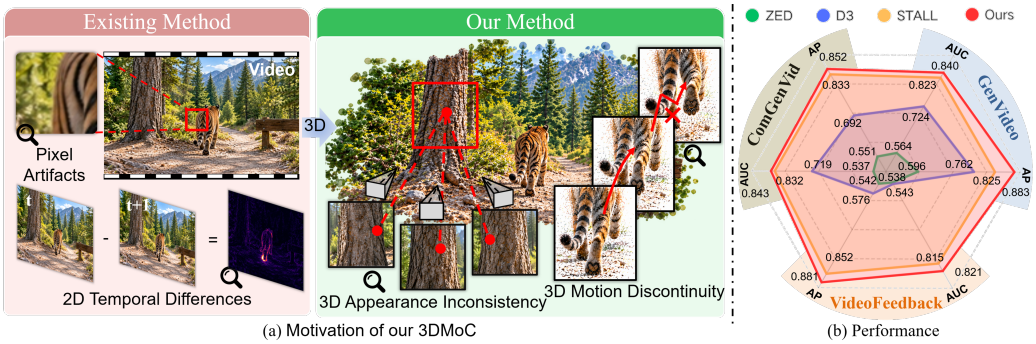
3DMoC: Training-Free AI-Generated Video Detection via 3D Motion Consistency

Anonymous authors
Paper under double-blind review



(a) Motivation of our 3DMoC (b) Performance

Figure 1: (a) Existing methods mainly exploit local pixel artifacts or 2D temporal differences, which are increasingly weakened by advanced video generators. Our method reveals both tracked appearance inconsistencies and 3D motion discontinuities in the scene elements, (b) thereby maintaining excellent AUC and AP scores across multiple benchmark tests.

Abstract

Existing methods for detecting AI-generated videos can identify common 2D spatiotemporal anomalies. However, there remain two limitations for detecting high-quality videos: (1) These methods primarily detect local pixel artifacts but lack spatial modeling, making it difficult to identify appearance differences from different views. (2) They primarily measure temporal inconsistencies based on inter-frame differences, thus struggling to identify motion anomalies in space. To address the above issues, we propose a training-free generated video detection via 3D Motion Consistency (**3DMoC**), which captures 3D appearance and motion anomalies in video, thereby achieving more accurate identification of AI-generated videos. Specifically, to characterize the appearance consistency of elements, we devise a **Tracked Appearance Consistency** module that aligns the same region across views through trajectory tracking to capture appearance differences in scene. Furthermore, we design a **3D Motion Continuity** module that measures the degree of motion variation between tracked 3D world points to identify motion anomalies. Benefiting from the collaborative designs, our 3DMoC effectively captures both anomalies in 3D appearance and motion within generated videos. Extensive experiments demonstrate that our method achieves state-of-the-art detection accuracy on different generated-video benchmarks.

1 Introduction

Recent advances in text-to-video and image-to-video generation have substantially improved the realism and temporal coherence of generated videos, enabling the creation of high-fidelity dynamic content from simple textual or visual prompts Chen et al. (2024a); Wan et al. (2025). However, such powerful generation capabilities also introduce serious security risks: highly realistic synthetic videos can be maliciously used for misinformation, fraud, and impersonation Chen et al. (2024b); Zheng et al. (2025); Hayun et al. (2026). Therefore, it is imperative to develop reliable AI-generated video detection methods that can effectively distinguish fake videos from real ones and remain robust against continuously emerging generation models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Early studies on AI-generated video detection have predominantly adopted supervised learning paradigms, where detectors are trained on labeled real and generated videos to learn discriminative spatial or spatiotemporal artifacts Zhang et al. (2025). For example, some methods inherit image forgery detection and focus on spatial appearance artifacts Zheng et al. (2021), while subsequent studies increasingly incorporate temporal modeling to capture inter-frame inconsistencies Gu et al. (2021). More recent methods further leverage large-scale pretrained visual or video encoders to jointly model appearance and temporal anomalies, achieving strong detection performance on known generation models and benchmark datasets Ni et al. (2022); Chen et al. (2024b); Ma et al. (2026). Despite these advances, these methods exhibit significant performance degradation when encountering unseen generators, as they rely heavily on the features specific to training data.

To alleviate the dependence on generated training data, recent studies have increasingly explored training-free detection, which relies on frozen pretrained representations and intrinsic statistics of real videos for detecting AI-generated video Feng et al. (2026). One research approach focuses on exploring the local spatiotemporal consistency of video representations Zheng et al. (2025); Internò et al. (2025); Hyun & Kim (2026), characterizing the inter-frame differences in generated videos. Another research approach models the statistical distribution of real videos in the representation space Hayun et al. (2026), and improves detection accuracy based on the likelihood deviation between generated content and real data. However, the aforementioned methods still face two key limitations when facing high-quality generated videos:(1) Existing methods primarily focus on local visual artifacts, making it difficult to detect appearance changes of elements in space. (2) These methods primarily characterize temporal anomalies through inter-frame differences, thus struggling to identify abrupt changes of spatial motion.

