Physics-Driven Spatiotemporal Modeling for AI-Generated Video Detection

Shuhai Zhang[1 4*] **Zihao Lian**[1 *] **Jiahao Yang**[1] **Daiyuan Li**[1] **Guoxuan Pang**[2]
Feng Liu[5] **Bo Han**[7] **Shutao Li**[6†] **Mingkui Tan**[1 3†]
[1]South China University of Technology, [2]University of Science and Technology of China
[3]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, [4]Pazhou Lab
[5]University of Melbourne, [6]Hunan University, [7]Hong Kong Baptist University

Abstract

AI-generated videos have achieved near-perfect visual realism (*e.g.*, Sora), urgently necessitating reliable detection mechanisms. However, detecting such videos faces significant challenges in modeling high-dimensional spatiotemporal dynamics and identifying subtle anomalies that violate physical laws. In this paper, we propose a physics-driven AI-generated video detection paradigm based on probability flow conservation principles. Specifically, we propose a statistic called *Normalized Spatiotemporal Gradient* (NSG), which quantifies the ratio of spatial probability gradients to temporal density changes, explicitly capturing deviations from natural video dynamics. Leveraging pre-trained diffusion models, we develop an NSG estimator through spatial gradients approximation and motion-aware temporal modeling without complex motion decomposition while preserving physical constraints. Building on this, we propose an NSG-based video detection method (NSG-VD) that computes the *Maximum Mean Discrepancy* (MMD) between NSG features of the test and real videos as a detection metric. Last, we derive an upper bound of NSG feature distances between real and generated videos, proving that generated videos exhibit amplified discrepancies due to distributional shifts. Extensive experiments confirm that NSG-VD outperforms state-of-the-art baselines by 16.00% in Recall and 10.75% in F1-Score, validating the superior performance of NSG-VD. The source code is available at https://github.com/ZSHsh98/NSG-VD.

1 Introduction

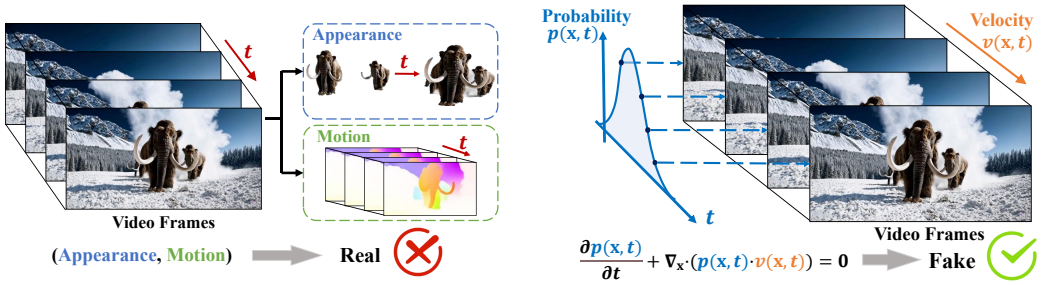
The rapid advancement of generative models [1, 2, 3, 4, 5, 6], particularly diffusion-based frameworks (*e.g.*, Sora [3]), has achieved unprecedented capabilities in synthesizing photorealistic video content. While these breakthroughs enable transformative applications in content creation for creative industries [7, 8, 9, 10], they simultaneously pose critical societal risks through malicious manipulation (*e.g.*, deepfake disinformation [11, 12, 13, 14, 15, 16, 17], synthetic media fraud [7, 18]). As AI-generated videos become increasingly realistic in both spatial and temporal domains, developing effective video detection methods becomes critically urgent for preserving societal trust in digital media.

A fundamental challenge for AI-generated video detection lies in modeling the spatiotemporal dynamics of video evolution. Intuitively, natural videos inherently obey physical laws like motion coherence and texture continuity [19, 20], while AI-generated videos often exhibit subtle yet systematic inconsistencies in spatiotemporal coherence [21]. This observation raises a crucial question:

How can we model intrinsic spatiotemporal dynamics of natural videos to expose synthetic anomalies?

*Equal contribution. Email: shuhaizhangshz@gmail.com, lianzihaolzh@gmail.com
†Corresponding author. Email: mingkuitan@scut.edu.cn, shutao_li@hnu.edu.cn

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Traditional Spatiotemporal Modeling (b) Physics-Driven Spatiotemporal Modeling (Ours)

Figure 1: Comparisons of traditional and physics-driven paradigms for spatiotemporal modeling in AI-generated video detection. (a) Traditional methods [22, 23, 24] often rely on specific artifacts like appearance consistency and optical flow-based motion modeling, struggling with highly realistic content yet physically implausible (*e.g.*, Sora). (b) Our physics-driven approach explicitly models video dynamics via physics conservation laws, effectively identifying violations of physical laws.

Two critical difficulties arise: 1) Video content inherently contains complex *spatial domain correlations* (*e.g.*, texture structure) and *temporal domain dependencies* (*e.g.*, motion trajectories), requiring modeling frameworks that jointly capture both spatial structures and temporal dynamics characteristics. 2) AI-generated videos are rapidly approaching the perceptual quality of natural videos, with discrepancies that may become vanishingly *subtle* in both visual appearance and temporal evolution.

Existing AI-generated video detection methods primarily rely on local feature inconsistencies (*e.g.*, optical flow-based motion modeling [22], appearance consistency modeling [23]) or supervised learning with large-scale datasets [25, 24, 26, 27]. However, they often ignore physics-driven constraints governing spatiotemporal evolution inherent to natural videos. This limitation exhibits inherent vulnerabilities when confronting synthetic anomalies that violate physical laws, *e.g.*, non-physical motion patterns in Sora-generated videos [3] (Figure 1-a), leading to inferior performance.

In this paper, we propose a physics-driven paradigm based on probability flow conservation principles [28, 29]. By modeling video dynamics as fluid mechanics, we formulate video evolution through a probability flow velocity field governed by continuity equations (see Figure 1-b and Section 3.1). This reveals a key insight: *natural video dynamics preserve the product between the velocity field and the ratio of spatial probability gradients to temporal density changes*. Inspired by this, we introduce a **Normalized Spatiotemporal Gradient (NSG)** statistic, which quantifies the ratio of spatial probability gradients to temporal density changes. NSG captures fundamental discrepancies in how videos adhere to physical constraints while eliminating reliance on specific artifacts, enabling sensitive detection even when visual differences are imperceptible to humans or conventional models.

To enable practical estimation, we develop an NSG estimator leveraging pre-trained diffusion models' inherent gradient estimation ability [30, 31] in Section 3.2. By approximating spatial gradients with learned score functions (*i.e.*, the gradient of the log probability density) from the diffusion models and temporal derivatives through motion-aware temporal dynamics via a brightness constancy constraint [32], our method avoids explicit flow computation while preserving essential physical constraints. This estimator eliminates reliance on complex motion modeling by physics-inspired priors while maintaining sensitivity to subtle spatiotemporal inconsistencies inherent to synthetic content.

Building on this foundation, we propose an **NSG-based video detection method (NSG-VD)** in Section 3.3, which computes the *Maximum Mean Discrepancy* (MMD) [33, 34] between NSG features of real videos and the test video as the detection metric, as illustrated in Figure 2. We further theoretically derive an upper bound of the distance between NSG features of real and generated data in Section 3.4, showing this bound expands with increasing distribution shifts in generated videos. This implies that the MMD between NSGs of real videos tends to be smaller than that between real and generated videos, establishing the theoretical basis for the effectiveness of NSG-VD. Extensive experiments show that NSG-VD achieves 16.00% higher Recall and 10.75% higher F1-score than baselines, validating the superior performance of NSG-VD. Our contributions are summarized as:

- A physics-driven NSG statistic: We formulate the video evolution through a probability flow velocity field with a continuity equation and propose a novel statistic Normalized Spatiotemporal Gradient (NSG) that explicitly models spatiotemporal dynamics of videos. By quantifying the ratio

of spatial probability gradients to temporal density changes, NSG fundamentally captures violations of physical continuity in AI-generated videos without reliance on artifact-specific supervision.

- A diffusion-guided NSG estimation with physical priors: We develop an NSG estimator by spatial gradients approximation and motion-aware temporal dynamics modeling using pre-trained diffusion models. By avoiding explicit flow modeling and instead enforcing brightness constancy constraints, our method achieves effective NSG approximation without domain-specific motion modeling.
- An AI-generated video detection method with theoretical and empirical justifications: We propose an NSG-based video detection method (NSG-VD), which quantifies distributional shifts in NSG features using *Maximum Mean Discrepancy* (MMD). We derive an upper bound of NSG feature distances between real and generated videos, proving that generated videos exhibit amplified discrepancies under distribution shifts. Empirical results also show the superiority of our NSG-VD.

2 Related Work

AI-Generated Video Detection. Early generated video detection methods primarily focus on identifying synthetic facial videos. Yang et al. [35] and Amerini et al. [22] exploit auxiliary facial motion cues (landmark dynamics vs. optical flow) for deepfake detection. Gu et al. [36] separately model spatial and temporal inconsistencies, and introduce a vertical slicing feature fusion mechanism to establish a more comprehensive spatial-temporal representation. Wang et al. [23] propose an alternating-freezing strategy with spatiotemporal augmentation for facial consistency modeling. Xu et al. [24] transform consecutive frames into a predefined layout via masking/resizing to enable efficient spatiotemporal modeling. Peng et al. [37] integrate multi-feature fusion of facial perspectives, textures, and attributes. While most methods utilize facial priors, their reliance on domain-specific features limits their generalizability to more general AI-generated content detection.

With the rapid advancement in video generation, detecting general AI-generated content has become challenging. Bai et al. [38] fuse frame-level and optical flow predictions to detect spatial-temporal anomalies. To jointly capture spatiotemporal cues, Ma et al. [39] and Chen et al. [25] propose Transformer- and mamba-based frameworks to model spatiotemporal relationships in video frame features for detection. Song et al. [40] exploit the cross-modal perception and reasoning in vision-language large models to learn general forgery features. Despite this progress, these methods mainly focus on appearance inconsistencies, while overlooking the intrinsic spatiotemporal dynamics cues, thereby struggling to tackle visual cues from diverse video generative models.

Diffusion Models. Diffusion models [41, 30, 42, 31] have emerged as powerful probabilistic generative models, benefiting from their diffusion-denoising paradigm that perturb data into noise through Gaussian processes and reconstruct samples via iterative denoising. Intuitively, the high-quality and diverse generative capabilities of diffusion models come from their ability to capture and exploit the distributional characteristics of natural data, enabling effective discrimination between natural samples and outliers. Motivated by this, a growing body of research has leveraged diffusion models for the detection of adversarial [43, 44, 45] and generated samples [46, 47, 48], wherein the score model emerges as a powerful discriminative tool. Nevertheless, it remains challenging to simultaneously capture and integrate spatiotemporal features when relying solely on score models.

3 Modeling Spatiotemporal Dynamics for AI-Generated Video Detection

AI-Generated Video Detection. Let $\mathbb{P}$ be a Borel probability measure on a separable spatiotemporal metric space $\mathcal{X}\subset\mathbb{R}^{T\times d}$, where T is the number of frames and d is the spatial dimension. Given independent and identically distributed (i.i.d.) samples $S_{\mathbb{P}}=\{\mathbf{x}^{(i)}\}_{i=1}^{n}$ from the real video distribution $\mathbb{P}$, we aim to determine whether each sample $\mathbf{y}^{(j)}$ in $S_{\mathbb{Q}}=\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^{m}$ originates from $\mathbb{P}$.

Challenges for Video Detection. The complex spatiotemporal dynamics in high-dimensional video data often require modeling both spatial irregularities (*e.g.*, unnatural textures) and temporal inconsistencies (*e.g.*, implausible motions). Moreover, the diversity of generative paradigms (*e.g.*, diffusion models [1] and generative adversarial networks [49]) introduces heterogeneous distribution shifts that exhibit as subtle statistical inconsistencies rather than explicit artifacts. These challenges are further worsened by rapidly evolving generative techniques (*e.g.*, Sora [3]), which continuously produce novel spatiotemporal patterns surpassing existing detection mechanisms.

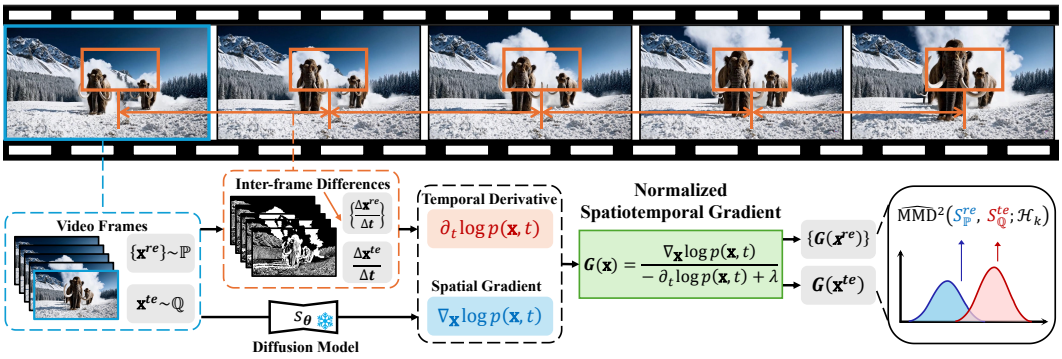


Figure 2: Overview of the proposed NSG-VD. Given a reference set of real videos $\{\mathbf{x}^{re}\}$ and a test video $\mathbf{x}^{te}$, we estimate their spatial gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ and temporal derivatives $\partial_t \log p(\mathbf{x}, t)$ via a pre-trained diffusion model s_θ, from which we derive their Normalized Spatiotemporal Gradients (NSGs) and calculate the MMD between NSG features of real and test videos as a detection metric.

Method Overview. To address these challenges, we propose a physics-driven method based on physical conservation principles to model *spatiotemporal dynamics* and introduce a novel statistic **Normalized Spatiotemporal Gradient (NSG)**, which quantifies the ratio of spatial probability gradients to temporal density changes, capturing subtle anomalies in videos (Section 3.1). Leveraging diffusion models, we develop an effective NSG estimator by spatial gradients approximation and motion-aware temporal dynamics modeling (Section 3.2). Building on this, we develop an **NSG-based video detection method (NSG-VD)**, which computes the *Maximum Mean Discrepancy* (MMD) between NSG features of the test video and real videos as a detection characteristic (Section 3.3), where its framework is shown in Figure 2. Last, we theoretically show that the MMD between NSGs of real videos tends to be smaller than that between real and generated videos (Section 3.4).

3.1 Modeling Spatiotemporal Dynamics via Normalized Spatiotemporal Gradient

Detecting AI-generated videos requires capturing both spatial irregularities and temporal inconsistencies in synthetic content. Inspired by conservation laws in physics (*e.g.*, mass or energy transport), we initially formulate the probability flow velocity field $\mathbf{v}(\mathbf{x}, t)$ to model the evolution of probability density $p(\mathbf{x}, t)$, which satisfies a continuity equation for global consistency across spatiotemporal domains. However, solving $\mathbf{v}$ faces challenges due to its underdetermined nature (Eqn. (4)). To address this, we propose **Normalized Spatiotemporal Gradient (NSG)** $\mathbf{g}(\mathbf{x}, t)$, a *dual* field statistic of $\mathbf{v}$ combining both spatial gradients and temporal dynamics of $p(\mathbf{x}, t)$, as defined in Eqn. (6).

Probability Flow Velocity Field $\mathbf{v}(\mathbf{x}, t)$. We begin to conceptualize the *probability flow* (also called *probability current*) as the movement of probability mass of $\mathbf{x}$ over time t [50, 51]. To formalize this flow, we define the *probability flow density* $\mathbf{J}(\mathbf{x}, t)$, analogous to fluid mechanics [52]:

$$\mathbf{J}(\mathbf{x}, t) = p(\mathbf{x}, t) \cdot \mathbf{v}(\mathbf{x}, t), \tag{1}$$

where $p(\mathbf{x}, t)$ denotes the probability density and $\mathbf{v}(\mathbf{x}, t)$ represents the velocity field guiding the flow of probability mass. The conservation of probability mass [28, 29] implies the continuity equation:

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} + \nabla_{\mathbf{x}} \cdot \mathbf{J}(\mathbf{x}, t) = 0, \tag{2}$$

where $\nabla_{\mathbf{x}} \cdot \mathbf{J} = \sum_i \frac{\partial \mathbf{J}_i}{\partial \mathbf{x}_i}$ denotes the divergence of the vector field $\mathbf{J}$ [51]. This is not a video-specific assumption but a universal mathematical formulation of probability mass conservation, which holds for any time-evolving probability density $p(\mathbf{x}, t)$ [53, 29]. Intuitively, this equation shows that the rate of change in probability density $\partial_t p$ at a point equals the difference between *inflow* (negative divergence) or *outflow* (positive divergence) of the probability flow $\mathbf{J}$. Substituting $\mathbf{J}(\mathbf{x}, t)$ into Eqn. (2), dividing by $p(\mathbf{x}, t)$, and applying the chain rule to $\log p(\mathbf{x}, t)$, yields:

$$\partial_t \log p(\mathbf{x}, t) + \nabla_{\mathbf{x}} \cdot \mathbf{v}(\mathbf{x}, t) + \mathbf{v}(\mathbf{x}, t) \cdot \nabla_{\mathbf{x}} \log p(\mathbf{x}, t) = 0. \tag{3}$$

This expression reveals how the velocity field $\mathbf{v}(\mathbf{x}, t)$ simultaneously encodes temporal evolution ($\partial_t \log p(\mathbf{x}, t)$) and spatial gradients ($\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$) of the probability distribution.

