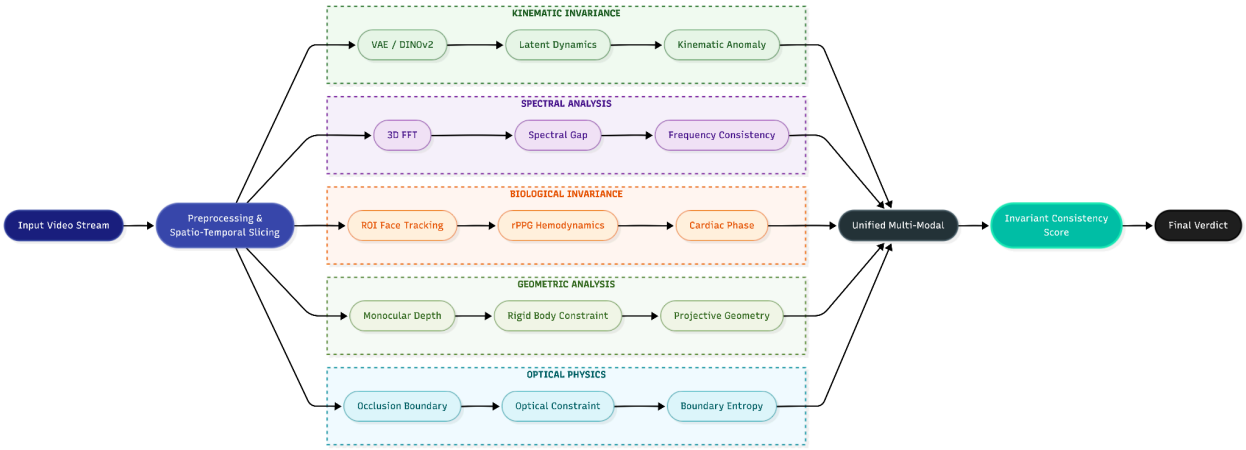
Bio-Physical Inconsistency Detection in Diffusion-Generated Videos

Abstract

Recent advances in diffusion-based video generation models have produced outputs that are visually indistinguishable from real footage to human observers. However, these models prioritize perceptual plausibility over adherence to the physical and biological laws governing the real world. In this paper, we propose a theoretical framework for detecting state-of-the-art synthetic videos by identifying violations of bio-physical invariants. We introduce five complementary detection principles: (i) Latent Trajectory Incoherence, (ii) Spatio-Temporal Spectral Gaps, (iii) Biometric Volumetric Inconsistency via remote photoplethysmography (rPPG), (iv) Projective Geometry Violations under Rigid Body Constraints, and (v) Occlusion Boundary Entropy. Rather than framing detection as an image classification task, we argue that deepfake detection must be reformulated as a computational physics problem, where the objective is to identify inconsistencies in the underlying generative universe rather than surface-level artifacts.



1. Introduction

Diffusion-based video generation models, including Diffusion Transformers (DiT), have demonstrated unprecedented ability to synthesize temporally coherent, high-fidelity videos conditioned on text, pose, or motion guidance. Models such as Kling and Higgsfield represent a paradigm shift from frame-based generation toward holistic spatio-temporal modeling. Despite this progress, these systems remain probabilistic predictors trained to approximate appearance distributions, not simulators of the physical or biological world.

Current deepfake detection methods largely rely on supervised classifiers trained on visual artifacts, texture inconsistencies, or frequency-domain fingerprints. As generative models improve, such artifacts rapidly disappear, rendering detector–generator arms races ineffective. We posit that a more fundamental asymmetry exists: while appearance can be learned statistically, the laws of physics and biology impose hard constraints that are difficult to approximate implicitly.

This paper presents a unified theoretical framework for detecting synthetic videos by identifying violations of these constraints. We shift the detection objective from recognizing "fake people" to identifying "broken universes"—generated realities that subtly but measurably deviate from physical law.

2. Background and Related Work

2.1 Diffusion-Based Video Generation

Diffusion models generate data by iteratively denoising samples from a learned distribution. Video diffusion models extend this framework to the temporal domain, often conditioning on pose skeletons, optical flow, or keyframes. Motion in these systems emerges from the denoising dynamics rather than explicit physical simulation.