To tackle the above challenges, we propose a training-free AI-generated video detection framework based on **3D Motion Consistency** (**3DMoC**), which exploits 3D appearance and motion anomalies in scenes to identify AI-generated videos, as shown in Fig. 1. Specifically, to characterize the appearance consistency of scene, we design a Tracked Appearance Consistency module that aligns features in the same region across different views based on trajectories for detecting 3D appearance changes. Furthermore, we devise a 3D Motion Continuity module that models the motion variations of 3D points in the same region to identify motion discontinuities in scene. Thanks to these two dedicated designs, our method captures 3D appearance and motion anomalies to achieve more accurate AI-generated video detection, without relying on generator-specific artifacts for training. Extensive experiments demonstrate that our method achieves competitive detection performance across diverse generated-video benchmarks and exhibits strong generalization.

The main contributions of this paper are as follows:

- We propose 3DMoC, a training-free framework that detects AI-generated videos by identifying 3D appearance and motion anomalies. To the best of our knowledge, this is the first method to explore 3D motion consistency for training-free AI-generated video detection.
- A Tracked Appearance Consistency module is designed to identify cross-view appearance inconsistencies based on features aligned from 3D trajectories.
- We devise a 3D motion continuity module that detects abnormal motion by analyzing the variations of tracked points in 3D scene.
- Extensive experiments on multiple AI-generated video benchmarks demonstrate the effectiveness and cross-generator generalization of our method.

2 Related Work

Supervised AI-Generated Video Detection. Early studies on AI-generated video detection follow the paradigm of image forgery detection, focusing on spatial artifacts such as abnormal textures, frequency patterns, and generation traces within individual frames Tan et al. (2024); Zheng et al. (2024). With the rapid development of video generation models, subsequent works increasingly incorporate temporal information to capture inconsistencies between frames. For instance, STIL Gu et al. (2021) jointly models spatial and temporal inconsistencies for video forgery detection, while FTCN Zheng et al. (2021) and MINTIME Coccomini et al. (2024) further exploit temporal coherence across consecutive frames. More recent methods leverage large-scale pretrained vision or video

representations to enhance spatiotemporal modeling. DeMamba Chen et al. (2024b) introduces sequential modeling to capture local spatiotemporal inconsistencies. In addition, physics-driven approaches further attempt to exploit dynamic priors beyond conventional visual artifacts Zhang et al. (2025). However, they generally require labeled videos for training and tend to capture generator- or dataset-specific characteristics, resulting in degraded generalization when encountering unseen generators.

Training-Free AI-Generated Video Detection. To alleviate the dependence on generated training samples, recent studies have increasingly explored training-free AI-generated video detection by leveraging frozen pretrained representations and intrinsic characteristics of real videos Zheng et al. (2025); Internò et al. (2025); Hayun et al. (2026); Hyun & Kim (2026). For example, D3 Zheng et al. (2025) exploits temporal dynamics encoded in frozen feature representations, providing effective cues for detecting videos from unseen generators. Following this direction, ReStraV Internò et al. (2025) further investigates intrinsic temporal regularities of real and generated videos. More recently, SPLIT Hyun & Kim (2026) combines global video representations with local patch-level spatiotemporal consistency, while STALL Hayun et al. (2026) further extends training-free detection from temporal dynamics to joint spatial-temporal statistics to capture complementary forgery cues across different spatial scales.

However, existing methods still rely primarily on pixel-level artifacts and two-dimensional temporal discrepancies for detection, whereas these clues are gradually being eliminated in increasingly realistic generated videos. In contrast, we propose a specialized detection framework from the perspective of 3D motion consistency, which detects 3D appearance differences and motion anomalies in tracked scene elements to achieve more accurate detection.

3 METHOD

3.1 OVERVIEW

Given an input video $V = \{I_t\}_{t=1}^{T}$, where I_t denotes the t-th sampled frame and T is the number of frames, we detect generated videos from two complementary perspectives: **Tracked Appearance Consistency(TAC)** and **3D Motion Continuity(3DMC)**, as shown in Fig. 2.

The first module evaluates whether the same scene content preserves consistent visual appearance under viewpoint changes. Frozen DINOv3 features are used to characterize global and local appearance statistics, while VGGT correspondences provide spatial associations across frames for comparing feature differences in the same scene elements. The second module analyzes reconstructed 3D point trajectories and measures whether their motion exhibits regular displacement and variation. The two complementary scores are finally normalized and fused into a video-level anomaly score.

3.2 TRACKED APPEARANCE CONSISTENCY

For each frame I_t, a frozen DINOv3 encoder extracts a global CLS feature $\mathbf{e}_t \in \mathbb{R}^D$, and N local patch features: $\mathbf{F}_t = \{\mathbf{f}_{t,j}\}_{j=1}^{N}, \mathbf{f}_{t,j} \in \mathbb{R}^D$, where D denotes the feature dimension. To determine whether elements in the same scene maintain a consistent appearance across different viewpoints, we simultaneously model global appearance statistics and tracked local appearance consistency.