Normalized Spatiotemporal Gradient $\mathbf{g}(\mathbf{x}, t)$ as Dual Field of $\mathbf{v}(\mathbf{x}, t)$. To solve $\mathbf{v}(\mathbf{x}, t)$, we focus on the dominant components of Eqn. (3). Assuming that the divergence term $\nabla_{\mathbf{x}} \cdot \mathbf{v}$ is subdominant in smoothly varying distributions (*e.g.*, incompressible flow approximations [28, 54]), a condition commonly used in fluid dynamics [28] and quantum mechanics [55], Eqn. (3) simplifies to:

$$\mathbf{v}(\mathbf{x}, t) \cdot \nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \approx -\partial_t \log p(\mathbf{x}, t). \tag{4}$$

Considering the non-uniqueness of solutions to $\mathbf{v}(\mathbf{x}, t)$ in Eqn. (4), we normalize both sides into

$$\mathbf{v}(\mathbf{x}, t) \cdot \frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)}{-\partial_t \log p(\mathbf{x}, t)} \approx 1. \tag{5}$$

Definition 1. *(**Normalized Spatiotemporal Gradient (NSG).**) The relation in Eqn. (5) reveals that natural video dynamics preserve the product between the velocity field and the ratio of spatial probability gradients to temporal density changes. We formalize this constrained ratio as the Normalized Spatiotemporal Gradient (NSG), defined as:*

$$\mathbf{g}(\mathbf{x}, t) = \frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)}{-\partial_t \log p(\mathbf{x}, t) + \lambda}. \tag{6}$$

Here, $\lambda > 0$ prevents numerical instability. Eqn. (5) and (6) imply that $\mathbf{g}(\mathbf{x}, t)$ acts as a *dual field* to $\mathbf{v}(\mathbf{x}, t)$, satisfying $\mathbf{v} \cdot \mathbf{g} \approx 1$. The formulation of $\mathbf{g}(\mathbf{x}, t)$ bypasses the ill-posed velocity $\mathbf{v}(\mathbf{x}, t)$ inversion problem while preserving the critical information about spatiotemporal gradient dynamics.

Interpretation and Advantages. The NSG statistic $\mathbf{g}(\mathbf{x}, t)$ quantifies the directional sensitivity of probability flow per unit temporal variation, driven by both spatial gradients ($\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$) and temporal derivatives ($\partial_t \log p(\mathbf{x}, t)$). This statistic captures both spatial irregularities (via $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$) and temporal inconsistencies (via $\partial_t \log p(\mathbf{x}, t)$), enabling comprehensive analysis of video dynamics. Moreover, by modeling fundamental probability flow dynamics, NSG avoids dependencies on specific artifacts, making it suitable for detecting generated videos across diverse generation paradigms.

3.2 Estimating NSG with Diffusion Models

The NSG statistic $\mathbf{g}(\mathbf{x}, t)$ in Eqn. (6) requires estimating two key components: the spatial gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ and the temporal derivatives $\partial_t \log p(\mathbf{x}, t)$. Using diffusion models' inherent gradient estimation ability [30, 31], we propose an effective estimator combining spatial gradients from pre-trained diffusion models with motion-aware temporal dynamics using Eqn. (8) and (9), yielding:

$$\mathbf{g}(\mathbf{x}, t) \approx \frac{\mathbf{s}_\theta(\mathbf{x}_t)}{\mathbf{s}_\theta(\mathbf{x}_t) \cdot \frac{\mathbf{x}_{t+\Delta t} - \mathbf{x}_t}{\Delta t} + \lambda}, \tag{7}$$

where $\mathbf{s}_\theta$ denotes the learned score function from diffusion models and $\mathbf{x}_t$ represents the t-th video frame. This estimator eliminates the need for explicit flow computation while preserving critical spatiotemporal dynamics through physics-inspired modeling. Below, we detail its derivation.

Spatial Gradients Estimation. Diffusion models [31, 56] explicitly learn a score network $\mathbf{s}_\theta$ through score matching [30] or denoising diffusion modeling [41] to approximate $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$. For a given video $\mathbf{x}$ at t-th frame, the spatial gradient is estimated by:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \approx \mathbf{s}_\theta(\mathbf{x}_t), \tag{8}$$

where $\mathbf{x}_t$ is the t-th frame of the video $\mathbf{x}$. Here, we omit the diffusion timestep to make the notation clearer. This estimation allows direct computation of $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ in NSG via a single forward pass of the pre-trained diffusion model, eliminating the need for numerical differentiation.

Temporal Derivatives Approximation. To estimate $\partial_t \log p(\mathbf{x}, t)$, we exploit the temporal coherence of video sequences under the *brightness constancy assumption* [32], which posits that the probability density along motion trajectories remains constant. This leads to the following approximation:

Proposition 1. *Under the brightness constancy assumption $p(\mathbf{x} + \Delta\mathbf{x}, t + \Delta t) \approx p(\mathbf{x}, t)$ with small inter-frame motion ($\Delta t \to 0$) and inter-frame displacement ($\Delta\mathbf{x} \to 0$), we have*

$$\partial_t \log p(\mathbf{x}, t) \approx -\frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \cdot \Delta\mathbf{x}}{\Delta t}. \tag{9}$$

3.3 Exploring NSG for Detecting AI-Generated Videos

To effectively distinguish AI-generated content from real videos, it is crucial to design metrics that capture subtle distributional discrepancies in high-dimensional spatiotemporal features. Recent studies show *Maximum Mean Discrepancy* (MMD) [33]—a non-parametric statistic for distribution alignment—has demonstrated remarkable capabilities in measuring distributional differences, *e.g.*, AI-text detection [48, 46] and adversarial samples detection [57, 58]. Building upon MMD's theoretical foundation and NSG's unique strength in modeling spatiotemporal dynamics, we propose an **NSG-based video detection method (NSG-VD)** that integrates MMD with the NSG feature representation.

MMD Formulation with NSG Features. We aggregate NSG features across T frames in each video as $\mathbf{G}(\mathbf{x})=\{\mathbf{g}(\mathbf{x},t)\}_{t=1}^{T}$. Let $S_{\mathbb{P}}^{re}=\{\mathbf{x}^{(i)}\}_{i=1}^{n}$ denote a reference set of real videos and $S_{\mathbb{Q}}^{te}=\{\tilde{\mathbf{y}}\}$ represent a test video. The MMD [33] between $S_{\mathbb{P}}^{re}$ and $S_{\mathbb{Q}}^{te}$ in terms of NSG is computed as:

$$\widehat{\mathrm{MMD}}_b^2\left[S_{\mathbb{P}}^{re}, S_{\mathbb{Q}}^{te}; \mathcal{H}_k\right] = \frac{1}{n^2}\sum_{i,j=1}^{n} k\Big(\mathbf{G}^{(i)}, \mathbf{G}^{(j)}\Big) - \frac{2}{n}\sum_{i=1}^{n} k\Big(\mathbf{G}^{(i)}, \mathbf{G}^{(\text{test})}\Big) + k\Big(\mathbf{G}^{(\text{test})}, \mathbf{G}^{(\text{test})}\Big), \quad (10)$$

where $\mathbf{G}^{(i)} = \mathbf{G}(\mathbf{x}^{(i)})$ and $\mathbf{G}^{(\text{test})} = \mathbf{G}(\tilde{\mathbf{y}})$ are NSG features extracted from real and test videos. The kernel $k : \mathcal{G} \times \mathcal{G} \to \mathbb{R}$ maps NSG features to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$, such as the Gaussian kernel $k(\mathbf{a}, \mathbf{b}) = \exp\left(-\|\mathbf{a}-\mathbf{b}\|^2 / (2\sigma^2)\right)$. Note that while MMD is conventionally used for distribution-level comparisons, recent studies [48, 46, 58] validate its efficacy in single-sample detection by quantifying deviations from reference distributions. Crucially, while MMD provides a viable solution for distributional comparison, the core advantage of NSG-VD stems from the NSG itself modeling fundamental spatiotemporal dynamics (see details in Appendix E.3).

Detection Protocol with MMD Metric. Let $f(\tilde{\mathbf{y}}; S_{\mathbb{P}}, k_\omega, \tau) = \mathbb{I}\left(\widehat{\mathrm{MMD}}_b^2 > \tau\right)$, where $\mathbb{I}$ is the indicator function and τ is a threshold for the decision. Given a test video $\tilde{\mathbf{y}}$, we compute the MMD with NSG against a referenced real video set and give the decision:

$$f(\tilde{\mathbf{y}}) = \begin{cases} \text{Fake}, & \text{if } f(\tilde{\mathbf{y}}; S_{\mathbb{P}}, k_\omega, \tau) = 1, \\ \text{Real}, & \text{if } f(\tilde{\mathbf{y}}; S_{\mathbb{P}}, k_\omega, \tau) = 0. \end{cases} \quad (11)$$

Optimization for NSG-VD. To enhance discriminative power, we use a deep kernel [34] for MMD:

$$k_\omega(\mathbf{x}, \mathbf{y}) = \left[(1-\epsilon)\kappa\left(\phi_{\mathbf{G}}(\mathbf{x}), \phi_{\mathbf{G}}(\mathbf{y})\right) + \epsilon\right] \cdot \Phi\left(\mathbf{G}(\mathbf{x}), \mathbf{G}(\mathbf{y})\right), \quad (12)$$

where $\phi_{\mathbf{G}}(\mathbf{x}) = \phi(\mathbf{G}(\mathbf{x}))$ is a deep neural network, κ and Φ are Gaussian kernels with bandwidths σ_ϕ and σ_Φ, and $\epsilon \in (0,1)$. The kernel parameters $\omega=\{\epsilon, \phi, \sigma_\phi, \sigma_\Phi\}$ will be optimized by Eqn. (13) to maximize the detection ability. Considering the multiple-population scenarios across diverse video distributions [48], we adopt a *multi-population aware optimization* for the kernel training:

$$k_\omega^* = \arg\max_{k_\omega} \frac{\widehat{\mathrm{MPP}}_u(S_{\mathbb{P}}^{tr}, S_{\mathbb{Q}}^{tr}; k_\omega)}{\sqrt{\hat{\sigma}^2(S_{\mathbb{P}}^{tr}, S_{\mathbb{Q}}^{tr}; k_\omega) + \lambda}}, \quad \hat{\sigma}^2 = \frac{4}{N^3}\sum_{i=1}^{N}\left(\sum_{j=1}^{N} H_{ij}^*\right)^2 - \frac{4}{N^4}\left(\sum_{i=1}^{N}\sum_{j=1}^{N} H_{ij}^*\right)^2, \quad (13)$$

where $S_{\mathbb{P}}^{tr}$ and $S_{\mathbb{Q}}^{tr}$ denote the training real and generated videos, respectively, $\widehat{\mathrm{MPP}}_u(S_{\mathbb{P}}^{tr}, S_{\mathbb{Q}}^{tr}; k_\omega) = \frac{1}{N(N-1)}\sum_{i\neq j} H_{ij}^*$ and $H_{ij}^* = k_\omega(\mathbf{x}_i, \mathbf{x}_j) - k_\omega(\mathbf{x}_i, \mathbf{y}_j) - k_\omega(\mathbf{y}_i, \mathbf{x}_j)$.

3.4 Theoretical Guarantees for NSG-VD

The effectiveness of NSG-VD relies on ensuring the MMD between NSG features of real videos is smaller than that between real and generated videos. To formalize this, we analyze the MMD formulation in Eqn. (10), where the key discriminative information lies in the cross-term $k\left(\mathbf{G}^{(i)}, \mathbf{G}^{(\text{test})}\right)$ since the first and third terms remain invariant for fixed reference sets. Under the Gaussian kernel, this cross-term is dominated by the exponential squared distance between NSG features. Note that analyzing $\frac{\nabla_{\mathbf{x}} \log p(\mathbf{x},t)}{-\partial_t \log p(\mathbf{x},t)+\lambda}$ under practical distributions can be very difficult and infeasible, we adopt a common practice [59, 60] by assuming Gaussian-distributed data to derive theoretical insights. Below, we first characterize the NSG statistics for real and generated videos under Gaussian assumptions.

Proposition 2. *Let the real video distribution be* $p(\mathbf{x},t)=\mathcal{N}(\mathbf{0},\sigma(t)^2\mathbf{I}_d)$ *and the generated video distribution be* $q(\mathbf{y},t)=\mathcal{N}(\boldsymbol{\mu},\sigma(t)^2\mathbf{I}_d)$*, respectively, where* $\mathbf{I}_d\in\mathbb{R}^{d\times d}$ *is an identity matrix and* $\boldsymbol{\mu}\neq\mathbf{0}\in\mathbb{R}^d$ *is the distribution shift and* $\sigma(t)\neq\mathbf{0}$*, the NSG* $\mathbf{g}(\mathbf{x},t)$ *and* $\mathbf{g}(\mathbf{y},t)$ *satisfy:*

$$\mathbf{g}(\mathbf{x},t)=-\frac{\mathbf{x}/\sigma(t)^2}{D_r(\mathbf{x})},\ -\frac{\mathbf{x}}{\sigma(t)^2}\sim\mathcal{N}\Big(\mathbf{0},\sigma(t)^2\mathbf{I}_d\Big),\ D_r(\mathbf{x})\sim\lambda+\frac{d\dot{\sigma}(t)}{\sigma(t)}-\frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d);$$
$$\mathbf{g}(\mathbf{y},t)=-\frac{\mathbf{y}/\sigma(t)^2}{D_f(\mathbf{y})},\ -\frac{\mathbf{y}}{\sigma(t)^2}\sim\mathcal{N}\Big(-\frac{\boldsymbol{\mu}}{\sigma(t)},\sigma(t)^2\mathbf{I}_d\Big),\ D_f(\mathbf{y})\sim\lambda+\frac{d\dot{\sigma}(t)}{\sigma(t)}-\frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d,\varphi),$$

where $D_r(\mathbf{x})=\lambda+\frac{d\dot{\sigma}(t)}{\sigma(t)}-\frac{\|\mathbf{x}\|^2\dot{\sigma}(t)}{\sigma(t)^3}$, $D_f(\mathbf{y})=\lambda+\frac{d\dot{\sigma}(t)}{\sigma(t)}-\frac{\|\mathbf{y}\|^2\dot{\sigma}(t)}{\sigma(t)^3}$, $\dot{\sigma}(t)\triangleq\frac{d}{dt}\sigma(t)$*, and* $\varphi=\frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$*,* $\chi^2(d)$ *is the central chi-squared distribution with* d *degrees of freedom and* $\chi^2(d,\varphi)$ *is the noncentral chi-squared distribution with noncentrality parameter* φ *and* d *degrees of freedom [61].*

Proposition 2 reveals that the distribution shift $\boldsymbol{\mu}$ in generated videos introduces deviations in both the numerator and denominator of the NSG, *i.e.*, *noncentral* Gaussian and chi-squared distributions. To quantify this deviation, we derive an upper bound on the squared distance between NSGs:

Theorem 1. *Let the real video distribution be* $\mathbf{x}\sim\mathcal{N}(\mathbf{0},\sigma(t)^2\mathbf{I}_d)$ *and the generated video distribution be* $\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu},\sigma(t)^2\mathbf{I}_d)$*, respectively, where* $\mathbf{I}_d\in\mathbb{R}^{d\times d}$ *is an identity matrix and* $\boldsymbol{\mu}\neq\mathbf{0}\in\mathbb{R}^d$ *is the distribution shift. Given* $\mathbf{G}(\mathbf{x})=\{\mathbf{g}(\mathbf{x},t)\}_{t=1}^T$*, denote* $\varphi=\|\boldsymbol{\mu}\|^2/\sigma(t)^2$ *and assume* $|-\partial_t\log p(\mathbf{x},t)+\lambda|\geq C>0$ *and* $|-\partial_t\log p(\mathbf{y},t)+\lambda|\geq C>0$*, with probability at least* $1-\delta$*, we have*

$$\|\mathbf{G}(\mathbf{x})-\mathbf{G}(\mathbf{y})\|^2\leq\mathcal{O}\left(\frac{T}{C^4\sigma(t)^2}\left[\varphi d+d^2+\varphi+\log\frac{T}{\delta}\cdot(\varphi+d)+\log^2\frac{T}{\delta}\right]\right).$$

Theorem 1 reveals that the bound of the squared distance between NSG features of real and fake data will be smaller if the distribution shift term $\varphi=\|\boldsymbol{\mu}\|^2/\sigma(t)^2$ is closer to zero for a given δ. This formalizes the intuition that small distribution shifts produce small geometric distortions in NSG space, while significant deviations in synthetic content lead to large separations from real data. Under the Gaussian kernel, this implies that the real data have a larger $k(\mathbf{G}(\mathbf{x}),\mathbf{G}(\mathbf{y}))$ than the fake data since the distribution shift term $\varphi=0$ for real data. Therefore, when substituted into Eqn. (10), the MMD between NSG features of real videos is smaller than that between real and generated videos.

