Physics-Driven Spatiotemporal Modeling for AI-Generated Video Detection

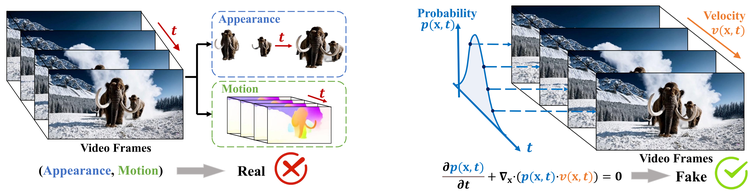
Shuhai Zhang*, Zihao Lian*, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li†, Mingkui Tan†

shuhaizhangshz@gmail.com mingkuitan@scut.edu.cn

Background

- AI-generated videos have achieved near-perfect visual realism, urgently necessitating reliable detector. However, detecting such videos faces challenges in **modeling high-dimensional spatiotemporal dynamics** and identifying subtle anomalies that **violate physical laws**.
- **Challenges**: Modeling spatiotemporal dynamics, diverse and evolving generative paradigms.

Modeling Spatiotemporal Dynamics via NSG



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

- **Probability Flow Velocity Field** $\mathbf{v}(\mathbf{x},t)$: Define **probability flow** $\mathbf{J}(\mathbf{x},t)=p(\mathbf{x},t)\cdot\mathbf{v}(\mathbf{x},t)$ as the movement of probability mass of $\mathbf{x}$ over time t, the conservation of probability mass implies:

$$\frac{\partial p(\mathbf{x},t)}{\partial t}+\nabla_{\mathbf{x}}\cdot[p(\mathbf{x},t)\cdot\mathbf{v}(\mathbf{x},t)]=0, \tag{1}$$

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 $\mathbf{J}$. Applying the chain rule:

$$\partial_t\log p(\mathbf{x},t)+\nabla_{\mathbf{x}}\cdot\mathbf{v}(\mathbf{x},t)+\mathbf{v}(\mathbf{x},t)\cdot\nabla_{\mathbf{x}}\log p(\mathbf{x},t)=0. \tag{2}$$

Under incompressible flow approximations used in fluid dynamic, *i.e.*, $\nabla_{\mathbf{x}}\cdot\mathbf{v}\approx 0$, we have

$$\mathbf{v}(\mathbf{x},t)\cdot\frac{\nabla_{\mathbf{x}}\log p(\mathbf{x},t)}{-\partial_t\log p(\mathbf{x},t)}\approx 1. \tag{3}$$

- **Definition (Normalized Spatiotemporal Gradient (NSG).)** The relation in Eqn. (3) reveals that *natural video dynamics preserve the product between **the velocity field** and **the ratio of spatial probability gradients to temporal density changes**.* Formalize this constrained ratio as NSG:

$$\mathbf{g}(\mathbf{x},t)=\frac{\nabla_{\mathbf{x}}\log p(\mathbf{x},t)}{-\partial_t\log p(\mathbf{x},t)+\lambda}. \tag{4}$$

Here, $\lambda>0$ prevents numerical instability, $\mathbf{g}(\mathbf{x},t)$ acts as a *dual field* to $\mathbf{v}(\mathbf{x},t)$, satisfying $\mathbf{v}\cdot\mathbf{g}\approx 1$.

Estimating NSG with Diffusion Models

- **Spatial Gradients Estimation.** For a video $\mathbf{x}$ at t-th frame, the spatial gradient is estimated by:

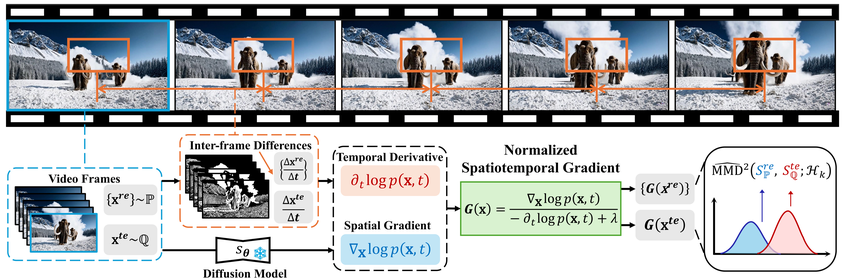
$$\nabla_{\mathbf{x}}\log p(\mathbf{x},t)\approx\mathbf{s}_\theta(\mathbf{x}_t), \tag{5}$$

where $\mathbf{s}_\theta$ denotes the learned *score function* from diffusion models.