Global Appearance Statistics. We evaluate the global representation based on a reference feature distribution constructed from an independent calibration set Hayun et al. (2026) to determine whether it conforms to the typical visual distribution of real videos. Let $\boldsymbol{\mu}_s \in \mathbb{R}^D$ denotes the reference center and $\mathbf{W}_s \in \mathbb{R}^{D\times D}$ the corresponding decorrelation transformation. The frame-level appearance difference is defined as:

$$\ell_t^s = \|(\mathbf{e}_t - \boldsymbol{\mu}_s)\mathbf{W}_s\|_2^2 + C, \tag{1}$$

where C is a constant independent of the input frame. This score measures how closely the global representation of I_t conforms to the reference distribution of real videos. We aggregate the frame-level statistics into a video-level spatial statistic:

$$L^s = \max_{1\le t\le T} \ell_t^s. \tag{2}$$

L^s captures the maximum deviation of the video to the reference appearance distribution.

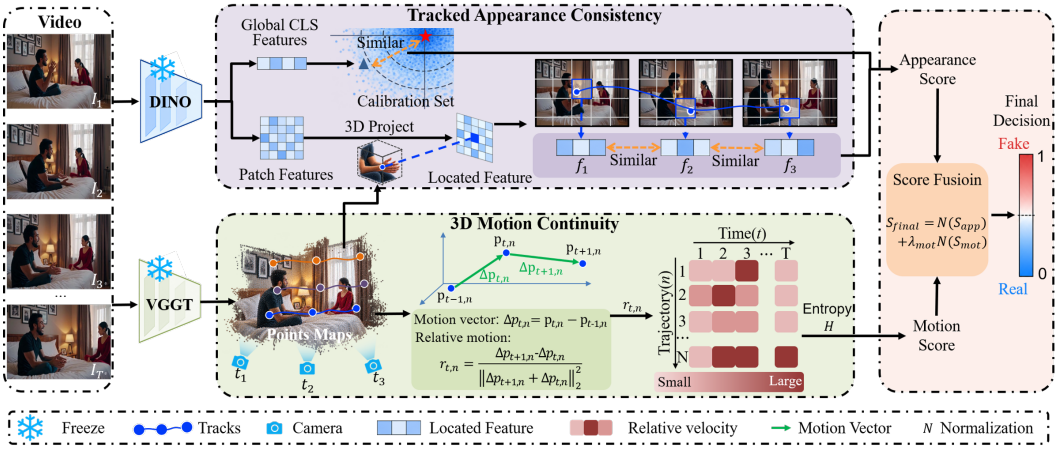


Figure 2: Overview of our 3DMoC. The Tracked Appearance Consistency branch projects 3D tracking points onto local patches and tracks corresponding regions across frames to measure appearance consistency. The 3D Motion Continuity branch reconstructs point trajectories and characterizes their temporal motion variations using relative motion statistics.

To obtain a comparable video-level score, we further normalize L^s with respect to the statistics collected from the independent calibration set:

$$p^s = \frac{1}{|\mathcal{C}_s|} \sum_{c \in \mathcal{C}_s} \mathbb{I}(c \leq L^s), \tag{3}$$

where $\mathcal{C}_s$ denotes the corresponding reference statistics, $\mathbb{I}(\cdot)$ is the indicator function. A larger p^s indicates a greater discrepancy from the distribution of real videos.

Local Appearance Consistency. Direct comparison of local features at fixed image locations is unreliable under object motion and viewpoint changes, since the same image position may correspond to different scene content across frames. We therefore leverage VGGT correspondences to associate the same scene content between adjacent frames.

For the n-th correspondence in frame I_t, let $\mathbf{x}_{t,n} = (u_{t,n}, v_{t,n}) \in \mathbb{R}^2$ denotes its 2D image-space location. By mapping $\mathbf{x}_{t,n}$ onto the DINO patch grid, we obtain the corresponding local appearance representation $\mathbf{g}_{t,n} \in \mathbb{R}^D$. For each valid correspondence between frames t and $t+1$, we characterize its appearance variation as:

$$\Delta \mathbf{g}_{t,n} = \mathbf{g}_{t+1,n} - \mathbf{g}_{t,n}, \tag{4}$$

and normalize the variation direction by:

$$\widehat{\Delta \mathbf{g}}_{t,n} = \frac{\Delta \mathbf{g}_{t,n}}{\|\Delta \mathbf{g}_{t,n}\|_2 + \epsilon}, \tag{5}$$

where ϵ is a very small value that prevents the denominator from 0 in static points. We then evaluate whether these trajectory-aligned appearance variations conform to the temporal statistics of real videos. Let $\boldsymbol{\mu}_t \in \mathbb{R}^D$ and $\mathbf{W}_t \in \mathbb{R}^{D \times D}$ denote the temporal reference center and decorrelation transformation obtained from the independent calibration set Hayun et al. (2026). The local appearance difference is defined as:

$$\ell_{t,n}^{\text{local}} = \left\| \left(\widehat{\Delta \mathbf{g}}_{t,n} - \boldsymbol{\mu}_t \right) \mathbf{W}_t \right\|_2^2 + C, \tag{6}$$

where C is a constant independent of the input video.

Similar to the global appearance branch, we first aggregate the frame-pair appearance-consistency responses into a video-level temporal statistic following Eq. 2. We then normalize the resulting statistic with respect to the calibration reference distribution following Eq. 3, yielding the temporal appearance score p^t.