4 Experiments

Datasets. We evaluate our methods on the GenVideo benchmark [25], a large-scale dataset for AI-generated video detection that includes diverse real-world videos and synthetic content from multiple generative models. We use Kinetics-400 [62] as the real video source, SEINE [63] or Pika [64] as the AI-generated videos for training. The test set comprises MSR-VTT [65] and 10 diverse AI-generated datasets from different generation paradigms. More details are in Appendix C.1.

Evaluation Metrics. We evaluate the performance of video detection on Recall, Accuracy, F1-score [66] and AUROC [67] metrics. More details are provided in Appendix C.2. We use **bold** numbers to indicate the best results and underlined numbers to denote the second-best results in tables.

Baselines. We compare our NSG-VD with following baselines: TALL [24], NPR [27], STIL [36], and Demamba [25]. These baselines are implemented based on the codebase provided by Demamba [25].

4.1 Comparisons on Standard Evaluation

We start by comparing our NSG-VD with baselines using $10,000$ real videos from Kinetics-400 and $10,000$ generated videos from Pika (Table 1) and SENIE (Table 2) for training, respectively.

Results on Trained with Kinetics-400 and Pika. From Table 1, existing methods exhibit critical limitations. For instance, Demamba struggles with generative paradigms like HotShot (40.60% Recall) and Sora (48.21% Recall), while NPR shows unstable performance with Accuracy ranging from 57.20% to 98.20%. TALL fails on synthetic outliers (*e.g.*, 25.00% Recall on Sora) and STIL collapses completely on critical cases (*e.g.*, 1.40% Recall on HotShot and 1.79% Recall on Sora), revealing limitations of their inherent dependencies on generator-specific artifacts.

Table 1: Comparisons with baselines on *a standard evaluation* (%), where we train all models with $10,000$ real and generated videos from Kinetics-400 and Pika, respectively.

Method	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
DeMamba	Recall	87.00	93.60	98.80	40.60	48.40	98.00	88.40	59.00	48.21	58.20	72.02
	Accuracy	91.70	95.00	97.60	68.50	72.40	97.20	92.40	77.70	72.32	77.30	84.21
	F1	91.29	94.93	97.63	56.31	63.68	97.22	92.08	72.57	63.53	71.94	80.12
	AUROC	98.04	98.82	99.68	87.84	90.12	99.46	97.81	91.32	88.36	87.38	93.88
NPR	Recall	61.20	80.00	98.00	16.00	33.00	91.20	80.60	34.60	35.71	43.20	57.35
	Accuracy	79.80	89.20	98.20	57.20	65.70	94.80	89.50	66.50	67.86	70.80	77.96
	F1	75.18	88.11	98.20	27.21	49.03	94.61	88.47	50.81	52.63	59.67	68.39
	AUROC	93.05	97.18	99.66	82.97	90.50	99.13	97.87	87.54	90.47	91.84	93.02
TALL	Recall	51.20	65.20	93.40	32.00	61.60	94.80	81.80	49.20	25.00	53.60	60.78
	Accuracy	75.10	82.10	96.20	65.50	80.30	96.90	90.40	74.10	61.61	76.30	79.85
	F1	67.28	78.46	96.09	48.12	75.77	96.83	89.50	65.51	39.44	69.34	72.63
	AUROC	95.82	97.14	99.73	92.55	97.36	99.79	99.09	94.84	86.67	93.75	95.67
STIL	Recall	73.80	70.80	43.40	1.40	2.00	45.00	13.20	7.20	1.79	11.60	27.02
	Accuracy	86.90	85.40	71.70	50.70	51.00	72.50	56.60	53.60	50.89	55.80	63.51
	F1	84.93	82.90	60.53	2.76	3.92	62.07	23.32	13.43	3.51	20.79	35.82
	AUROC	96.43	97.77	99.34	86.66	90.56	98.88	97.04	88.16	92.57	87.52	93.49
NSG-VD (Ours)	Recall	68.33	98.33	100.00	92.50	87.50	80.00	98.33	94.17	78.57	82.50	**88.02**
	Accuracy	81.67	98.33	96.67	91.67	90.83	88.33	95.83	94.17	88.39	88.75	**91.46**
	F1	78.85	98.33	96.77	91.74	90.52	87.27	95.93	94.17	87.13	88.00	**90.87**
	AUROC	92.26	98.66	98.15	94.45	96.38	94.83	98.16	97.41	96.40	94.73	**96.14**

Table 2: Comparisons with baselines on *a standard evaluation* (%), where we train all models with $10,000$ real and generated videos from Kinetics-400 and SEINE, respectively.

Method	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
DeMamba	Recall	47.40	87.80	88.20	77.40	75.00	85.60	91.60	68.60	42.86	48.00	71.25
	Accuracy	72.80	93.00	93.20	87.80	86.60	91.90	94.90	83.40	68.75	73.10	84.54
	F1	63.54	92.62	92.84	86.38	84.84	91.36	94.73	80.52	57.83	64.09	80.87
	AUROC	88.29	98.39	98.76	97.84	96.89	98.76	99.35	96.87	80.93	88.11	94.42
NPR	Recall	46.40	76.40	69.80	63.80	56.00	75.00	83.80	58.80	35.71	27.40	59.31
	Accuracy	71.40	86.40	83.10	80.10	76.20	85.70	90.10	77.60	66.96	61.90	77.95
	F1	61.87	84.89	80.51	76.22	70.18	83.99	89.43	72.41	51.95	41.83	71.33
	AUROC	85.73	96.01	93.79	91.44	89.96	95.13	96.87	89.46	84.15	76.66	89.92
TALL	Recall	58.60	75.00	79.40	60.20	62.00	77.80	88.20	43.80	33.93	35.80	61.47
	Accuracy	78.80	87.00	89.20	79.60	80.50	88.40	93.60	71.40	66.07	67.40	80.20
	F1	73.43	85.23	88.03	74.69	76.07	87.02	93.23	60.50	50.00	52.34	74.05
	AUROC	97.10	98.12	98.63	96.37	96.45	97.76	99.38	94.80	83.35	89.45	95.14
STIL	Recall	28.60	57.40	78.40	46.80	18.80	66.40	69.00	24.80	14.29	19.00	42.35
	Accuracy	64.20	78.60	89.10	73.30	59.30	83.10	84.40	62.30	57.14	59.40	71.08
	F1	44.41	72.84	87.79	63.67	31.60	79.71	81.56	39.68	25.00	31.88	55.81
	AUROC	95.53	97.91	99.40	96.49	92.79	98.06	98.86	91.00	92.79	86.58	94.94
NSG-VD (Ours)	Recall	91.67	100.00	100.00	100.00	100.00	98.33	100.00	97.50	94.64	89.17	**97.13**
	Accuracy	82.50	88.33	89.58	84.58	86.25	87.08	86.67	87.92	89.29	78.33	**86.05**
	F1	83.97	89.55	90.57	86.64	87.91	88.39	88.24	88.97	89.83	80.45	**87.45**
	AUROC	90.67	97.62	98.38	95.88	96.69	97.87	97.64	95.09	96.14	88.65	**95.46**

In contrast, our NSG-VD achieves state-of-the-art performance across all metrics, significantly outperforming baselines despite not being pre-trained on large-scale videos. Remarkably, NSG-VD demonstrates exceptional reliability on challenging closed-source generators like Sora (78.57% Recall vs. 48.21% for Demamba) and emerging paradigms like HotShot (92.50% Recall vs. 40.60% for Demamba), and maintains reliability across other diverse domains (*e.g.*, MorphStudio, MoonValley). Notably, our NSG-VD achieves $16.00\% \uparrow$ average Recall and $10.75\% \uparrow$ F1-score over Demamba, and $55.05\% \uparrow$ F1-score over STIL. These results confirm its generalization across both open-source and closed-source generated models, highlighting the advantages of physics-driven modeling.

Results on Trained with Kinetics-400 and SENIE. As shown in Table 2, our NSG-VD achieves superior detection performance across all metrics compared to baselines. Notably, it attains near-perfect Recall ($\geq 98.33\%$) on models like MoonValley, HotShot and Show1, while maintaining balanced performance across diverse domains (*e.g.*, ModelScope, WildScrape). These results are consistent with the results on Pika in Table 1, further demonstrating the effectiveness of our proposed method. In contrast, existing baselines exhibit pronounced limitations under this setting. Demamba's performance is more constrained ($\leq 85.60\%$ Recall on most models), and NPR's F1-score varies widely ($41.83\% \sim 89.43\%$). TALL shows instability on models like Sora (33.93% Recall), while

Table 3: Comparisons with baselines under *data-imbalanced scenarios* (%), where we train all models with $10,000$ real and $1,000$ generated videos from Kinetics-400 and SEINE, respectively.

Method	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
DeMamba	Recall	56.80	80.40	82.60	65.60	63.80	78.20	83.00	53.40	33.93	43.20	64.09
	Accuracy	78.10	89.90	91.00	82.50	81.60	88.80	91.20	76.40	65.18	71.30	81.60
	F1	72.17	88.84	90.17	78.94	77.62	87.47	90.41	69.35	49.35	60.08	76.44
	AUROC	93.01	98.17	98.90	96.42	95.36	98.38	98.74	95.50	86.51	87.49	94.85
NPR	Recall	25.40	52.20	42.40	26.00	21.40	48.20	66.60	22.00	10.71	12.20	32.71
	Accuracy	62.40	75.80	70.90	62.70	60.40	73.80	83.00	60.70	55.36	55.80	66.09
	F1	40.32	68.32	59.30	41.07	35.08	64.78	79.67	35.89	19.35	21.63	46.54
	AUROC	83.64	94.85	92.44	86.68	83.77	94.33	95.77	84.34	84.60	70.52	87.10
TALL	Recall	28.20	45.20	41.20	26.20	33.80	60.20	60.20	22.60	25.00	18.20	36.08
	Accuracy	64.00	72.50	70.50	63.00	66.80	80.00	80.00	61.20	62.50	59.00	67.95
	F1	43.93	62.17	58.27	41.46	50.45	75.06	75.06	36.81	40.00	30.74	51.40
	AUROC	93.34	94.56	94.25	91.64	91.63	94.99	97.60	91.46	84.92	85.20	91.96
STIL	Recall	25.80	64.80	68.40	46.20	29.20	67.20	70.00	44.40	26.79	25.00	46.78
	Accuracy	62.70	82.20	84.00	72.90	64.40	83.40	84.80	72.00	63.39	62.30	73.21
	F1	40.89	78.45	81.04	63.03	45.06	80.19	82.16	61.33	42.25	39.87	61.43
	AUROC	85.14	95.74	96.87	89.46	83.22	96.09	96.23	90.36	89.89	78.99	90.20
NSG-VD (Ours)	Recall	85.83	99.17	100.00	99.17	97.50	95.83	99.17	91.67	82.14	81.67	**93.21**
	Accuracy	84.58	92.50	93.75	89.58	89.58	90.83	92.50	90.00	86.61	81.67	**89.16**
	F1	84.77	92.97	94.12	90.49	90.35	91.27	92.97	90.16	85.98	81.67	**89.48**
	AUROC	90.76	98.18	98.18	95.03	95.48	96.97	96.53	95.11	95.73	87.13	**94.91**

STIL fails entirely on critical cases (*e.g.*, 19.00% Recall on WildScrape). These failures highlight the fragility of artifact-based approaches in capturing subtle spatiotemporal inconsistencies.

Quantitatively, NSG-VD surpasses Demamba by $25.88\% \uparrow$ in average Recall (97.13% vs. 71.25%) and NPR by $16.12\% \uparrow$ in average F1-score (87.45% vs. 71.33%). On closed-source models like Sora, it achieves 94.64% Recall—nearly twice Demamba's (42.86%) and sextuple STIL's (14.29%). This improvement highlights NSG-VD's sensitivity to synthetic anomalies, especially in near-photorealistic videos (*e.g.*, Sora), where subtle spatiotemporal inconsistencies are amplified by the NSG but not effectively captured by baselines, indicating reliable detection across diverse generation paradigms.

4.2 Comparisons on Challenging Data-Imbalanced Scenarios

In real-world scenarios, natural videos are often abundant and accessible, while collecting sufficient AI-generated videos remains challenging due to rapidly evolving generation techniques. To thoroughly assess reliability under these conditions, we train all models using $10,000$ Kinetics-400 real videos and only $1,000$ SENIE-generated videos. As shown in Table 3, all baselines exhibit significant limitations. Demamba fails catastrophically on challenging generators like Sora (33.93% Recall) and WildScrape (43.20% Recall), while NPR exhibits fluctuations in Accuracy ($55.36\% \sim 83.00\%$). TALL fails completely on emerging paradigms like WildScrape (18.20% Recall) and Lavie (22.60% Recall), and STIL shows highly variable performance, *e.g.*, $25.00\% \sim 70.00\%$ Recall. Such instability indicates over-reliance on synthetic data volume or sensitivity to superficial artifacts.

In contrast, NSG-VD achieves strong generalization across 10 diverse generations. Notably, NSG-VD attains superior performance on critical test cases: 82.14% Recall on Sora (vs. $10.71\% \sim 33.93\%$ for baselines) and 81.67% Recall on WildScrape (vs. $12.20\% \sim 43.20\%$). Critically, NSG-VD achieves $29.12\% \uparrow$ higher average Recall than Demamba and $38.08\% \uparrow$ higher F1-score than TALL. These results confirm NSG-VD's reliable generalization from limited synthetic data without compromising discriminative power, demonstrating that adherence to universal physical principles outperforms domain-specific feature reliance even when synthetic training data is severely constrained.

4.3 Impact of Spatial Gradients and Temporal Derivatives for NSG-VD

To investigate the impact of the spatial gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ and temporal derivatives $\partial_t \log p(\mathbf{x}, t)$ for our NSG-VD, we evaluate these components as independent detection statistics for AI-generated video detection. To this end, we train these separate models with $10,000$ real videos from Kinetics-400 and generated videos from Pika. From Table 4, the spatial

Table 4: Impact of spatial gradients and temporal derivatives on average metrics (%).

Method	Recall	Accuracy	F1	AUROC
Spatial Gradients	87.99	82.84	83.40	91.85
Temporal Derivatives	60.35	71.09	66.97	78.95
NSG-VD (Ours)	**88.02**	**91.46**	**90.87**	**96.14**

gradient achieves moderate performance (*e.g.*, 87.99% Recall, 83.40% F1-score), suggesting its ability to capture spatial anomalies, which may arise from its sensitivity to localized variations in texture or geometry. The temporal derivative, however, shows limited detection power (*e.g.*, 60.35% Recall, 66.97% F1-score), likely due to its sensitivity to transient noise in dynamic modeling. In contrast, our NSG-VD integrating both components achieves significantly enhanced performance (*e.g.*, 88.02% Recall, 90.87% F1-score). This demonstrates that the interplay between spatial gradients and temporal derivatives formalized via physical conservation principles is critical for video detection.

4.4 Impact of Decision Threshold for NSG-VD

We evaluate the decision threshold τ in Eqn. (11) for NSG-VD by testing $\tau \in [0.4, 1.3]$ under the same settings as Table 1. As shown in Figure 3, our NSG-VD maintains remarkably stable performance across a wide range of τ values without requiring fine-grained tuning. Specifically, NSG-VD consistently shows high detection performance as $\tau \in [0.7, 1.1]$ for average Recall, Accuracy and F1-Score across diverse generators. These results indicate that NSG features create a clear separation between real and fake distributions. We set $\tau = 1.0$ as the default throughout all settings.

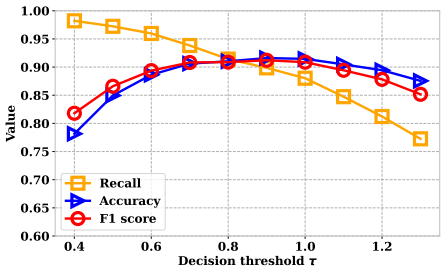


Figure 3: Impact of decision threshold.

5 Conclusion

In this paper, we propose a physics-driven AI-generated video detection paradigm by modeling spatiotemporal dynamics through the Normalized Spatiotemporal Gradient (NSG), a novel statistic based on probability flow conservation principles. Leveraging pre-trained diffusion models, we propose an NSG-based video detection method (NSG-VD). Theoretical analyses and extensive experiments validate the superiority of our NSG-VD in detecting advanced generated videos.

Acknowledgements

This work was partially supported by the Joint Funds of the National Natural Science Foundation of China (Grant No.U24A20327), RGC Young Collaborative Research Grant No. C2005-24Y, RGC General Research Fund No. 12200725, and NSFC General Program No. 62376235.

References

[1] 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] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[3] 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. *OpenAI Blog*, 1:8, 2024.

[4] Tao Wu, Yong Zhang, Xintao Wang, Xianpan Zhou, Guangcong Zheng, Zhongang Qi, Ying Shan, and Xi Li. Customcrafter: Customized video generation with preserving motion and concept composition abilities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 8469–8477, 2025.

[5] Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? 2024.

[6] Hsin-Ping Huang, Yu-Chuan Su, Deqing Sun, Lu Jiang, Xuhui Jia, Yukun Zhu, and Ming-Hsuan Yang. Fine-grained controllable video generation via object appearance and context. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3698–3708. IEEE, 2025.

[7] 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.

[8] 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, 2024.

[9] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[10] Yaowei Li, Xintao Wang, Zhaoyang Zhang, Zhouxia Wang, Ziyang Yuan, Liangbin Xie, Ying Shan, and Yuexian Zou. Image conductor: Precision control for interactive video synthesis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5031–5038, 2025.

[11] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Towards open-set identity preserving face synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6713–6722, 2018.