2.2 Deepfake Detection

Traditional detection approaches include spatial artifact analysis, temporal inconsistency detection, and frequency-domain signatures. Recent work explores physiological cues such as eye blinking or heart rate estimation. However, most methods treat these cues independently rather than as manifestations of deeper physical processes.

2.3 Physics-Based Invariants

Physical systems are governed by conservation laws, rigid-body constraints, and biomechanical dynamics. Violations of these invariants have been used in computer vision for anomaly detection but remain underexplored in generative media analysis.

3. Theory of Latent Trajectory Incoherence

3.1 Hypothesis

We hypothesize the existence of a fundamental phase drift between motion guidance signals and the generated latent manifold in diffusion-based video models.

3.2 Latent Dynamics Formulation

Let x_t denote a video frame at time t, and let $z_t = E(x_t)$ be its latent representation obtained via a pretrained encoder (e.g., VAE or DINOv2). We define latent velocity and acceleration as:

$$v_t = z_{t+1} - z_t, \quad a_t = v_{t+1} - v_t$$

In real videos, a_t is constrained by deterministic physical motion. In synthetic videos, we expect increased variance and micro-stutter due to stochastic denoising dynamics.

3.3 Detection Metric

We measure the variance of latent acceleration over time. Elevated high-frequency fluctuations indicate latent trajectory incoherence.

4. Spatio-Temporal Frequency Analysis: The Spectral Gap

4.1 Biological Motion as 1/f Noise

Human motion exhibits pink noise characteristics due to neuromuscular control systems. Camera sensors further impose band-limited frequency responses.

4.2 3D Fourier Analysis

We perform a 3D Fast Fourier Transform over the video volume (X, Y, T). Synthetic videos exhibit spectral spikes in the temporal dimension corresponding to diffusion sampling steps or interpolation artifacts.

4.3 Cyclostationary Anomalies

These anomalies manifest as unnatural regularity in high-frequency temporal bands, invisible perceptually but statistically distinct from real motion.

5. Biometric Volumetric Analysis via rPPG

5.1 Sub-Surface Scattering and Hemodynamics

Human skin is a translucent volume. Blood volume changes modulate light absorption according to the Beer–Lambert law, producing a measurable Blood Volume Pulse (BVP).

5.2 Spatial Phase Consistency

In real subjects, the cardiac pulse propagates across facial regions with measurable phase delays. Generative models produce either spatially uniform or incoherent pseudo-pulse signals.

5.3 Detection Signal

We extract rPPG signals from multiple facial regions and analyze inter-region phase coherence. Deviations from physiological norms indicate synthetic generation.

6. Projective Geometry and Rigid Body Constraints

6.1 Epipolar Geometry

Under rigid motion, 2D projections of 3D points must satisfy the fundamental matrix constraint.

6.2 Monocular Depth Consistency

Using monocular depth estimation, we reconstruct frame-wise 3D geometry. In synthetic videos, depth maps of facial features often drift or deform during rotations.

6.3 Reprojection Error

We compute the reprojection error of tracked 3D landmarks. Errors exceeding known limits of human skin elasticity signal violations of rigid-body constraints.

7. The Physics of Occlusion: Boundary Entropy

7.1 Occlusion as a Stress Test

Occlusions represent adversarial scenarios for generative models due to layer disentanglement challenges.

7.2 Point Spread Function Constraints

In real cameras, occlusion boundaries are governed by the optical PSF. Synthetic videos often exhibit pixel halos or temporal smearing.

7.3 Entropy Gradient Analysis

We measure spatial entropy gradients along moving boundaries. Persistent boundary memory indicates attention-based leakage rather than physical occlusion.

8. Unified Detection Framework

We propose combining the above signals into an invariant consistency score spanning biological, kinematic, geometric, and optical domains. Unlike supervised classifiers, this framework generalizes across architectures by targeting fundamental constraints.

9. Limitations and Future Work

Future generative models may explicitly incorporate differentiable physics, hemodynamic simulation, or rigid-body solvers, reducing detectable violations. However, such integration introduces significant computational cost and training complexity. Future work includes empirical validation across datasets and adversarial robustness analysis.

10. Conclusion

We argue that next-generation deepfake detection must transcend appearance-based analysis and embrace computational physics. Diffusion-based video models generate convincing surfaces but lack an internalized model of the biological and physical world. By identifying violations of invariant laws, we can detect synthetic media not by what it looks like, but by how its universe breaks.

You are not looking for a fake person; you are looking for a broken universe.