Finally, the spatial appearance statistic p^s and temporal appearance statistic p^t are combined as

$$S_{\text{app}} = \lambda_1 p^s + (1 - \lambda_1) p^t, \tag{7}$$

where λ_1 is the weighting parameter. A higher value of S_{app} indicates a greater deviation from the calibrated appearance statistics, and therefore a higher probability that the content is fake.

3.3 3D Motion Continuity

We further examine whether the tracked scene points exhibit continuity in the reconstructed 3D space. Specifically, VGGT provides the world-point map, point trajectories, visibility, and confidence for the same sampled frame sequence. We associate each 2D trajectory with its corresponding 3D world coordinates and analyze the short-term motion of the same tracked scene point across consecutive frames.

3D Motion Evolution. Let $\mathbf{p}_{t,n} \in \mathbb{R}^3$ denotes the reconstructed 3D position of the n-th tracked scene point in frame I_t. Its displacement vector between adjacent frames is defined as:

$$\Delta\mathbf{p}_{t,n} = \mathbf{p}_{t+1,n} - \mathbf{p}_{t,n}. \tag{8}$$

To describe the trend of local 3D motion over time, we further compute the difference vector between two consecutive motion vectors:

$$\Delta^2\mathbf{p}_{t,n} = \Delta\mathbf{p}_{t+1,n} - \Delta\mathbf{p}_{t,n}. \tag{9}$$

Directly using this variation may overemphasize trajectories with naturally large motion. We normalize the second-order 3D motion variation by the local motion magnitude. Specifically, the local motion scale is defined as:

$$m_{t,n} = \frac{\|\Delta\mathbf{p}_{t,n}\|_2 + \|\Delta\mathbf{p}_{t+1,n}\|_2}{2}. \tag{10}$$

We therefore normalize it by the local motion scale of the same trajectory and define the relative motion as:

$$r_{t,n} = \rho\left(\frac{\|\Delta^2\mathbf{p}_{t,n}\|_2}{\rho(m_{t,n}) + \epsilon}\right), \tag{11}$$

where ϵ is a very small value that prevents the denominator from 0 in static points. $\rho(x) = \log(1+x)$ represents the logarithmic function that compress the range of motion, thereby preventing large displacements from excessively suppressing abnormal motion changes.

Entropy-normalized Motion. A large local motion variation is not necessarily abnormal in every video. As shown in Fig. 3, real videos exhibit more concentrated motion that reflects the coherent evolution of the scene, whereas generated videos exhibit a more dispersed distribution caused by inconsistent local 3D motion. To account for entire video-specific context, we first organize all valid relative motion responses into a video-level motion matrix $\mathbf{R} = [r_{t,n}]$. Then, we further normalize the motion matrix to obtain the probability distribution of relative motion, as follows:

$$q_{t,n} = \frac{r_{t,n}}{\sum_{(i,j)} r_{i,j} + \epsilon}, \tag{12}$$

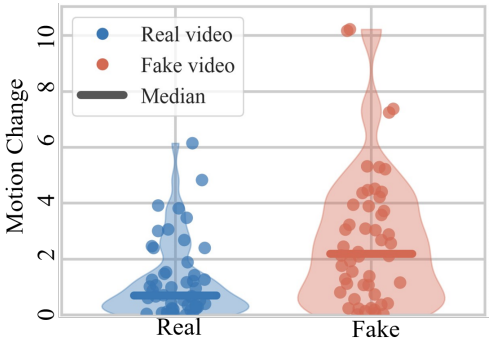


Figure 3: Comparison of motion changes between the real and fake videos.

where ϵ is also a minimum value, which prevents calculation errors caused by videos with absolutely no motion. The entropy of the relative-motion matrix can be calculated as follows:

$$H = -\sum_{(t,n)} q_{t,n} \log q_{t,n}. \tag{13}$$

Entropy H reflects the overall dispersion of motion in a video. A higher entropy value indicates greater motion complexity in the video itself.

We further aggregate local motion responses according to the video-specific motion dispersion:

$$S_{\text{mot}} = \sum_{t,n} \left(\frac{\max(r_{t,n}, 0)}{H + \epsilon}\right)^{\gamma}, \tag{14}$$

where γ controls how local abnormality responses contribute to the video-level risk. A larger S_{mot} indicates that the 3D trajectory is less smooth, local motion changes are more abnormal, and the video is more likely to be fake.

3.4 Score Fusion

The Tracked Appearance Consistency and 3D Motion Continuity modules capture complementary appearance and motion cues. Since their scores have different numerical scales, we first apply a robust normalization based on the median and median absolute deviation(MAD):

$$\mathcal{N}(x_i) = \tanh\left(\frac{x_i - \text{median}(x)}{\text{MAD}(x) + \varepsilon}\right), \tag{15}$$

where $x = \{x_i\}_{i=1}^{N}$ denotes a set of scores in batch. This mechanism is used to align the numerical scales of two heterogeneous exception branches. The two normalized scores are then fused as:

$$S_{\text{final}} = \mathcal{N}(S_{\text{app}}) + \lambda_{\text{mot}}\mathcal{N}(S_{\text{mot}}), \tag{16}$$

where λ_{mot} controls the contribution of the 3D motion branch.