[12] Zinan Guo, Yanze Wu, Chen Zhuowei, Peng Zhang, Qian He, et al. Pulid: Pure and lightning id customization via contrastive alignment. *Advances in neural information processing systems*, 37:36777–36804, 2024.

[13] Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. In *NeurIPS*, 2023.

[14] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 1136–1142. International Joint Conferences on Artificial Intelligence Organization, 2021.

[15] Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023.

[16] Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27102–27112, 2024.

[17] Jincheng Li, Shuhai Zhang, Jiezhang Cao, and Mingkui Tan. Learning defense transformations for counterattacking adversarial examples. *Neural Networks*, 164:177–185, 2023.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

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

generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21807–21818, June 2024.

[20] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

[21] Hao Ouyang, Qiuyu Wang, Yuxi Xiao, Qingyan Bai, Juntao Zhang, Kecheng Zheng, Xiaowei Zhou, Qifeng Chen, and Yujun Shen. Codef: Content deformation fields for temporally consistent video processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8089–8099, 2024.

[22] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019.

[23] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4129–4138, 2023.

[24] 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.

[25] 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.

[26] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[27] 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.

[28] George Keith Batchelor. *An introduction to fluid dynamics*. Cambridge university press, 2000.

[29] Michel Rieutord. *Fluid dynamics: an introduction*. Springer, 2014.

[30] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[32] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[33] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.

[34] Feng Liu, Wenkai Xu, Jie Lu, Guangquan Zhang, Arthur Gretton, and Danica J Sutherland. Learning deep kernels for non-parametric two-sample tests. In *International Conference on Machine Learning*, pages 6316–6326. PMLR, 2020.

[35] 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.

[36] 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.

[37] Chunlei Peng, Zimin Miao, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Where deepfakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2024.

[38] 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.

[39] 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.

[40] Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, LEI BAI, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

[41] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[42] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *NeurIPS*, volume 33, pages 12438–12448, 2020.

[43] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *ICML*, pages 16805–16827. PMLR, 2022.

[44] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *ICML*, pages 12062–12072. PMLR, 2021.

[45] Shuhai Zhang, Feng Liu, Jiahao Yang, Yifan Yang, Changsheng Li, Bo Han, and Mingkui Tan. Detecting adversarial data by probing multiple perturbations using expected perturbation score. In *International Conference on Machine Learning*, pages 41429–41451. PMLR, 2023.

[46] Yiliao Song, Zhenqiao Yuan, Shuhai Zhang, Zhen Fang, Jun Yu, and Feng Liu. Deep kernel relative test for machine-generated text detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

[47] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[48] Shuhai Zhang, Yiliao Song, Jiahao Yang, Yuanqing Li, Bo Han, and Mingkui Tan. Detecting machine-generated texts by multi-population aware optimization for maximum mean discrepancy. In *International Conference on Learning Representations*, 2024.

[49] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022.

[50] Frank Wilczek. Quantum field theory. *Reviews of Modern Physics*, 71(2):S85, 1999.

[51] John Lighton Synge and Alfred Schild. *Tensor calculus*, volume 5. Courier Corporation, 1978.

[52] WB Hodge, SV Migirditch, and William C Kerr. Electron spin and probability current density in quantum mechanics. *American Journal of Physics*, 82(7):681–690, 2014.

[53] Hannes Risken. Fokker-planck equation. In *The Fokker-Planck equation: methods of solution and applications*, pages 63–95. Springer, 1989.

[54] Ronald L Panton. *Incompressible flow*. John Wiley & Sons, 2024.

[55] Arno Böhm. *Quantum mechanics: foundations and applications*. Springer Science & Business Media, 2013.

[56] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[57] Ruize Gao, Feng Liu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, and Masashi Sugiyama. Maximum mean discrepancy test is aware of adversarial attacks. In *International Conference on Machine Learning*, pages 3564–3575. PMLR, 2021.

[58] Shuhai Zhang, Feng Liu, Jiahao Yang, Yifan Yang, Changsheng Li, Bo Han, and Mingkui Tan. Detecting adversarial data by probing multiple perturbations using expected perturbation score. In *International conference on machine learning*, pages 41429–41451. PMLR, 2023.

[59] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. *Advances in neural information processing systems*, 36:54046–54060, 2023.

[60] Yiming Wang, Pei Zhang, Baosong Yang, Derek Wong, Zhuosheng Zhang, and Rui Wang. Embedding trajectory for out-of-distribution detection in mathematical reasoning. *Advances in Neural Information Processing Systems*, 37:42965–42999, 2024.

[61] SH Abdel-Aty. Approximate formulae for the percentage points and the probability integral of the non-central χ 2 distribution. *Biometrika*, 41(3/4):538–540, 1954.

[62] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[63] 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.

[64] Leijie Wang, Nicholas Vincent, Julija Rukanskaitė, and Amy Xian Zhang. Pika: Empowering non-programmers to author executable governance policies in online communities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–18, 2024.

[65] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[66] Nancy Chinchor and Beth M Sundheim. Muc-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993.

[67] Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

[68] Lucien Birgé and Pascal Massart. Minimum contrast estimators on sieves: exponential bounds and rates of convergence. 1998.

[69] Patrik Róbert Gerber, Tianze Jiang, Yury Polyanskiy, and Rui Sun. Kernel-based tests for likelihood-free hypothesis testing. *Advances in Neural Information Processing Systems*, 36:15680–15715, 2023.

[70] Florian Kalinke and Zoltán Szabó. Nyström m-hilbert-schmidt independence criterion. In *Uncertainty in Artificial Intelligence*, pages 1005–1015. PMLR, 2023.

[71] Bo Han, Jiangchao Yao, Tongliang Liu, Bo Li, Sanmi Koyejo, Feng Liu, et al. Trustworthy machine learning: From data to models. *Foundations and Trends® in Privacy and Security*, 7(2-3):74–246, 2025.

[72] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29(2):429–443, 1997.

[73] Haoang Chi, Feng Liu, Wenjing Yang, Long Lan, Tongliang Liu, Bo Han, William K. Cheung, and James T. Kwok. TOHAN: A one-step approach towards few-shot hypothesis adaptation. In *NeurIPS*, 2021.

[74] Ilya O Tolstikhin, Bharath K Sriperumbudur, and Bernhard Schölkopf. Minimax estimation of maximum mean discrepancy with radial kernels. *Advances in Neural Information Processing Systems*, 29, 2016.

[75] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *The Annals of Statistics*, 50(1):225–251, 2022.

[76] 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.

[77] Zeroscope. 2023. URL https://huggingface.co/cerspense/zeroscope_v2_XL.

[78] 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.

[79] Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. Pia: Your personalized image animator via plug-and-play modules in text-to-image models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7747–7756, 2024.

[80] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[81] 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. *arXiv preprint arXiv:2412.20404*, 2024.

[82] 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.

[83] Hotshot. 2023. URL https://github.com/hotshotco/Hotshot-XL.

[84] 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, 2023.

[85] 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.

[86] 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.

[87] 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.

[88] Alberto Jiménez-Valverde. Insights into the area under the receiver operating characteristic curve (auc) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography*, 21(4):498–507, 2012.

[89] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[90] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[91] Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. *Advances in Neural Information Processing Systems*, 2023.

[92] Junyi Wu, Haoxuan Wang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Ptq4dit: Post-training quantization for diffusion transformers. *Advances in Neural Information Processing Systems*, 2024.

[93] Muyang Li, Yujun Lin, Zhekai Zhang, Tianle Cai, Xiuyu Li, Junxian Guo, Enze Xie, Chenlin Meng, Jun-Yan Zhu, and Song Han. Svdqunat: Absorbing outliers by low-rank components for 4-bit diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[94] Yang Zhao, Yanwu Xu, Zhisheng Xiao, Haolin Jia, and Tingbo Hou. Mobilediffusion: Instant text-to-image generation on mobile devices. In *European Conference on Computer Vision*, pages 225–242. Springer, 2024.

[95] Ling-An Zeng, Guohong Huang, Gaojie Wu, and Wei-Shi Zheng. Light-t2m: A lightweight and fast model for text-to-motion generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 9797–9805, 2025.

[96] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4230–4239, 2023.

[97] Alexander Denker, Francisco Vargas, Shreyas Padhy, Kieran Didi, Simon Mathis, Riccardo Barbano, Vincent Dutordoir, Emile Mathieu, Urszula Julia Komorowska, and Pietro Lio. Deft: Efficient fine-tuning of diffusion models by learning the generalised h-transform. *Advances in Neural Information Processing Systems*, 37:19636–19682, 2024.

[98] Zhixing Zhong, Junchen Hou, Zhixian Yao, Lei Dong, Feng Liu, Junqiu Yue, Tiantian Wu, Junhua Zheng, Gaoliang Ouyang, Chaoyong Yang, et al. Domain generalization enables general cancer cell annotation in single-cell and spatial transcriptomics. *Nature Communications*, 15(1):1929, 2024.

[99] Xiaoxu Guo, Fanghe Lin, Chuanyou Yi, Juan Song, Di Sun, Li Lin, Zhixing Zhong, Zhaorun Wu, Xiaoyu Wang, Yingkun Zhang, et al. Deep transfer learning enables lesion tracing of circulating tumor cells. *Nature Communications*, 13(1):7687, 2022.

[100] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15762–15772, 2024.

[101] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[102] Team Seawead, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.

APPENDIX

Contents

A Theoretical Analysis

A.1 Basic Theorems and Corollaries Related to Statistics

We start to provide some basic theoretical results, laying the foundation for establishing the bounds of the statistics in Appendix A.6 and A.7.

Theorem 2. *Let $X \sim \chi^2(d, \varphi)$ follow a **noncentral** chi-squared distribution with d degrees of freedom and noncentrality parameter φ. For any $t > 0$, the following tail bounds hold:*

$$P\left\{X - (d+\varphi) \geq 2\sqrt{(d+2\varphi)t} + 2t\right\} \leq e^{-t},$$
$$P\left\{X - (d+\varphi) \leq -2\sqrt{(d+2\varphi)t}\right\} \leq e^{-t}.$$

Proof. The moment-generating function of X satisfies

$$E[e^{sX}] = \frac{e^{\frac{\varphi s}{1-2s}}}{(1-2s)^{d/2}} \quad (s < 1/2).$$

The log-moment generating function of $X - (d+\varphi)$ is:

$$\begin{aligned}\log \mathbb{E}[e^{s(X-(d+\varphi))}] &= \log \mathbb{E}[e^{sX}] - s(d+\varphi) \\ &= -\frac{d}{2}\log(1-2s) + \frac{\varphi s}{1-2s} - s(d+\varphi). \end{aligned} \tag{14}$$

For $0 < s < 1/2$, we have

$$-s - \frac{1}{2}\log(1-2s) \leq \frac{s^2}{1-2s}, \tag{15}$$

which holds because the function $\psi(s) = -s - \frac{1}{2}\log(1-2s) - \frac{s^2}{1-2s}$ satisfies $\psi'(s) = -1 + \frac{1}{1-2s} - \frac{2s-s^2}{(1-2s)^2} = -\frac{2s^2}{(1-2s)^2} \leq 0$, implying $\max_{0<s<1/2}\psi(s) < \psi(0^+) = 0$, *i.e.*, $\psi(s) \leq 0$.

Substituting Eqn. (15) into Eqn.(14), we get

$$\log \mathbb{E}[e^{s(X-(d+\varphi))}] \leq \frac{ds^2}{1-2s} + \frac{2\varphi s^2}{1-2s} = \frac{(d+2\varphi)s^2}{1-2s}. \tag{16}$$

According to the result in [68], if $\exists\, v, c > 0$, s.t.$\log \mathbb{E}[e^{uZ}] \leq \frac{vu^2}{2(1-cu)}$, then for $\forall\, t > 0$, the following inequality holds:

$$P(Z \geq ct + \sqrt{2vt}) \leq e^{-t}.$$

Applying this result to Eqn. (16), we set $Z = X - (d+\varphi)$, $v = 2(d+2\varphi)$ and $c = 2$, then

$$P\left\{X - (d+\varphi) \geq 2\sqrt{(d+2\varphi)t} + 2t\right\} \leq e^{-t}.$$

For $-1/2 < s < 0$, we have

$$-s - \frac{1}{2}\log(1-2s) \leq s^2, \tag{17}$$

which holds because the function $h(s) = -s - \frac{1}{2}\log(1-2s) - s^2$ satisfies $h'(s) = -1 + \frac{1}{1-2s} - 2s = \frac{4s^2}{1-2s} \geq 0$, implying $\max_{-1/2<s<0}\psi(s) < \psi(0^-) = 0$, *i.e.*, $h(s) \leq 0$.

Substituting Eqn. (17) into Eqn.(14), we get

$$\log \mathbb{E}[e^{s(X-(d+\varphi))}] \leq ds^2 + \frac{2\varphi s^2}{1-2s} = (d + \frac{2\varphi}{1-2s})s^2 \leq (d+2\varphi)s^2. \tag{18}$$

According to the result in [68], if $\exists\, v > 0$, s.t., $\log \mathbb{E}[e^{sZ}] \leq \frac{vs^2}{2}$, then for $\forall\, t > 0$, the following inequality holds:

$$P\left(Z \leq -\sqrt{2vt}\right) \leq e^{-t}.$$

Applying this result to Eqn. (18), we set $Z = X - (d+\varphi)$, $v = 2(d+2\varphi)$, then

$$P\left\{X - (d+\varphi) \leq -2\sqrt{(d+2\varphi)t}\right\} \leq e^{-t}.$$

□

Corollary 1. *Given $X \sim \chi^2(d, \varphi)$, a noncentralchi-squared distribution with d degrees of freedom and the noncentrality parameter φ, with probability at least $1-\delta$, we have*

$$|X| \leq d + \varphi + \sqrt{4(d+2\varphi)\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right).$$

Proof. By Theorem 2, setting $e^{-t} = \frac{\delta}{2}$ yields the following inequalities:

$$P\left\{X - (d+\varphi) \geq 2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)\right\} \leq \frac{\delta}{2},$$

$$P\left\{X - (d+\varphi) \leq -2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}\right\} \leq \frac{\delta}{2}.$$

Combining these two inequalities, we obtain:

$$P\left\{X \geq d+\varphi+2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}+2\log\left(\frac{2}{\delta}\right) \text{ or } X \leq d+\varphi-2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}\right\} \leq \delta.$$

Taking the complement of the above event, we have:

$$P\left\{d+\varphi-2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)} \leq X \leq d+\varphi+2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}+2\log\left(\frac{2}{\delta}\right)\right\} \geq 1-\delta.$$

By relaxing the lower bound of X, we conclude

$$P\left\{|X| \leq d+\varphi+2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}+2\log\left(\frac{2}{\delta}\right)\right\} \geq 1-\delta.$$

□

A.2 Proof of Proposition 1

Proposition 1. *Under the brightness constancy assumption $p(\mathbf{x}+\Delta\mathbf{x}, t+\Delta t) \approx p(\mathbf{x}, t)$ with small inter-frame motion ($\Delta t \to 0$) and inter-frame displacement ($\Delta\mathbf{x} \to 0$), we have*

$$\partial_t \log p(\mathbf{x}, t) \approx -\frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \cdot \Delta\mathbf{x}}{\Delta t}. \tag{19}$$

Proof. We apply the Taylor expansion $\log p(\mathbf{x}+\Delta\mathbf{x}, t+\Delta t)$ around $(\mathbf{x}, t)$ to first order:

$$\log p(\mathbf{x}+\Delta\mathbf{x}, t+\Delta t) = \log p(\mathbf{x}, t) + \nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \cdot \Delta\mathbf{x} + \partial_t \log p(\mathbf{x}, t) \cdot \Delta t + o(\|\Delta\mathbf{x}\|^2 + \Delta t^2),$$

where $o(\|\Delta\mathbf{x}\|^2 + \Delta t^2)$ represents higher-order infinitesimal terms.

By assumption, $\log p(\mathbf{x}+\Delta\mathbf{x}, t+\Delta t) \approx \log p(\mathbf{x}, t)$. Subtracting $\log p(\mathbf{x}, t)$ from both sides:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \cdot \Delta\mathbf{x} + \partial_t \log p(\mathbf{x}, t) \cdot \Delta t + o(\|\Delta\mathbf{x}\|^2 + \Delta t^2) \approx 0.$$

Under $\Delta t \to 0$ and $\Delta\mathbf{x} \to 0$, we obtain:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \cdot \Delta\mathbf{x} + \partial_t \log p(\mathbf{x}, t) \cdot \Delta t \approx 0.$$

Rearranging terms gives:

$$\partial_t \log p(\mathbf{x}, t) \approx -\frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) \cdot \Delta\mathbf{x}}{\Delta t}.$$

□

A.3 Derivations of Chain Rule to the Conservation of Probability Mass

For $\frac{\partial p}{\partial t} + \nabla_{\mathbf{x}} \cdot \mathbf{J} = 0$, we substitute $\mathbf{J} = p\mathbf{v}$ and then divide the entire equation by p (which is strictly positive everywhere in its support), yielding:

$$\frac{1}{p}\partial_t p + \frac{1}{p}\nabla_{\mathbf{x}} \cdot (p\mathbf{v}) = 0.$$

Applying the vector calculus product rule $\nabla_{\mathbf{x}} \cdot (p\mathbf{v}) = p(\nabla_{\mathbf{x}} \cdot \mathbf{v}) + \mathbf{v} \cdot (\nabla_{\mathbf{x}} p)$ and the chain rule of calculus, $\frac{1}{p}\frac{\partial p}{\partial t} = \partial_t \log p$ and $\frac{\nabla_{\mathbf{x}} p}{p} = \nabla_{\mathbf{x}} \log p$, we obtain Eqn.(3):

$$\partial_t \log p + \nabla_{\mathbf{x}} \cdot \mathbf{v} + \mathbf{v} \cdot \nabla_{\mathbf{x}} \log p = 0.$$

This transformation does not alter the underlying fluid constraint—it is a variable change making explicit how velocity couples to log-density's temporal and spatial gradients.