- **Temporal Derivatives Approximation.** Under the brightness constancy assumption, which posits that the probability density along motion trajectories remains constant, we can approximatie the temporal derivatives as

$$\partial_t\log p(\mathbf{x},t)\approx-\frac{\nabla_{\mathbf{x}}\log p(\mathbf{x},t)\cdot\Delta\mathbf{x}}{\Delta t}. \tag{6}$$

Exploring NSG for Detecting AI-Generated Videos



- **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 $S^{re}_{\mathbb{P}}=\{\mathbf{x}^{(i)}\}_{i=1}^n$ denote a reference set of real videos and $S^{te}_{\mathbb{Q}}=\{\tilde{\mathbf{y}}\}$ represent a test video, compute their MMD as:

$$\widehat{\mathrm{MMD}}_b^2[S^{re}_{\mathbb{P}},S^{te}_{\mathbb{Q}};\mathcal{H}_k]=\frac{1}{n^2}\sum_{i,j=1}^n k\Big(\mathbf{G}(\mathbf{x}^{(i)}),\mathbf{G}(\mathbf{x}^{(j)})\Big)-\frac{2}{n}\sum_{i=1}^n k\Big(\mathbf{G}(\mathbf{x}^{(i)}),\mathbf{G}(\tilde{\mathbf{y}})\Big)+k(\mathbf{G}(\tilde{\mathbf{y}}),\mathbf{G}(\tilde{\mathbf{y}})), \tag{7}$$

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

$$f(\tilde{\mathbf{y}})=\begin{cases}\text{Fake}, & \text{if } f(\tilde{\mathbf{y}};S_{\mathbb{P}},k_\omega,\tau)=1,\\ \text{Real}, & \text{if } f(\tilde{\mathbf{y}};S_{\mathbb{P}},k_\omega,\tau)=0.\end{cases} \tag{8}$$

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

$$k_\omega(\mathbf{x},\mathbf{y})=[(1-\epsilon)\kappa\left(\phi_{\mathbf{G}}(\mathbf{x}),\phi_{\mathbf{G}}(\mathbf{y})\right)+\epsilon]\cdot\Phi\left(\mathbf{G}(\mathbf{x}),\mathbf{G}(\mathbf{y})\right), \tag{9}$$

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:

$$k^*_\omega=\arg\max_{k_\omega}\frac{\widehat{\mathrm{MPP}}_u(S^{tr}_{\mathbb{P}},S^{tr}_{\mathbb{Q}};k_\omega)}{\sqrt{\hat{\sigma}^2(S^{tr}_{\mathbb{P}},S^{tr}_{\mathbb{Q}};k_\omega)+\lambda}},\ \hat{\sigma}^2=\frac{4}{N^3}\sum_{i=1}^N\left(\sum_{j=1}^N H^*_{ij}\right)^2-\frac{4}{N^4}\left(\sum_{i=1}^N\sum_{j=1}^N H^*_{ij}\right)^2, \tag{10}$$

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

Contributions

- We formulate the video evolution as a probability flow with a continuity equation and propose a statistic **Normalized Spatiotemporal Gradient (NSG)** that explicitly models spatiotemporal dynamics of videos. NSG fundamentally captures violations of physical continuity without artifact supervision.
- We develop an NSG estimator by **spatial gradients approximation** and **motion-aware temporal dynamics modeling** using pre-trained diffusion models with physical priors.
- We propose **an NSG-based video detection method (NSG-VD)**, which quantifies distributional shifts in NSG features using ***Maximum Mean Discrepancy* (MMD)**. Theoretical analyses and extensive experiments validate the superiority of our NSG-VD in detecting advanced generated videos.

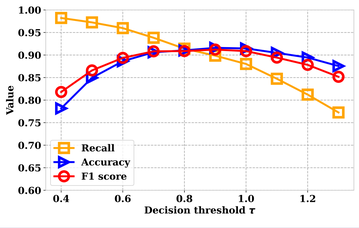
Experiments

- **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**

- **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
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**

- **Ablations on $\nabla_{\mathbf{x}}\log p(\mathbf{x},t)$, $\partial_t\log p(\mathbf{x},t)$ and τ.**

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**