4 Experiments

4.1 Experimental Setup

Datasets and Generative Models. We conduct experiments on three widely used video generation benchmarks with different data scales and real-fake distributions. GenVideo Chen et al. (2024b) contains 10,000 real and 8,276 fake samples, ComGenVid Hayun et al. (2026) includes 1,700 real and 3,400 fake samples, and VideoFeedback He et al. (2024a) comprises 80 real and 680 fake samples. These datasets cover a variety of sample ratio scenarios, as well as a wide range of visual content, motion patterns, and generation paradigms, providing a comprehensive assessment of the generalization capabilities. In particular, these synthetic videos were generated by various video generation models, including SoRA[1], AnimateDiff Guo et al. (2024), ModelScope Wang et al. (2023), Pika[2], and others.

Baselines. We compare our 3DMoC with representative state-of-the-art detection methods, including AEROBLADE Ricker et al. (2024), RIGID He et al. (2024b), ZED Cozzolino et al. (2025), D3 Zheng et al. (2025), and STALL Hayun et al. (2026). These methods exploit different detection cues, such as spatial generation artifacts, temporal inconsistencies, or spatial-temporal feature distributions.

Evaluation Metrics. We employ the Area Under the Receiver Operating Characteristic Curve (AUC) and Average Precision (AP) as the primary evaluation metrics. AUC measures the overall discriminative ability of a detector across different decision thresholds, while AP evaluates the precision-recall trade-off and is particularly informative for assessing detection performance under varying class distributions. Higher AUC and AP indicate better discrimination between real and generated videos. All methods are evaluated under the same dataset and evaluation protocol to ensure a fair comparison.

Implementation Details. All experiments are conducted on a single NVIDIA RTX 3090 GPU. We employ frozen DINOv3 Siméoni et al. (2025) and VGGT Wang et al. (2025) models to extract appearance representations and reconstruct cross-frame 3D point trajectories, respectively, without additional model fine-tuning. All hyperparameters and evaluation protocols are kept fixed across different datasets and generative models. More implementation details and algorithms are provided in the **Supplementary**.

4.2 Comparisons

We compare our 3DMoC with existing detection approaches, focusing on their generalization ability across different video generation models on three benchmark datasets, as shown in Table 1. It can be

[1] https://openai.com/index/sora

[2] https://pika.art

Table 1: Comparison with SOTA detection methods across different benchmarks. The best result is highlighted in red, and the second-best result is highlighted in orange.

Method	VideoFeedback		GenVideo		ComGenVid		Overall	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC
AEROBLADE	0.5832	0.5728	0.6291	0.6059	0.6234	0.6797	0.6120	0.6195
RIGID	0.6248	0.6197	0.6502	0.6719	0.5709	0.5625	0.6153	0.6181
ZED	0.5419	0.5378	0.5961	0.5639	0.5513	0.5374	0.5631	0.5465
D3 (L2)	0.5769	0.5337	0.7733	0.7385	0.6822	0.7192	0.6775	0.6638
D3 (cos)	0.5755	0.5427	0.7624	0.7241	0.6919	0.7201	0.6768	0.6620
STALL	0.8519	0.8152	0.8251	0.8230	0.8328	0.8315	0.8363	0.8232
3DMoC (Ours)	**0.8814**	**0.8205**	**0.8830**	**0.8403**	**0.8515**	**0.8428**	**0.8720**	**0.8345**

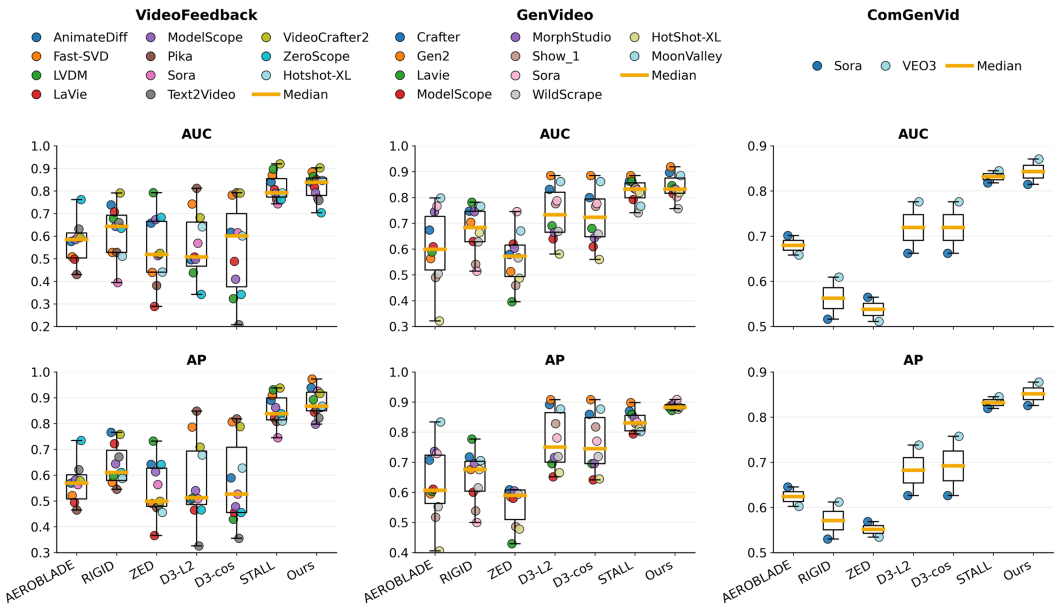


Figure 4: Performance comparison across different video generators on three datasets. Each colored dot represents a single video generator, and the orange horizontal line indicates the median value.