A.4 Derivations of Gradients in NSG

In the following, we provide the specific forms of the terms $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$, $-\partial_t \log p(\mathbf{x}, t) + \lambda$ and $\mathbf{g}(\mathbf{x}, t)$ when the data are from Gaussian distributions, which will be used in Appendix A.5, A.6, A.7.

Proposition 3. *Given the real video distribution $p(\mathbf{x}, t) = \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and the generated video distribution $q(\mathbf{y}, t)$=$\mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$ with $\boldsymbol{\mu}$≠$\mathbf{0}$ and $\sigma(t)$≠$\mathbf{0}$, the gradients $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ and $\nabla_{\mathbf{y}} \log p(\mathbf{y}, t)$ are:*

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) = -\frac{\mathbf{x}}{\sigma(t)^2} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\sigma(t)^2}\mathbf{I}_d),$$
$$\nabla_{\mathbf{y}} \log p(\mathbf{y}, t) = -\frac{\mathbf{y}}{\sigma(t)^2} \sim \mathcal{N}(-\frac{\boldsymbol{\mu}}{\sigma(t)}, \sigma(t)^2\mathbf{I}_d).$$

Proof. Recall that for a Gaussian distribution $p(\mathbf{z}) = \mathcal{N}(\boldsymbol{\nu}, \sigma^2\mathbf{I}_d)$, the probability density function is

$$p(\mathbf{z}) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|\mathbf{z} - \boldsymbol{\nu}\|^2}{2\sigma^2}\right).$$

The log-density is

$$\log p(\mathbf{z}) = -\frac{d}{2}\log(2\pi\sigma^2) - \frac{\|\mathbf{z} - \boldsymbol{\nu}\|^2}{2\sigma^2}.$$

Thus, we have

$$\nabla_{\mathbf{z}} \log p(\mathbf{z}) = -\frac{\mathbf{z} - \boldsymbol{\nu}}{\sigma^2}.$$

For the real video distribution $p(\mathbf{x}, t) = \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$, we have $\boldsymbol{\nu} = \mathbf{0}$. Taking the gradient w.r.t. $\mathbf{x}$:

$$\nabla_{\mathbf{x}} \log p(\mathbf{x}, t) = \nabla_{\mathbf{x}}\left(-\frac{\|\mathbf{x}\|^2}{2\sigma(t)^2}\right) = -\frac{\mathbf{x}}{\sigma(t)^2} \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{\sigma(t)^2}\mathbf{I}_d\right).$$

For the generated video distribution $q(\mathbf{y}, t) = \mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$, evaluated under $p(\mathbf{y}, t)$), the gradient w.r.t. $\mathbf{y}$ is

$$\nabla_{\mathbf{y}} \log p(\mathbf{y}, t) = -\frac{\mathbf{y}}{\sigma(t)^2} \sim \mathcal{N}\left(-\frac{\boldsymbol{\mu}}{\sigma(t)^2}, \frac{1}{\sigma(t)^2}\mathbf{I}_d\right).$$

□

Proposition 4. *Given the real video distribution $p(\mathbf{x}, t) = \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and the generated video distribution $q(\mathbf{y}, t)$=$\mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$ with $\boldsymbol{\mu}$≠$\mathbf{0}$ and $\sigma(t)$≠$\mathbf{0}$, the partial derivatives $-\partial_t \log p(\mathbf{x}, t)$ and $-\partial_t \log p(\mathbf{y}, t)$ are:*

$$-\partial_t \log p(\mathbf{x}, t) = \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\|\mathbf{x}\|^2\dot{\sigma}(t)}{\sigma(t)^3} \sim \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d),$$
$$-\partial_t \log p(\mathbf{y}, t) = \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\|\mathbf{y}\|^2\dot{\sigma}(t)}{\sigma(t)^3} \sim \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d, \varphi),$$

where $\dot{\sigma}(t) \triangleq \frac{d}{dt}\sigma(t)$ and $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$. Here, $\chi^2(d)$ is the central chi-squared distribution and $\chi^2(d, \varphi)$ is the noncentral chi-squared distribution with noncentrality parameter φ [61].

Proof. We first derive the expression for the real video distribution $p(\mathbf{x}, t)$. The log-density is

$$\log p(\mathbf{x}, t) = -\frac{d}{2}\log(2\pi\sigma(t)^2) - \frac{\|\mathbf{x}\|^2}{2\sigma(t)^2}.$$

Taking the time derivative ∂_t (denoted by dot notation):

$$\begin{aligned}\partial_t \log p(\mathbf{x}, t) &= -\frac{d}{2}\cdot\frac{1}{2\pi\sigma(t)^2}\cdot 2\pi\cdot 2\sigma(t)\dot{\sigma}(t) + \frac{\|\mathbf{x}\|^2}{\sigma(t)^3}\dot{\sigma}(t)\\ &= -\frac{d\dot{\sigma}(t)}{\sigma(t)} + \frac{\|\mathbf{x}\|^2\dot{\sigma}(t)}{\sigma(t)^3}.\end{aligned}$$

Thus, we get

$$-\partial_t \log p(\mathbf{x}, t) = \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\|\mathbf{x}\|^2\dot{\sigma}(t)}{\sigma(t)^3} \sim \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d),$$

where the last formula is based on $\|\frac{\mathbf{x}}{\sigma(t)}\|^2 \sim \chi^2(d)$.

For the generated video distribution $q(\mathbf{y}, t) = \mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$, under $p(\mathbf{y}, t)$ with $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$, we have

$$-\partial_t \log p(\mathbf{y}, t) = \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\|\mathbf{y}\|^2\dot{\sigma}(t)}{\sigma(t)^3} \sim \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d, \varphi),$$

where the last formula is based on $\|\frac{\mathbf{y}}{\sigma(t)}\|^2 \sim \chi^2(d, \varphi)$. □

A.5 Proof of Proposition 2

Proposition 2. *Let the real video distribution be $p(\mathbf{x}, t) = \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and the generated video distribution be $q(\mathbf{y}, t)$=$\mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$, respectively, where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is an identity matrix and $\boldsymbol{\mu}\neq\mathbf{0} \in \mathbb{R}^d$ is the distribution shift and $\sigma(t)\neq\mathbf{0}$, the NSG $\mathbf{g}(\mathbf{x}, t)$ and $\mathbf{g}(\mathbf{y}, t)$ satisfy:*

$$\mathbf{g}(\mathbf{x}, t) = -\frac{\mathbf{x}/\sigma(t)^2}{D_r(\mathbf{x})},\ -\frac{\mathbf{x}}{\sigma(t)^2} \sim \mathcal{N}\Big(\mathbf{0}, \sigma(t)^2\mathbf{I}_d\Big),\ D_r(\mathbf{x}) \sim \lambda + \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d);$$

$$\mathbf{g}(\mathbf{y}, t) = -\frac{\mathbf{y}/\sigma(t)^2}{D_f(\mathbf{y})},\ -\frac{\mathbf{y}}{\sigma(t)^2} \sim \mathcal{N}\Big(-\frac{\boldsymbol{\mu}}{\sigma(t)}, \sigma(t)^2\mathbf{I}_d\Big),\ D_f(\mathbf{y}) \sim \lambda + \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\dot{\sigma}(t)}{\sigma(t)}\chi^2(d, \varphi),$$

where $D_r(\mathbf{x}) = \lambda + \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\|\mathbf{x}\|^2\dot{\sigma}(t)}{\sigma(t)^3}$, $D_f(\mathbf{y}) = \lambda + \frac{d\dot{\sigma}(t)}{\sigma(t)} - \frac{\|\mathbf{y}\|^2\dot{\sigma}(t)}{\sigma(t)^3}$, $\dot{\sigma}(t) \triangleq \frac{d}{dt}\sigma(t)$, and $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$, $\chi^2(d)$ is the central chi-squared distribution with d degrees of freedom and $\chi^2(d, \varphi)$ is the noncentral chi-squared distribution with noncentrality parameter φ and d degrees of freedom [61].

Proof. According to the definition of NSG,

$$\mathbf{g}(\mathbf{x}, t) = \frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)}{-\partial_t \log p(\mathbf{x}, t) + \lambda}, \tag{20}$$

we can substitute the results of $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ in Proposition 3 and $-\partial_t \log p(\mathbf{x}, t)$ in Proposition 4 into Eqn. (20) and directly contribute to the results. □

A.6 Derivations of Upper Bounds for Gradients

Next, we present some propositions on the upper bounds that will be used in Appendix A.7.

Proposition 5. *Given the real video distribution $p(\mathbf{x}, t) = \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and the generated video distribution $q(\mathbf{y}, t)$=$\mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$ with $\boldsymbol{\mu}\neq\mathbf{0}$ and $\sigma(t)\neq\mathbf{0}$, let $D_r(\mathbf{x}) = \lambda - \partial_t \log p(\mathbf{x}, t)$ and $D_f(\mathbf{y}) = \lambda - \partial_t \log p(\mathbf{y}, t)$, with probability at least $1-\delta$, we have*

$$|D_r(\mathbf{x}) - D_f(\mathbf{y})| \leq \varphi + 2\sqrt{(d + 2\varphi)\log\frac{4}{\delta}} + 2\sqrt{d\log\frac{4}{\delta}} + 2\log\frac{4}{\delta},$$

where $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$.

Proof. From Proposition 4, we obtain

$$|D_r(\mathbf{x}) - D_f(\mathbf{y})| = \frac{|\dot{\sigma}(t)|}{\sigma(t)^3} \left| \|\mathbf{x}\|^2 - \|\mathbf{y}\|^2 \right|.$$

where $\dot{\sigma}(t) \triangleq \frac{d}{dt}\sigma(t)$.

Let $Z = \frac{\|\mathbf{x}\|^2}{\sigma(t)^2} \sim \chi^2(d)$ and $W = \frac{\|\mathbf{y}\|^2}{\sigma(t)^2} \sim \chi^2(d, \varphi)$, where $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$. The difference becomes

$$|D_r(\mathbf{x}) - D_f(\mathbf{y})| = \frac{|\dot{\sigma}(t)|}{\sigma(t)} |W - Z|.$$

To bound $|W - Z|$, we use concentration inequalities for chi-squared distributions by Theorem 2. For $Z \sim \chi^2(d)$, we have

$$P\left\{ Z - d \geq 2\sqrt{d \log \tfrac{4}{\delta}} + 2 \log \tfrac{4}{\delta} \right\} \leq \frac{\delta}{4},$$
$$P\left\{ Z - d \leq -2\sqrt{d \log \tfrac{4}{\delta}} \right\} \leq \frac{\delta}{4}.$$

Combining these two events, we obtain

$$P\left\{ -2\sqrt{d \log \tfrac{4}{\delta}} \leq Z - d \leq 2\sqrt{d \log \tfrac{4}{\delta}} + 2 \log \tfrac{4}{\delta} \right\} \geq 1 - \frac{\delta}{2}. \tag{21}$$

Similarly, for $W \sim \chi^2(d, \varphi)$, we have

$$P\left\{ \varphi - 2\sqrt{(d + 2\varphi) \log \tfrac{4}{\delta}} \leq W - d \leq \varphi + 2\sqrt{(d + 2\varphi) \log \tfrac{4}{\delta}} + 2 \log \tfrac{4}{\delta} \right\} \geq 1 - \frac{\delta}{2}. \tag{22}$$

Combining the bounds in Eqn. (22) and (21), we have with probability $1 - \delta$:

$$|W - Z| \leq \varphi + 2\sqrt{(d + 2\varphi) \log \tfrac{4}{\delta}} + 2\sqrt{d \log \tfrac{4}{\delta}} + 2 \log \tfrac{4}{\delta}.$$

Substituting this into the expression for $|D_r(\mathbf{x}) - D_f(\mathbf{y})|$ completes the proof. □

Proposition 6. *Given the real video distribution $p(\mathbf{x}, t) = \mathcal{N}(\mathbf{0}, \sigma(t)^2 \mathbf{I}_d)$ and the generated video distribution $q(\mathbf{y}, t) = \mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2 \mathbf{I}_d)$ with $\boldsymbol{\mu} \neq \mathbf{0}$ and $\sigma(t) \neq \mathbf{0}$, the following inequalities hold with probability at least $1 - \delta$:*

1) $$\frac{\|\mathbf{x}\|^2}{\sigma(t)^2} \leq d + \sqrt{4d \log\left(\frac{2}{\delta}\right)} + 2 \log\left(\frac{2}{\delta}\right),$$

2) $$\frac{\|\mathbf{y}\|^2}{\sigma(t)^2} \leq d + \varphi + \sqrt{4(d + 2\varphi) \log\left(\frac{2}{\delta}\right)} + 2 \log\left(\frac{2}{\delta}\right),$$

3) $$\frac{\|\mathbf{x} - \mathbf{y}\|^2}{2\sigma(t)^2} \leq d + \frac{\varphi}{2} + \sqrt{4(d + \varphi) \log\left(\frac{2}{\delta}\right)} + 2 \log\left(\frac{2}{\delta}\right),$$

where $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$.

Proof. 1) Since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma(t)^2 \mathbf{I}_d)$, $\frac{\|\mathbf{x}\|^2}{\sigma(t)^2}$ follows a central chi-squared distribution $\chi^2(d)$. By the concentration inequality for central chi-squared distributions (Corollary 1), with probability $1 - \delta$:

$$\frac{\|\mathbf{x}\|^2}{\sigma(t)^2} \leq d + \sqrt{4d \log\left(\frac{2}{\delta}\right)} + 2 \log\left(\frac{2}{\delta}\right).$$

2) For $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2 \mathbf{I}_d)$, $\frac{\|\mathbf{y}\|^2}{\sigma(t)^2}$ follows a noncentral chi-squared distribution $\chi^2(d, \varphi)$, where $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$. By Corollary 1, with probability $1 - \delta$, we have

$$\frac{\|\mathbf{y}\|^2}{\sigma(t)^2} \leq d + \varphi + \sqrt{4(d + 2\varphi) \log\left(\frac{2}{\delta}\right)} + 2 \log\left(\frac{2}{\delta}\right).$$

3) Since $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$, their difference $\mathbf{z}$ satisfies:

$$\mathbf{x} - \mathbf{y} \sim \mathcal{N}(-\boldsymbol{\mu}, 2\sigma(t)^2\mathbf{I}_d).$$

Thus, $\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma(t)^2}$ follows a noncentral chi-squared distribution $\chi^2(d, \varphi/2)$, where $\varphi/2 = \frac{\|\boldsymbol{\mu}\|^2}{2\sigma(t)^2}$.

By Corollary 1, with probability $1-\delta$, we have

$$\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma(t)^2} \leq d + \frac{\varphi}{2} + \sqrt{4\,(d+\varphi)\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right).$$

□

A.7 Proof of Theorem 1

Given a video $\mathbf{x} \in \mathbb{R}^{T\times d}$, its NSG Feature is $\mathbf{G}(\mathbf{x}) = \{\mathbf{g}(\mathbf{x}, t)\}_{t=1}^{T}$, where $\mathbf{g}(\mathbf{x}, t)$ is defined as:

$$\mathbf{g}(\mathbf{x}, t) = \frac{\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)}{-\partial_t \log p(\mathbf{x}, t) + \lambda}, \tag{23}$$

Here, $\lambda > 0$ and $p(\mathbf{x}, t)$ is the probability density of the real video parameterized by time t.

Note that Theorem 1 share a common lower bound C on both $D_r(\mathbf{x}){=}\lambda{-}\partial_t \log p(\mathbf{x}, t)$ and $D_f(\mathbf{y}){=}\lambda{-}\partial_t \log p(\mathbf{y}, t)$. We first derive the conditions for $D_r(\mathbf{x}){>}C{>}0$ and $D_f(\mathbf{y}){>}C{>}0$ to hold in Proposition 7. The same analytical approach can be extended to examine other cases.

Proposition 7. *Let the real video distribution be $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and the generated video distribution be $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$, respectively, where $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ is an identity matrix and $\boldsymbol{\mu} \neq \mathbf{0} \in \mathbb{R}^d$ is the distribution shift, we have $D_r(\mathbf{x}) > C > 0$ and $D_f(\mathbf{y}) > C > 0$ with probability at least $1-\delta$, provided C and λ meet the following conditions:*

1) Case 1 ($\frac{\dot\sigma(t)}{\sigma(t)} > 0$):

$$C = \lambda - \frac{\dot\sigma(t)}{\sigma(t)}\cdot\varphi + \frac{2\dot\sigma(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)} + \frac{2\dot\sigma(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right),$$

$$\lambda > \frac{\dot\sigma(t)}{\sigma(t)}(d+\varphi) - \frac{2\dot\sigma(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)} - \frac{2\dot\sigma(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right).$$

where $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$.

