang Yao, Yan Teng, Xingjun Ma, Yingchun Wang, and Yu-Gang Jiang. Safevid: Toward safety aligned video large multimodal models. *arXiv preprint arXiv:2505.11926*, 2025. 2

[59] 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. *arXiv preprint arXiv:2505.12620*, 2025. 2

[60] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2, 4

[61] Shiyu Wu, Jing Liu, Jing Li, and Yequan Wang. Few-shot learner generalizes across ai-generated image detection. *arXiv e-prints*, pages arXiv–2501, 2025. 2, 3

[62] Juncong Xu, Yang Yang, Han Fang, and Honggu Liu. Famsec: A few-shot-sample-based general ai-generated image detection method. *arXiv preprint arXiv:2410.13156*, 2024. 3

[63] Jaehong Yoon, Shoubin Yu, Vaidehi Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. *arXiv preprint arXiv:2410.12761*, 2024. 2

[64] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025. 1

[65] Yuan Zhang, Jiacheng Jiang, Guoqing Ma, Zhiying Lu, Haoyang Huang, Jianlong Yuan, and Nan Duan. Generative pre-trained autoregressive diffusion transformer. *arXiv preprint arXiv:2505.07344*, 2025. 1

[66] Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free ai-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12852–12862, 2025. 3

[67] Yueying Zou, Peipei Li, Zekun Li, Huaibo Huang, Xing Cui, Xuannan Liu, Chenghanyu Zhang, and Ran He. Survey on ai-generated media detection: From non-mllm to mllm. *arXiv preprint arXiv:2502.05240*, 2025. 1