seen that our method achieves consistently strong performance in both AUC and AP. In particular, our 3DMoC obtains an overall AUC of 0.8345 and AP of 0.8720, outperforming the compared methods in terms of average detection performance. Existing methods such as AEROBLADE, RIGID, and ZED achieve relatively lower overall results, while D3(L2) and D3(cos) improve the detection accuracy but still exhibit clear performance gaps across different benchmarks. Although STALL performs competitively and achieves relatively stable results, our method further improves the overall detection performance, especially in terms of AP. Our method performs consistently across these three datasets, indicating that our 3DMoC demonstrates strong robustness despite significant differences in the proportions of positive and negative samples.

Moreover, in Fig. 4, existing methods exhibit noticeable performance variations across different generators. Specifically, AEROBLADE, RIGID, and ZED show relatively scattered AUC and AP distributions, indicating that their detection performance is sensitive to the underlying generation models. D3(L2) and D3(cos) achieve improved performance on several subsets, but still present considerable cross-generator fluctuations. Although STALL produces stable detection results and generally maintains higher scores than other baselines, its performance still varies across different benchmarks. In contrast, our 3DMoC consistently achieves high AUC and AP across all generative models with most results concentrated in the high-performance region.



Figure 5: Examples of appearance inconsistency and motion anomalies.

As further illustrated in Fig. 5, generated videos from different models commonly exhibit local appearance inconsistencies and abnormal motion patterns across frames, while real videos preserve more stable appearance and coherent motion. By exploiting these model-agnostic 3D consistency cues, 3DMoC is less dependent on generator-specific artifacts and therefore maintains more stable detection performance across diverse generated videos.

4.3 Robustness Evaluation

We further evaluate the robustness of our method under common video perturbations. As shown in Fig. 6, we apply Gaussian blur, Gaussian noise, JPEG compression, and random cropping with progressively increased severity levels. Our method maintains relatively stable performance under Gaussian blur and random cropping, with only minor fluctuations in both AP and AUC. Gaussian noise causes a moderate degradation as the perturbation becomes stronger, while JPEG compression introduces performance drop at high severity levels. Nevertheless, the detector preserves competitive detection performance across all settings, indicating its reliance on intrinsic appearance and 3D motion cues rather than pixel-level artifacts or specific sampling configurations.

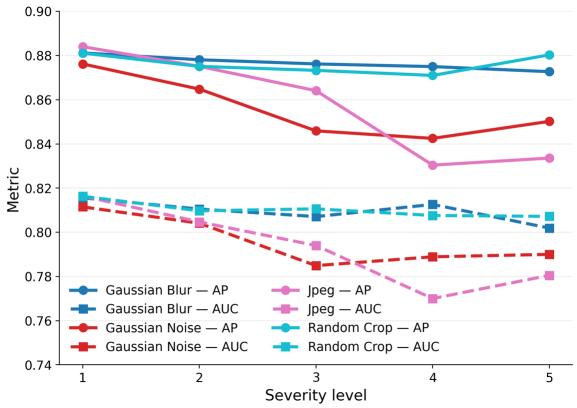


Figure 6: Robustness to common video perturbations. Detection performance under different severity levels of Gaussian blur, Gaussian noise, JPEG compression, and random cropping.

4.4 Ablation Studies

Ablation on Key Components. We perform an ablation study to evaluate the contribution of each component, as shown in Table 2. TAC alone already achieves stronger performance than 3DMC alone, indicating that appearance consistency provides more discriminative cues, while 3D motion information serves as an important complement. Removing Local Patch consistency causes performance drop, highlighting the importance of track-based alignment for capturing cross-frame appearance consistency. Similarly, removing Track results in a clear drop in both AP and AUC, highlighting the importance of trajectory-based alignment for capturing cross-frame appearance consistency. Furthermore, removing the entropy-based motion while retaining the other components results in a performance drop, indicating that the performance improvement stems not only from the pretrained VGGT representation but also from the proposed motion modeling. Overall, the full model achieves the strongest performance, validating the complementary contributions of TAC and 3DMC.

Table 2: Ablation study of key modules and components. TAC consists of Global CLS, Local Patch, and Track, while 3DMC consists of VGGT and Entropy. We report the AP and AUC under different component configurations to evaluate the contribution of each design.

Configuration	TAC			3DMC		AP	AUC
	Global CLS	Local Patch	Track	VGGT	Entropy		
w/o Global CLS		✓	✓	✓	✓	0.6221	0.5141
w/o Local Patch	✓		✓	✓	✓	0.7828	0.7218
w/o Track	✓	✓		✓	✓	0.8764	0.8189
w/o Entropy	✓	✓	✓	✓		0.8765	0.8129
TAC only	✓	✓	✓			0.8165	0.7929
3DMC only				✓	✓	0.7927	0.7894
Full model	✓	✓	✓	✓	✓	0.8814	0.8205

Table 3: Ablation study on the weight λ_{mot} and motion transformation ρ.