2) Case 2 ($\frac{\dot\sigma(t)}{\sigma(t)} < 0$):

$$C = \lambda - \frac{2\dot\sigma(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)},$$

$$\lambda > \frac{2\dot\sigma(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)}.$$

Proof. Let $W = \frac{\|\mathbf{y}\|^2}{\sigma(t)^2} \sim \chi^2(d, \varphi)$, where $\varphi = \frac{\|\boldsymbol{\mu}\|^2}{\sigma(t)^2}$. From Theorem 2, we have

$$P\left\{W - (d+\varphi) \geq 2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)} + 2\log\left(\frac{2}{\delta}\right)\right\} \leq \frac{\delta}{2}, \tag{24}$$

$$P\left\{W - (d+\varphi) \leq -2\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}\right\} \leq \frac{\delta}{2}. \tag{25}$$

1) **Case 1** ($\frac{\dot\sigma(t)}{\sigma(t)} > 0$):

Substituting $D_f(\mathbf{y}) = \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}(W-d)$ into the bound Eqn. (24):

$$P\left\{D_f(\mathbf{y}) \leq \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}\cdot\varphi + \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)} + \frac{2\dot{\sigma}(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right)\right\} \leq \frac{\delta}{2}.$$

Thus, the following inequalities hold with probability at least $1-\delta/2$:

$$D_f(\mathbf{y}) \geq \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}\cdot\varphi + \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)} + \frac{2\dot{\sigma}(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right),$$
$$D_r(\mathbf{x}) \geq \lambda + \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)} + \frac{2\dot{\sigma}(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right).$$

To ensure $D_r(\mathbf{x}) > C > 0$ and $D_f(\mathbf{y}) > C > 0$, we can select C and λ as:

$$C = \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}\cdot\varphi + \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)} + \frac{2\dot{\sigma}(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right),$$
$$\lambda > \frac{\dot{\sigma}(t)}{\sigma(t)}(d+\varphi) - \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)} - \frac{2\dot{\sigma}(t)}{\sigma(t)}\log\left(\frac{2}{\delta}\right).$$

2) **Case 2** ($\frac{\dot{\sigma}(t)}{\sigma(t)} < 0$):

Substituting $D_f(\mathbf{y}) = \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}(W-d)$ into the bound Eqn. (25):

$$P\left\{D_f(\mathbf{y}) \leq \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}\cdot\varphi - \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)}\right\} \leq \frac{\delta}{2}.$$

Thus, the following inequalities hold with probability at least $1-\delta/2$:

$$D_f(\mathbf{y}) \geq \lambda - \frac{\dot{\sigma}(t)}{\sigma(t)}\cdot\varphi - \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{(d+2\varphi)\log\left(\frac{2}{\delta}\right)},$$
$$D_r(\mathbf{x}) \geq \lambda - \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)}.$$

To ensure $D_r(\mathbf{x}) > C > 0$ and $D_f(\mathbf{y}) > C > 0$, we can select C and λ as:

$$C = \lambda - \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)},$$
$$\lambda > \frac{2\dot{\sigma}(t)}{\sigma(t)}\sqrt{d\log\left(\frac{2}{\delta}\right)}.$$

□

Building upon the established Propositions 2, 5, 6 and 7, we next prove Theorem 1.

Theorem 1. *Let the real video distribution be $\mathbf{x}\sim\mathcal{N}(\mathbf{0}, \sigma(t)^2\mathbf{I}_d)$ and the generated video distribution be $\mathbf{y}\sim\mathcal{N}(\boldsymbol{\mu}, \sigma(t)^2\mathbf{I}_d)$, respectively, where $\mathbf{I}_d\in\mathbb{R}^{d\times d}$ is an identity matrix and $\boldsymbol{\mu}\neq\mathbf{0}\in\mathbb{R}^d$ is the distribution shift. Given $\mathbf{G}(\mathbf{x})=\{\mathbf{g}(\mathbf{x},t)\}_{t=1}^T$, denote $\varphi=\|\boldsymbol{\mu}\|^2/\sigma(t)^2$ and assume $|-\partial_t\log p(\mathbf{x},t)+\lambda| \geq C > 0$ and $|-\partial_t\log p(\mathbf{y},t)+\lambda| \geq C > 0$, with probability at least $1-\delta$, we have*

$$\|\mathbf{G}(\mathbf{x})-\mathbf{G}(\mathbf{y})\|^2 \leq \mathcal{O}\left(\frac{T}{C^4\sigma(t)^2}\left[\varphi d + d^2 + \varphi + \log\frac{T}{\delta}\cdot(\varphi+d) + \log^2\frac{T}{\delta}\right]\right).$$

Proof. Based on the definition of $\mathbf{G}(\mathbf{x})$, we have

$$\|\mathbf{G}(\mathbf{x}) - \mathbf{G}(\mathbf{y})\|^2 = \sum_{t=1}^{T} \|\mathbf{g}(\mathbf{x}, t) - \mathbf{g}(\mathbf{y}, t)\|^2 . \tag{26}$$

Next, we focus on the bound of $\mathbf{g}(\mathbf{x}, t) - \mathbf{g}(\mathbf{y}, t)$.

Let $D_r(\mathbf{x}) = \lambda - \partial_t \log p(\mathbf{x}, t)$ and $D_f(\mathbf{y}) = \lambda - \partial_t \log p(\mathbf{y}, t)$, by Propositions 3 and 4, we have

$$\begin{aligned}
&\|\mathbf{g}(\mathbf{x}, t) - \mathbf{g}(\mathbf{y}, t)\|^2 \\
=&\left\| \frac{\mathbf{x}/\sigma(t)^2}{D_r(\mathbf{x})} - \frac{\mathbf{y}/\sigma(t)^2}{D_f(\mathbf{y})} \right\|^2 \\
=&\left\| \frac{\mathbf{x}/\sigma(t)^2}{D_r(\mathbf{x})} - \frac{\mathbf{x}/\sigma(t)^2}{D_f(\mathbf{y})} + \frac{\mathbf{x}/\sigma(t)^2}{D_f(\mathbf{y})} - \frac{\mathbf{y}/\sigma(t)^2}{D_f(\mathbf{y})} \right\|^2 \\
\leq& 2\left\| \frac{\mathbf{x}/\sigma(t)^2}{D_r(\mathbf{x})} - \frac{\mathbf{x}/\sigma(t)^2}{D_f(\mathbf{y})} \right\|^2 + 2\left\| \frac{\mathbf{x}/\sigma(t)^2}{D_f(\mathbf{y})} - \frac{\mathbf{y}/\sigma(t)^2}{D_f(\mathbf{y})} \right\|^2 \\
=& 2\left(\frac{1}{D_r(\mathbf{x})} - \frac{1}{D_f(\mathbf{y})} \right)^2 \cdot \frac{\|\mathbf{x}\|^2}{\sigma(t)^4} + 2\left(\frac{1}{D_f(\mathbf{y})} \right)^2 \cdot \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma(t)^4} \\
=& 2\left(\frac{D_f(\mathbf{y}) - D_r(\mathbf{x})}{D_r(\mathbf{x}) D_f(\mathbf{y})} \right)^2 \cdot \frac{\|\mathbf{x}\|^2}{\sigma(t)^4} + 2\left(\frac{1}{D_f(\mathbf{y})} \right)^2 \cdot \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma(t)^4} \\
\leq& 2\frac{|D_f(\mathbf{y}) - D_r(\mathbf{x})|^2}{C^4} \cdot \frac{\|\mathbf{x}\|^2}{\sigma(t)^4} + \frac{2}{C^2} \cdot \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma(t)^4}.
\end{aligned} \tag{27}$$

For the first term in Eqn. (27), according to Proposition 5, with probability at least $1 - 2\delta/3$, we have

$$\begin{aligned}
&2\frac{|D_f(\mathbf{y}) - D_r(\mathbf{x})|^2}{C^4} \cdot \frac{\|\mathbf{x}\|^2}{\sigma(t)^4} \\
\leq& \frac{2}{C^4\sigma(t)^2} \left(\varphi + 2\sqrt{(d + 2\varphi)\log\tfrac{12}{\delta}} + 2\sqrt{d\log\tfrac{12}{\delta}} + 2\log\tfrac{12}{\delta} \right) \cdot \left(d + \sqrt{4d\log\left(\tfrac{6}{\delta}\right)} + 2\log\left(\tfrac{6}{\delta}\right) \right) \\
\leq& \frac{2}{C^4\sigma(t)^2} \left(\varphi + 4\sqrt{(d + 2\varphi)\log\tfrac{12}{\delta}} + 2\log\tfrac{12}{\delta} \right) \cdot \left(d + \sqrt{4d\log\left(\tfrac{6}{\delta}\right)} + 2\log\left(\tfrac{6}{\delta}\right) \right).
\end{aligned} \tag{28}$$

For the second term in Eqn. (27), applying Proposition 6, with probability at least $1 - \delta/3$, we have

$$\frac{2}{C^2} \cdot \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma(t)^4} \leq \frac{4}{C^2\sigma(t)^2} \left(d + \frac{\varphi}{2} + \sqrt{4(d + \varphi)\log\left(\frac{6}{\delta}\right)} + 2\log\left(\frac{6}{\delta}\right) \right). \tag{29}$$

For simplicity, let $L = \log\frac{12}{\delta}$. Since $\log\frac{6}{\delta} = L - \log 2 < L$, we have

$$\begin{aligned}
2\frac{|D_f(\mathbf{y}) - D_r(\mathbf{x})|^2}{C^4} \cdot \frac{\|\mathbf{x}\|^2}{\sigma(t)^4} &\leq \frac{2}{C^4\sigma(t)^2} \left(\varphi + 4\sqrt{(d + 2\varphi)L} + 2L \right) \cdot \left(d + 2\sqrt{dL} + 2L \right) \\
&\leq \frac{2}{C^4\sigma(t)^2} \left(\varphi + 2(d + 2\varphi) + L + 2L \right) \cdot (d + d + L + 2L) \\
&= \frac{2}{C^4\sigma(t)^2} (5\varphi + 2d + 3L)(2d + 3L) \\
&= \frac{2}{C^4\sigma(t)^2} \left(10\varphi d + 4d^2 + 3L \cdot (5\varphi + 4d) + 9L^2 \right).
\end{aligned} \tag{30}$$

$$\begin{aligned}
\frac{2}{C^2} \cdot \frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma(t)^4} &\leq \frac{4}{C^2\sigma(t)^2} \left(d + \frac{\varphi}{2} + 2\sqrt{(d + \varphi)L} + 2L \right) \\
&\leq \frac{4}{C^2\sigma(t)^2} \left(d + \frac{\varphi}{2} + (d + \varphi)L + 2L \right) \\
&= \frac{4}{C^2\sigma(t)^2} \left(d + \frac{\varphi}{2} + L(d + \varphi + 2) \right).
\end{aligned} \tag{31}$$

Combining Eqn. (30) and (31) and (27), and substituting $L = \log \frac{12}{\delta}$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\|\mathbf{g}(\mathbf{x},t)-\mathbf{g}(\mathbf{y},t)\|^2 \\
\leq &\frac{2}{C^4\sigma(t)^2}\left(10\varphi d+4d^2+3L\cdot(5\varphi+4d)+9L^2\right)+\frac{4}{C^2\sigma(t)^2}\left(d+\frac{\varphi}{2}+L(d+\varphi+2)\right).
\end{aligned}$$

For further simplicity, note that $\frac{1}{C^2} \leq \frac{1}{C^4}$, we obtain

$$\begin{aligned}
&\|\mathbf{g}(\mathbf{x},t)-\mathbf{g}(\mathbf{y},t)\|^2 \\
\leq &\frac{2}{C^4\sigma(t)^2}\left(10\varphi d+4d^2+2d+\varphi+L\cdot(17\varphi+14d+4)+9L^2\right) \\
= &\frac{2}{C^4\sigma(t)^2}\left(10\varphi d+4d^2+2d+\varphi+\log\frac{12}{\delta}\cdot(17\varphi+14d+4)+9\log^2\frac{12}{\delta}\right). \qquad (32)
\end{aligned}$$

Summing over time steps $t=1,\ldots,T$, and replacing δ in Eqn. (32) into δ/T , we get:

$$\|\mathbf{G}(\mathbf{x})-\mathbf{G}(\mathbf{y})\|^2 \leq \frac{2T}{C^4\sigma(t)^2}\left(10\varphi d+4d^2+2d+\varphi+\log\frac{12T}{\delta}\cdot(17\varphi+14d+4)+9\log^2\frac{12T}{\delta}\right).$$

Thus, we obtain the final results

$$\|\mathbf{G}(\mathbf{x})-\mathbf{G}(\mathbf{y})\|^2 \leq \mathcal{O}\left(\frac{T}{C^4\sigma(t)^2}\left[\varphi d+d^2+\varphi+\log\frac{T}{\delta}\cdot(\varphi+d)+\log^2\frac{T}{\delta}\right]\right).$$

□

B More Related Work

Maximum Mean Discrepancy (MMD). Maximum Mean Discrepancy (MMD) is an effective metric for two-sample testing to assess whether two samples originate from the same distribution [69, 33, 70, 34, 71, 72, 73]. Originally introduced by Müller et al. [72] as an instance of an integral probability metric, MMD admits several sample-based estimators. Particularly, Gretton et al. [33] introduce the U-statistic estimator, which is unbiased for the squared MMD and achieves near-minimal variance among all unbiased alternatives. Further, Tolstikhin et al. [74] derive lower bounds on the estimation error of MMD under finite samples when employing a radial universal kernel.

Building upon the traditional formulation of MMD, recent advancements incorporate learnable kernels to enhance its discriminative capability. Liu et al. [34] develop a data-splitting strategy for kernel optimization and selection, effectively addressing the kernel adaptation challenges for complex-data scenarios. Kim et al. [75] propose an adaptive two-sample test designed for comparing two Hölder densities supported on the d-dimensional unit ball. In addition, Zhang et al. [48] introduce MMD-MP, a multi-population aware optimization framework to further improve the stability of kernel-based MMD training. At present, MMD has been extensively applied to distributional measurement and discrepancy detection tasks across both textual and visual modalities [45, 48, 46].

C More Details for Experiment Settings

C.1 More Details on Datasets

GenVideo [25] is a large-scale benchmark for AI-generated video detection, comprising 1.22 million real videos and 1.05 million AI-generated videos. The real video collection aggregates content from three established datasets: MSR-VTT (web video clips) [65], Kinetics-400 (human action videos) [62], and Youku-mPLUG (diverse online videos captured from Youku.com) [76]. The AI-generated portion features videos produced by 19 distinct generation models, including both open-source implementations (ZeroScope [77], I2VGen-XL [78], SVD [1], VideoCrafter [7], DynamiCrafter [8], Stable Diffusion(SD) [79], SEINE [63], Latte [80], OpenSora [81], ModelScope [82], HotShot [83], Show-1 [84], Gen2 [85], Crafter [86], Lavie [87]) and commercial closed-source systems (Pika, Sora, MoonValley, MorphStudio). This dataset spans multiple generation paradigms, including text-to-video and image-to-video synthesis. Throughout all experiments, we filter videos with less than 8 frames and only uniformly sample 8 frames for each video during training and testing.

C.2 More Details on Evaluation Metrics

Video generation detection is inherently a binary classification task. Here, we introduce four fundamental evaluation metrics of binary classification: **True Positive** (TP) means correctly predicted positive instances (ground truth is positive, prediction is positive).**True Negative** (TN) means correctly predicted negative instances (ground truth is negative, prediction is negative). **False Positive** (FP) means incorrectly predicted positive instances (ground truth is negative, prediction is positive). **False Negative** (FN) means incorrectly predicted negative instances (ground truth is positive, prediction is negative).

AUROC denotes the Area Under the Receiver Operating Characteristic Curve [67, 88], which is a widely used statistic for assessing the discriminatory capacity of distribution models. Formally,

$$AUROC = \int TP(t)FP(t)dt,$$

where $TP(t) = TP(t)/(TP(t) + FN(t))$ is the true positive rate and $FP(t) = FP(t)/(FP(t) + TN(t))$ is false positive rate with a threshed t.

Accuracy (ACC) measures the model's overall correctness in classification by calculating the ratio of correctly predicted instances (both true positive and true negative) to the total instances.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}.$$

Recall [66] evaluates the model's ability to identify all relevant instances of a class, measuring the proportion of true positives among all actual positive instances. It emphasizes minimizing false

negatives, ensuring comprehensive coverage of positive cases.

$$\text{Recall} = \frac{TP}{TP + FN}.$$

Precision [66] quantifies the model's capability to avoid false positives by calculating the proportion of true positives among all predicted positive instances. It ensures reliability in positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

F1-score [66] balances Precision and Recall using their harmonic mean, providing a robust metric for scenarios with imbalanced class distributions. It penalizes extreme biases toward either precision or recall.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

C.3 Implementation Details on NSG-VD

In our NSG-VD, we employ the pre-trained diffusion model s_θ of Guided Diffusion using the 256 × 256 unconditional checkpoint from the guided-diffusion library [3] following [56]. For a given video $\mathbf{x}$ at t-th frame, we compute its score feature $\nabla_{\mathbf{x}} \log p(\mathbf{x}, t)$ by diffusing $\mathbf{x}_t$ at diffusion timestep $5/1,000$ and passing it through s_θ. For the deep kernel $\phi_{\mathbf{G}}$, we employ a single-layer of Swin transformer [89], mapping input features of dimension $8 \times 224 \times 224$ to a 300-dimensional output.