Configuration	AP	AUC
λ_{mot}		
0.05	0.8810	0.8212
0.10	0.8814	0.8205
0.20	0.8816	0.8190
0.50	0.8776	0.8061
ρ		
None	0.8768	0.8157
Log	0.8814	0.8205
Sqrt	0.8779	0.8154

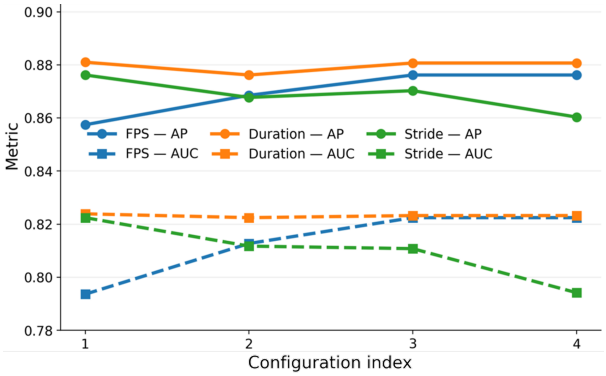


Figure 7: Sensitivity to video sampling configurations. AUC and AP under different configurations of FPS, video duration, and sampling stride.

Ablation on Hyperparameters. We further investigate the effect of the motion weighting factor λ_{mot} and motion transformation ρ, as reported in Table 3. A moderate increase in the motion-branch weight initially improves the detection accuracy, indicating that 3D structural cues provide complementary information to the appearance branch. However, assigning an excessively large weight to the motion branch degrades both AP and AUC. We therefore use $\lambda_{mot} = 0.1$ in the remaining experiments to achieve a balanced contribution from the appearance and 3D motion branches. We also compare several transformations for the motion statistic. Applying a logarithmic transformation provides the most consistent performance, as it suppresses extreme motion responses while preserving relative differences between normal and abnormal trajectories.

Temporal ablations. We further investigate the performance under different temporal sampling strategies, including FPS, clip duration, and frame stride, as shown in Fig. 7. The results show that the performance remains generally stable over a broad range of sampling configurations. Increasing FPS provides more temporal observations and consistently improves detection performance, while varying the clip duration introduces limited fluctuations. For frame stride, moderate temporal intervals maintain stable performance, indicating that sufficient temporal continuity is beneficial for capturing the continuity of 3D motion.

5 Conclusions

In this paper, we propose 3DMoC, a novel training-free framework that leverages the intrinsic 3D motion consistency of real-world scenes to detect AI-generated videos. Unlike existing methods that mainly focus on visual artifacts or inter-frame variations in the 2D space, our 3DMoC identifies appearance and motion anomalies in 3D space, enabling more accurate fake video detection. Specifically, we elaborate a Tracked Appearance Consistency module and a 3D Motion Continuity module that collaborate to capture appearance discrepancies and sudden motion changes of features aligned along 3D trajectories. Experimental results demonstrate that our 3DMoC effectively detects generated videos across diverse datasets and exhibits strong generalization without requiring fake samples for training.

AI USE STATEMENT

In this work, we used generative AI tools to polish the content. We did not use generative AI tools for data generation or concept design. We have reviewed all work completed with the assistance of AI. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code will be released after the paper is accepted.

REFERENCES

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7310–7320, 2024a.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark. *arXiv preprint arXiv:2405.19707*, 2024b. doi: 10.48550/arXiv.2405.19707.

Wenhan Chen, Sezer Karaoglu, and Theo Gevers. Grab-3D: Detecting AI-Generated Videos from 3D Geometric Temporal Consistency. *arXiv preprint arXiv:2512.13665*, 2025.

Davide Alessandro Coccomini, Giorgos Kordopatis Zilos, Giuseppe Amato, Roberto Caldelli, Fabrizio Falchi, Symeon Papadopoulos, and Claudio Gennaro. MINTIME: Multi-Identity Size-Invariant Video Deepfake Detection. *IEEE Transactions on Information Forensics and Security*, 19:6084–6096, 2024.

Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and Luisa Verdoliva. Zero-Shot Detection of AI-Generated Images. In *European Conference on Computer Vision*, pp. 54–72, 2025.

Huidong Feng, Wentao Chen, Jie Chen, Xinqi Cai, Ruolong Ma, Yinglin Zheng, Yuxin Lin, and Ming Zeng. CoCoVideo: The High-Quality Commercial-Model-Based Contrastive Benchmark for AI-Generated Video Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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 ACM international conference on multimedia*, pp. 3473–3481, 2021.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. In *International Conference on Learning Representations*, pp. 34630–34648, 2024.

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free Detection of Generated Videos via Spatial-Temporal Likelihoods. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16299–16310, 2026.

Haojin He, Hao Tan, Zichang Tan, Ajian Liu, and Jun Wan. MotionPhys: Detecting AI-Generated Videos via Physical Consistency of Optical-Flow Trajectories. *arXiv preprint arXiv:2608.20770*, 2026.

Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Bill Yuchen Lin, and Wenhu Chen. VideoScore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2105–2123, 2024a.

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. RIGID: A Training-free and Model-Agnostic Framework for Robust AI-Generated Image Detection. *arXiv preprint arXiv:2405.20112*, 2024b.