We conduct our experiments on a server with 1× NVIDIA RTX 3090 GPU using Python 3.10.17 and Pytorch 2.7.0. We use Adam optimizer [90] to optimize the kernel parameters ω with batchsize 24, learning rate 0.0001, weight decay 0.1, $\sigma_\phi = 0.1$ and $\sigma_\Phi = 100$. For the testing, we set the decision threshold $\tau = 1$ in Eqn. (11). The overall algorithms for training and testing are in Alg. 1 and 2.

C.4 Pseudo Code of NSG-VD

Algorithm 1 Training deep kernel of MMD

Input: Real and generated videos $S_{\mathbb{P}}^{tr}$, $S_{\mathbb{Q}}^{tr}$;
$\omega \leftarrow \omega_0$; $\lambda \leftarrow 10^{-10}$; learning rate η;
for $r = 1, 2, \ldots, r_{max}$ **do**
  $k_\omega \leftarrow$ kernel function using Eqn. (12);
  $M(\omega) \leftarrow \widehat{\text{MPP}}_u(S_{\mathbb{P}}^{tr}, S_{\mathbb{Q}}^{tr}; k_\omega)$;
  $V_\lambda(\omega) \leftarrow \hat{\sigma}^2(S_{\mathbb{P}}^{tr}, S_{\mathbb{Q}}^{tr}; k_\omega)$ using Eqn. (13);
  $\hat{J}_\lambda(\omega) \leftarrow M(\omega)/\sqrt{V_\lambda(\omega)}$;
  $\omega \leftarrow \omega + \eta \nabla_{\text{Adam}} \hat{J}_\lambda(\omega)$;
end for
Output: k_ω^*

Algorithm 2 Detecting videos with NSG-VD

Input: Referenced videos $S_{\mathbb{P}}^{re}$, testing videos $S_{\mathbb{Q}}^{te}$; decision $f(\cdot)$; deep kernel k_ω; threshold τ;
for $\mathbf{x}_i$ in $S_{\mathbb{Q}}^{te}$ **do**
  $Q_i \leftarrow \widehat{\text{MMD}}_b^2(S_{\mathbb{P}}^{re}, \{\mathbf{x}_i\}; k_\omega)$ using Eqn. (10);
  $f(\mathbf{x}_i; S_{\mathbb{P}}^{re}, k_\omega, \tau) = \mathbb{I}(Q_i > \tau)$;
  Obtaining $f(\mathbf{x}_i)$ using Eqn. (11);
end for
Output: Predictions $\{f(\mathbf{x}_i)\}$ of the testing set

[3] https://github.com/openai/guided-diffusion

D More Experimental Results

D.1 More Results on Standard Evaluation

To demonstrate the statistical robustness and reproducibility of our proposed NSG-VD method, we report standard deviations of four metrics across 10 datasets with three different seeds (Table 5). The table shows that our NSG-VD achieves consistently high performance with minimal variance (*e.g.*, 0.41% for Recall, 0.87% for Accuracy), indicating strong reliability and repeatability of our methods.

Table 5: Standard deviations of NSG-VD with three different random seeds (%), where we train all models with $10,000$ real and generated videos from Kinetics-400 and SEINE, respectively.

Dataset	Recall	Accuracy	F1	AUROC
ModelScope	92.78 ± 0.48	84.31 ± 1.88	85.54 ± 1.54	92.49 ± 3.07
MorphStudio	99.44 ± 0.96	86.53 ± 2.64	88.10 ± 2.14	97.08 ± 0.92
MoonValley	100.00 ± 0.00	87.64 ± 0.64	89.00 ± 0.50	99.05 ± 0.30
HotShot	100.00 ± 0.00	88.06 ± 2.30	89.35 ± 1.83	98.64 ± 0.49
Show1	100.00 ± 0.00	88.89 ± 2.55	90.03 ± 2.08	96.96 ± 0.93
Gen2	98.61 ± 1.73	87.08 ± 2.20	88.43 ± 1.86	95.27 ± 2.17
Crafter	99.72 ± 0.48	88.89 ± 1.05	89.98 ± 0.86	98.07 ± 0.41
Lavie	98.61 ± 0.96	88.75 ± 2.17	89.77 ± 1.85	95.32 ± 2.92
Sora	94.05 ± 1.03	84.12 ± 1.97	83.85 ± 1.56	93.10 ± 1.36
WildScrape	91.67 ± 2.21	85.41 ± 2.60	86.29 ± 2.37	92.75 ± 0.31
Avg.	97.49 ± 0.41	86.97 ± 0.87	88.04 ± 0.74	95.87 ± 0.87

D.2 More Results on Impact of Spatial Gradients and Temporal Derivatives

To comprehensively analyze how spatial gradients and temporal derivatives contribute to detection performance across diverse generative paradigms, we include detailed results across 10 diverse generative paradigms in Table 6. The spatial gradients achieve strong performance across most generated models (*e.g.*, 81.67% Recall on ModelScope, 97.50% Recall on MorphStudio), with only minor performance gaps on models like HotShot (72.23% Recall) and Show1 (74.17% Recall).

In contrast, the temporal derivatives show complementary strengths and relatively better detection capabilities on these challenging cases, notably achieving 75.40% Recall on HotShot and 77.60% Recall on Show1, where temporal dynamics (*e.g.*, rapid motion transitions in HotShot dataset) may play a more pronounced role in exposing synthetic anomalies.

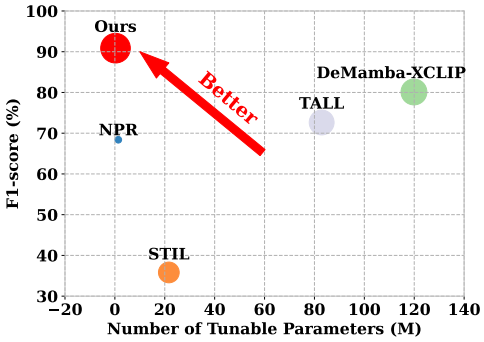
Notably, our proposed NSG-VD demonstrates superior reliability by integrating both components, achieving an average F1-score of 90.87%, a significant improvement over individual features. This highlights the complementary nature of spatiotemporal dynamics modeling, enabling consistent detection performance even when individual cues exhibit dataset-specific limitations.

Table 6: Impact of spatial gradients and temporal derivatives across 10 generated models (%), where we train all models with $10,000$ real and generated videos from Kinetics-400 and Pika, respectively.

Method	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
Spatial Gradients	Recall	81.67	97.50	100.00	72.23	74.17	98.33	98.13	77.12	98.21	82.50	87.99
	Accuracy	77.50	89.58	87.08	75.00	76.25	87.92	88.75	76.67	88.39	81.25	82.84
	F1	78.40	90.35	88.56	74.36	75.74	89.06	89.73	76.86	89.43	81.48	83.40
	AUROC	87.32	98.43	99.99	78.76	82.72	98.98	98.64	84.15	99.01	90.52	91.85
Tmporal Derivative	Recall	52.60	40.20	58.40	75.40	77.60	49.00	72.40	47.20	60.71	70.00	60.35
	Accuracy	67.40	61.20	70.30	78.80	79.90	65.60	77.30	64.70	69.64	76.10	71.09
	F1	61.74	50.89	66.29	78.05	79.43	58.75	76.13	57.21	66.67	74.55	66.97
	AUROC	72.97	68.64	81.46	87.09	87.80	74.48	84.24	72.97	76.28	83.62	78.95
NSG-VD (Ours)	Recall	68.33	98.33	100.00	92.50	87.50	80.00	98.33	94.17	78.57	82.50	**88.02**
	Accuracy	81.67	98.33	96.67	91.67	90.83	88.33	95.83	94.17	88.39	88.75	**91.46**
	F1	78.85	98.33	96.77	91.74	90.52	87.27	95.93	94.17	87.13	88.00	**90.87**
	AUROC	92.26	98.66	98.15	94.45	96.38	94.83	98.16	97.41	96.40	94.73	**96.14**

E More Discussions on NSG-VD

E.1 Efficiency of NSG-VD



Method	All Params(M) ↓	Tun. Params(M) ↓	F1(%) ↑
DeMamba	119.89	119.89	80.12
TALL	82.89	82.89	72.63
STIL	21.63	21.63	35.82
NPR	**1.37**	1.37	68.39
NSG-VD (Ours)	527.45	**0.25**	**90.87**

Figure 4: Comparisons with baselines in terms of training costs and performance (%), where we train all models with $10,000$ real and generated videos from Kinetics-400 and Pika, respectively.

Performance vs. Training Cost Analysis. We compare the number of trainable parameters and detection performance of NSG-VD against state-of-the-art baselines. As shown in Figure 4, our NSG-VD achieves a 90.87% F1-score with only 0.25 M trainable parameters, demonstrating superior parameter efficiency compared to methods like DeMamba (80.12% F1-score, 119.89 M parameters) and TALL (72.63% F1-score, 82.89 M parameters). This demonstrates that NSG-VD's physics-driven design enables high accuracy through minimal parameter tuning.

Existing baselines struggle to balance parameter scale and performance. NPR achieves only 68.39% F1-score despite its minimal trainable parameters (1.37 M), while STIL requires 21.63 M parameters to attain 35.82% F1-score—a suboptimal trade-off compared to NSG-VD's superior performance with 100× fewer parameters. These results underscore the limitations of conventional artifact-driven frameworks in effective parameter budget utilization, further validating the importance of our physics-guided spatiotemporal modeling paradigm for AI-generated video detection.

Performance vs. Inference Time Analysis. We evaluate the efficiency of our NSG-VD by analyzing both detection performance and computational overhead under the same setting as Table 1. As shown in Table 7, our NSG-VD achieves superior detection performance (*e.g.*, 97.13% Recall, 87.45% F1-score) with a practical inference latency of $0.3605s$ per video, which remains viable for non-real-time applications (*e.g.*, judicial video evidence analysis) despite being slower than other baselines. This latency stems from our current implementation of pre-trained diffusion models for gradient estimation—a design choice prioritizing theoretical validation over computational optimization.

Importantly, this current implementation prioritizes accuracy over speed to establish the theoretical and empirical validity of physics-guided spatiotemporal modeling. Empirically, we observe that the inference speed can be greatly enhanced with minimal performance degradation by scaling the resolution of pre-trained diffusion models, *e.g.*, reducing resolution to 128×128 and 64×64 cuts inference time by 67.73% and 91.73%, respectively, while retaining over 98.63% and 94.57% of the original AUROC (Table 7). This trade-off underscores the flexibility of our approach in balancing accuracy and efficiency according to application needs. Future work may further boost efficiency via diffusion model compression [91, 92, 93] or efficient architecture design [94, 95], highlighting NSG-VD's potential for practical deployment as video generation and detection requirements advance.

Table 7: Comparisons with baselines in terms of Inference time and performance, where we train all models with $10,000$ real and generated videos from Kinetics-400 and SEINE, respectively.

Method	Recall(%) ↑	Accuracy(%) ↑	F1(%) ↑	AUROC(%) ↑	Infer. Time (s) ↓
DeMamba	71.25	84.54	80.87	94.42	0.0311
NPR	59.31	77.95	71.33	89.92	**0.0036**
TALL	61.47	80.20	74.05	95.14	0.0044
STIL	42.35	71.08	55.81	94.94	0.0108
NSG-VD (64x64)	78.29	83.26	81.99	90.27	0.0298
NSG-VD (128x128)	84.93	**86.99**	86.25	94.15	0.1163
NSG-VD (256x256)	**97.13**	86.05	**87.45**	**95.46**	0.3605

E.2 Numerical Stability of NSG

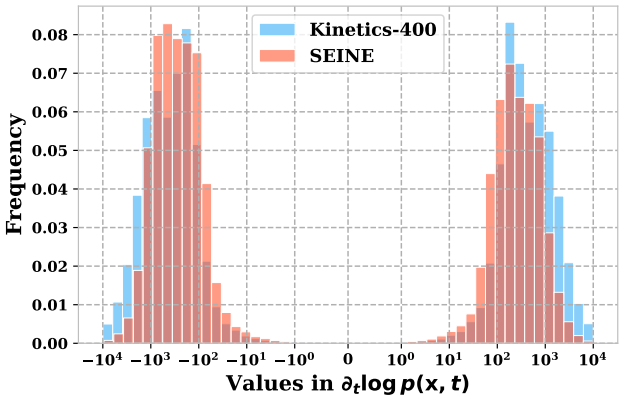


Figure 5: Distribution of the values of temporal derivatives $\partial_t \log p(\mathbf{x}, t)$ in the NSG statistic across $10,000$ real and generated videos from Kinetics-400 and SEINE, respectively.

To ensure the numerical stability of the NSG statistic with the temporal derivatives $\partial_t \log p(\mathbf{x}, t)$ in its denominator, we examine the distribution of values in $\partial_t \log p(\mathbf{x}, t)$ across $10,000$ real and generated videos from Kinetics-400 and SEINE, respectively. From Figure 5, nearly all values lie outside the critical near-zero range $[-0.1, 0.1]$. This indicates that almost no value in $\partial_t \log p(\mathbf{x}, t)$ approaches zero in practice, effectively mitigating instability risks from division by vanishingly small values.

The observed distribution aligns with the physical intuition that temporal density changes in real or synthetic videos are inherently non-stationary, resulting in measurable temporal derivatives. Additionally, the regularization term $\lambda > 0$ in the NSG denominator (Eqn. 6) further safeguards against edge cases where $\partial_t \log p(\mathbf{x}, t)$ might marginally approach zero. These design choices collectively ensure robust numerical stability for NSG across diverse video distributions.

E.3 Impact of MMD for NSG-VD

To validate the inherent superiority of the NSG statistic independent of the training objective, we compare our framework trained with both Maximum Mean Discrepancy (NSG-VD) and standard binary cross-entropy loss (NSG-BCE) against baselines using BCE. From Table 8, NSG-BCE achieves superior performance across all metrics compared to state-of-the-art baselines, even when adopting a conventional training paradigm. For example, it achieves 77.67% average Recall and 82.70% F1-score, significantly outperforming Demamba by $6.42\% \uparrow$ in Recall and $1.83\% \uparrow$ in F1-score, and TALL by $16.20\% \uparrow$ in Recall and $8.65\% \uparrow$ in F1-score. This demonstrates that the NSG statistic's ability to capture spatiotemporal dynamics remains effective regardless of the training objective.

Notably, NSG-BCE excels in challenging scenarios where other methods struggle. For instance, it achieves 64.29% Recall on Sora (vs. 42.86% for Demamba) and 63.60% Recall on WildScrape (vs. 48.00% for Demamba), highlighting its ability to generalize beyond superficial artifacts. The performance gap widens further in NSG-VD (97.13% Recall), where MMD explicitly models distributional shifts by the NSG feature in a producing kernel Hilbert space and enables more precise separation between real and synthetic videos. These results confirm that the NSG statistic's physics-driven design captures fundamental spatiotemporal dynamics, providing an intrinsic advantage over conventional features regardless of the training strategy.

Table 8: Impact of MMD in our NSG-VD across 10 generated paradigms (%), where we train all models with $10,000$ real and generated videos from Kinetics-400 and SEINE, respectively.

Method	Metric	Model Scope	Morph Studio	Moon Valley	HotShot	Show1	Gen2	Crafter	Lavie	Sora	Wild Scrape	Avg.
DeMamba	Recall	47.40	87.80	88.20	77.40	75.00	85.60	91.60	68.60	42.86	48.00	71.25
	Accuracy	72.80	93.00	93.20	87.80	86.60	91.90	94.90	83.40	68.75	73.10	84.54
	F1	63.54	92.62	92.84	86.38	84.84	91.36	94.73	80.52	57.83	64.09	80.87
	AUROC	88.29	98.39	98.76	97.84	96.89	98.76	99.35	96.87	80.93	88.11	94.42
NPR	Recall	46.40	76.40	69.80	63.80	56.00	75.00	83.80	58.80	35.71	27.40	59.31
	Accuracy	71.40	86.40	83.10	80.10	76.20	85.70	90.10	77.60	66.96	61.90	77.95
	F1	61.87	84.89	80.51	76.22	70.18	83.99	89.43	72.41	51.95	41.83	71.33
	AUROC	85.73	96.01	93.79	91.44	89.96	95.13	96.87	89.46	84.15	76.66	89.92
TALL	Recall	58.60	75.00	79.40	60.20	62.00	77.80	88.20	43.80	33.93	35.80	61.47
	Accuracy	78.80	87.00	89.20	79.60	80.50	88.40	93.60	71.40	66.07	67.40	80.20
	F1	73.43	85.23	88.03	74.69	76.07	87.02	93.23	60.50	50.00	52.34	74.05
	AUROC	97.10	98.12	98.63	96.37	96.45	97.76	99.38	94.80	83.35	89.45	95.14
STIL	Recall	28.60	57.40	78.40	46.80	18.80	66.40	69.00	24.80	14.29	19.00	42.35
	Accuracy	64.20	78.60	89.10	73.30	59.30	83.10	84.40	62.30	57.14	59.40	71.08
	F1	44.41	72.84	87.79	63.67	31.60	79.71	81.56	39.68	25.00	31.88	55.81
	AUROC	95.53	97.91	99.40	96.49	92.79	98.06	98.86	91.00	92.79	86.58	94.94
NSG-BCE (Ours)	Recall	53.40	96.40	94.80	90.60	77.60	79.40	83.20	73.40	64.29	63.60	77.67
	Accuracy	72.70	94.20	93.40	91.30	84.80	85.70	87.60	82.70	74.11	77.80	84.43
	F1	66.17	94.32	93.49	91.24	83.62	84.74	87.03	80.93	71.29	74.13	82.70
	AUROC	84.67	98.79	97.77	96.90	92.69	93.00	93.86	91.32	83.58	87.99	92.06
NSG-VD (Ours)	Recall	91.67	100.00	100.00	100.00	100.00	98.33	100.00	97.50	94.64	89.17	**97.13**
	Accuracy	82.50	88.33	89.58	84.58	86.25	87.08	86.67	87.92	89.29	78.33	**86.05**
	F1	83.97	89.55	90.57	86.64	87.91	88.39	88.24	88.97	89.83	80.45	**87.45**
	AUROC	90.67	97.62	98.38	95.88	96.69	97.87	97.64	95.09	96.14	88.65	**95.46**

E.4 Impact of Size of Reference Set for NSG-VD

We investigate the impact of reference set size by evaluating subsets containing between 10 and 500 samples, with other settings remaining consistent with Table 1. Performance is assessed using comprehensive criteria, including AUROC, Accuracy, F1 Score, and Recall. Intuitively, a larger reference set enables more accurate estimation of the underlying distribution of real videos, thereby supporting more stable and reliable detection. In contrast, smaller reference sets may introduce substantial sampling and estimation biases. As shown in Figure 6, our NSG-VD demonstrates consistently robust performance across varying reference set sizes, with the exception of extreme cases involving very limited samples (*e.g.*, $n = 10$). Consequently, we set $n = 100$ for all experiments.

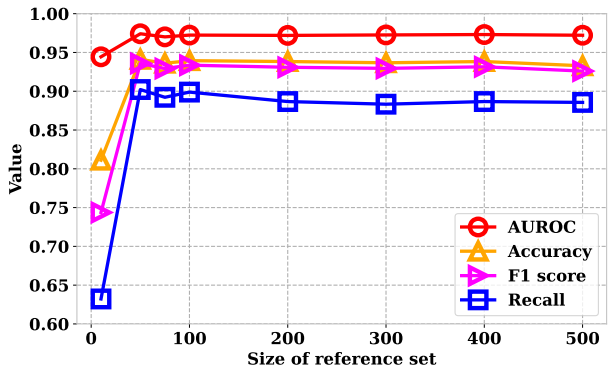


Figure 6: Impact of reference set size for NSG-VD, where we train all models with $10,000$ real and generated videos from Kinetics-400 and Pika, respectively.

E.5 Impact of Diversity of Real Videos in the Reference Set

We conduct additional ablation studies on *real-domain mixed* reference sets, revealing a key strength of NSG-VD: strong generalization to unseen generated video domains when most real test samples are covered by the reference distribution. Specifically, we train on Kinetics-400 (real) and SEINE (generated) videos, and test on MSR-VTT (real) and 10 generated videos using reference sets with varying ratios of MSR-VTT and Kinetics-400. From Table 9, even a small proportion ($3:7$) yields satisfactory performance (84.19% of Accuray, 81.12% of F1-Score) compared with baselines, which quickly saturates. This confirms that NSG-VD needs only modest in-domain real coverage, while the fake side can remain highly heterogeneous. These results will be included in our revision.

Table 9: Performance under different domain coverage ratios between MSR-VTT and Kinetics-400.

Domain Coverage (MSR-VTT : Kinetics-400)	0:10 (Low)	3:7 (Medium)	5:5 (Balanced)	7:3 (High)	10:0 (Full)	DeMamba	TALL
Average Accuracy (%)	77.82	84.19	85.57	**87.06**	86.05	84.21	80.20
Average F1-Score (%)	75.68	81.12	83.36	85.41	**87.45**	80.87	74.05

E.6 Discussions on Assumption of the Divergence Term

We assume $\nabla_{\mathbf{x}} \cdot \mathbf{v}$ is subdominant in smoothly varying video distributions for three reasons: **First**, its direct estimation is ill-posed in high-dimensional video data. Solving $\partial_t \mathbf{x} = \mathbf{v}(\mathbf{x}, t)$ is an underdetermined inverse problem, and video noise (*e.g.*, blur or compression) further amplifies estimation errors, making explicit divergence computation unstable and infeasible [32, 29]. **Second**, many physical flows approximate incompressibility ($\nabla_{\mathbf{x}} \cdot \mathbf{v} \approx 0$), a simplification grounded in fluid dynamics [29] and quantum mechanics [55] that preserves physical interpretability. **Third**, our NSG remains robust even if $\nabla_{\mathbf{x}} \cdot \mathbf{v} \neq 0$, as it captures cumulative spatiotemporal inconsistencies across all terms in Eqn. (3). Experiments confirm the resilience of NSG-VD to deviations from this assumption.

F Limitations and Future Directions

While our proposed NSG-VD method demonstrates strong performance across diverse AI-generated video detection scenarios, several limitations and opportunities for future work remain:

Limitations. First, the current formulation of the NSG statistic relies on simplified physical assumptions (*e.g.*, the incompressible flow approximation in continuity equations), which may fail to capture highly dynamic or discontinuous motion patterns in complex real-world scenarios. Second, the effectiveness of NSG-VD critically depends on the quality of pre-trained diffusion models used for score estimation; domain shifts or limited training data may degrade the reliability of estimated NSG features. Third, while NSG-VD achieves competitive performance, its reliance on diffusion models introduces computational overhead, making it less suitable for large-scale real-time detection tasks. Lastly, while our deep kernel design improves detection performance, its architecture could be further optimized to better adapt to heterogeneous spatiotemporal patterns.

Future Directions. To address these limitations, future work could explore more sophisticated physical models that account for compressible flows or discontinuous motion dynamics [54], enhancing the NSG statistic's adaptability to complex scenarios. Additionally, developing effective domain-specific fine-tuning strategies [96, 97, 98, 99] could improve the reliability of score estimation under distribution shifts. For real-time deployment, investigating lightweight diffusion model compression techniques (*e.g.*, pruning [91, 100], quantization [92, 93]) would reduce computational costs. Finally, advancing the design of the deep kernel network—such as incorporating attention mechanisms [101] or hierarchical feature fusion [89]—could further optimize the MMD-based detection framework, enabling better performance for AI-generated video detection.

G Broader Impacts

The development of AI-generated video detection methods like NSG-VD has significant societal, ethical, and technical implications. Our work contributes to mitigating the risks of malicious deepfake content, such as misinformation, identity theft, and political manipulation, by enabling more reliable verification of video authenticity. By leveraging physics-informed principles, NSG-VD provides a reliable framework for detecting synthetic videos that may otherwise evade traditional artifact-based detection methods. This could strengthen trust in digital media, support content moderation efforts, and aid legal or journalistic investigations involving video evidence.

This research aligns with broader efforts to establish trustworthy multimedia ecosystems. By bridging physics principles with machine learning, NSG-VD advances interpretable detection mechanisms—a critical step toward auditing AI-generated content while fostering public awareness of synthetic media risks. We encourage interdisciplinary collaboration among researchers, ethicists, and legislators to ensure such technologies serve as safeguards rather than instruments of control.

H Visualizations

Figure 7: Results of the detection on *real* videos from the MSR-VTT dataset.

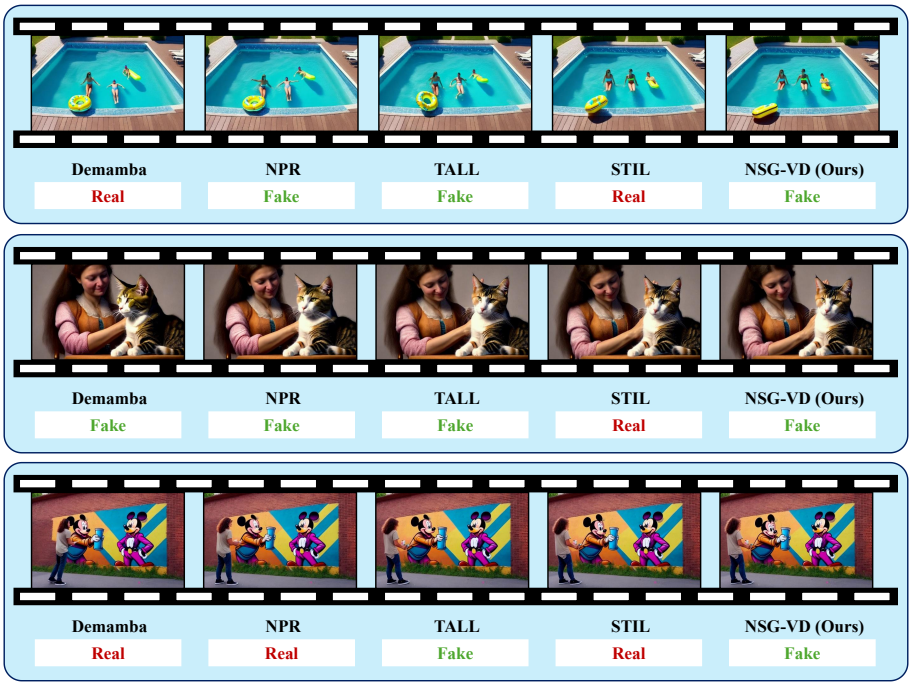


Figure 8: Results of the detection on *generated* videos from the Crafter dataset.

Figure 9: Results of the detection on *generated* videos from the Gen2 dataset.

Figure 10: Results of the detection on *generated* videos from the HotShot dataset.

Figure 11: Results of the detection on *generated* videos from the Lavie dataset.

Figure 12: Results of the detection on *generated* videos from the ModelScope dataset.

Figure 13: Results of the detection on *generated* videos from the MoonValley dataset.

Figure 14: Results of the detection on *generated* videos from the MorphStudio dataset.

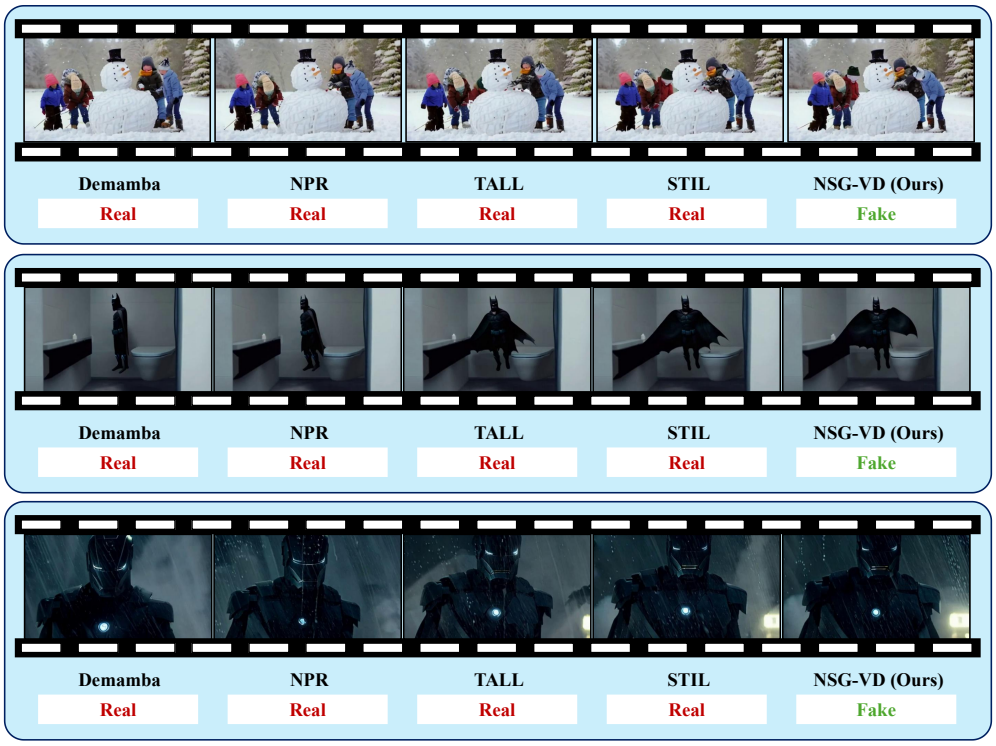


Figure 15: Results of the detection on *generated* videos from the Show1 dataset.

Figure 16: Results of the detection on *generated* videos from the Sora dataset.

Figure 17: Results of the detection on *generated* videos from the Seaweed dataset.

Figure 18: Results of the detection on *generated* videos from the Seaweed dataset.

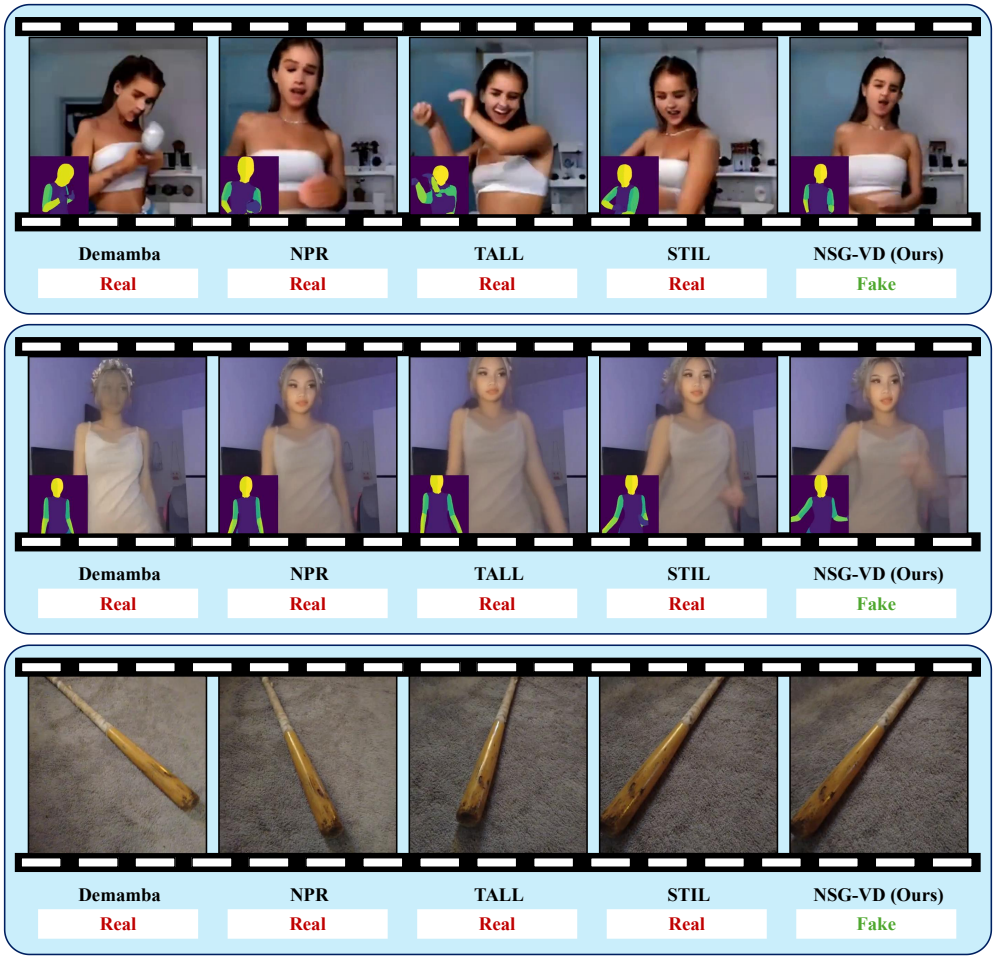


Figure 19: Results of the detection on *generated* videos from the WildScrape dataset.

To further demonstrate the excellent performance of our NSG-VD, we present visual detection results on both real and generated videos across all 10 datasets. As illustrated in Figures 7-19, both the baselines and NSG-VD demonstrate satisfactory detection on real video samples. For generated videos, the existing baselines achieve reasonable performance on early generation models (e.g., Crafter, Gen2, and MoonValley), but exhibit significant performance degradation when applied to more advanced generative models (e.g., Show1, Sora, and WildScrape). In contrast, NSG-VD consistently achieves strong detection performance across all generation levels.

On this basis, we consider the recently proposed Seaweed [102] method (as shown in Figures 17, 18), which generates highly realistic long-form videos. All four baselines exhibit near-complete failure on this dataset, whereas NSG-VD continues to deliver effective detection performance.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's contributions, including the proposed Normalized Spatiotemporal Gradient (NSG) statistic, the NSG-VD detection method based on probability flow conservation, and the theoretical analysis.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper provides complete theoretical results for the NSG feature lower bound (Theorem 1 in Section 3.4), including assumptions (*e.g.*, Gaussian-distributed real/fake videos and temporal derivatives constraints). The proof is detailed in Appendix A.7, including all mathematical derivations and references to supporting lemmas. All theorems and lemmas are numbered and cross-referenced, and proofs are accessible in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We fully disclose all information needed to reproduce the main experimental results of the paper, see Section 4 and Appendix C.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We will release our code upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide full experimental detail content in our experimental settings (see Section 4 and Appendix C).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We evaluate statistical significance by reporting mean and standard deviation across multiple runs with different random seeds (Appendix D.1).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide detailed information about on computing resources in Appendix C.3.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The paper meets the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in Appendix G.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We strictly follow the license of the assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: This paper does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.