Jongyeop Hyun and Hyounghun Kim. SPLIT: Training-Free AI-Generated and Partially Edited Video Detection via Spatial Patch-Level Incoherence and Temporal Roughness. In *European Conference on Computer Vision*, 2026.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-Generated Video Detection via Perceptual Straightening. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20672–20705, 2025.

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your One-Stop Solution for AI-Generated Video Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4458–4470, 2026.

Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding Language-Image Pretrained Models for General Video Recognition. In *European Conference on Computer Vision*, pp. 1–18, 2022.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. AEROBLADE: Training-Free Detection of Latent Diffusion Images Using Autoencoder Reconstruction Error. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Chuangchuang Tan, Huan Liu, 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual Geometry Grounded Transformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5294–5306, 2025.

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.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-Driven Spatiotemporal Modeling for AI-Generated Video Detection. In *Advances in Neural Information Processing Systems*, volume 38, pp. 174683–174730, 2025.

Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 59570–59596, 2024.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-Free AI-Generated Video Detection Using Second-Order Features. In *IEEE/CVF International Conference on Computer Vision*, pp. 1979–1989, 2025.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring Temporal Coherence for More General Video Face Forgery Detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 15024–15034, 2021.

Appendix of 3DMoC

A Local-score Aggregation Formula

For each adjacent frame pair, we first aggregate the local appearance differences over all valid tracked points:

$$L_t^{\text{local}} = \max_{n \in \Omega_t} \ell_{t,n}^{\text{local}}, \tag{17}$$

where Ω_t denotes the set of valid tracked correspondences between frames t and $t+1$. We then aggregate the frame-pair statistics over time:

$$L^{\text{local}} = \max_{1 \le t \le T-1} L_t^{\text{local}}. \tag{18}$$

Finally, we normalize L^{local} with respect to the calibration reference distribution $\mathcal{C}_t$:

$$p^t = \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} \mathbb{I}\left(c \le L^{\text{local}}\right). \tag{19}$$

B Experiments Settings

Dataset Details. We provide the detailed composition of the three evaluation benchmarks used in our experiments. These datasets cover different data scales and real-fake distributions, ranging from relatively balanced settings to highly imbalanced ones. GenVideo contains 18,276 videos in total, including 10,000 real videos from MSR-VTT and 8,276 generated videos from ten different generators. ComGenVid contains 5,100 videos, consisting of 1,700 real MSVD videos and 3,400 generated videos from Sora and VEO3. VideoFeedback contains 760 videos, including 80 real videos from DiDeMo and Panda70M and 680 generated videos produced by eleven different video generation models. The detailed subset statistics are summarized in Table 4.

Table 4: Detailed composition of the three evaluation datasets.

VideoFeedback

Type	Source	#
Real	DiDeMo	43
Real	Panda70M	37
Real Total		**80**
Fake	AnimateDiff	29
Fake	Fast-SVD	20
Fake	Hotshot-XL	72
Fake	LaVie-base	68
Fake	LVDM	81
Fake	ModelScope	96
Fake	Pika	87
Fake	SoRA-Clip	20
Fake	Text2Video-Zero	79
Fake	VideoCrafter2	91
Fake	ZeroScope-576w	37
Fake Total		**680**
Total		**760**

GenVideo

Type	Source	#
Real	MSR-VTT	10,000
Fake	Crafter	1,398
Fake	Gen2	1,380
Fake	HotShot	700
Fake	Lavie	1,400
Fake	ModelScope	700
Fake	MoonValley	626
Fake	MorphStudio	700
Fake	Show_1	700
Fake	Sora	56
Fake	WildScrape	616
Fake Total		**8,276**
Total		**18,276**

ComGenVid

Type	Source	#
Real	MSVD	1,700
Fake	Sora	1,700
Fake	VEO3	1,700
Fake Total		**3,400**
Total		**5,100**

Comparison Methods. We do not directly compare with Grab-3D Chen et al. (2025) and MotionPhys He et al. (2026) because they employ a supervised paradigm that differs from our training-free setting. Grab-3D introduces a learnable geometry-aware Transformer to model temporal consistency of vanishing-point-based 3D geometric cues, while MotionPhys extracts handcrafted motion descriptors from 2D optical-flow trajectories and employs a trained classifier for final prediction. In contrast, 3DMoC does not train any task-specific detector and directly performs detection using frozen representations and statistics derived from reconstructed 3D point trajectories.

Robustness Evaluation configurations. To evaluate robustness to common video degradations, we independently apply Gaussian blur, additive Gaussian noise, JPEG compression, and random cropping to the sampled frames while keeping the frozen DINO/VGGT encoders, temporal sampling strategy, calibration, and score-fusion rule unchanged. Each perturbation is evaluated at five severity levels: Gaussian blur uses $(r, \sigma, k) \in \{(1, 0.5, 3), (2, 1.0, 5), (3, 1.5, 7), (5, 2.5, 11), (10, 5.0, 21)\}$; Gaussian noise uses $\sigma \in \{0.02, 0.05, 0.10, 0.20, 0.50\}$; JPEG compression uses quality factors $q \in \{80, 50, 30, 10, 1\}$; and random cropping uses scale ranges $\{(0.85, 0.90), (0.70, 0.85), (0.50, 0.80), (0.30, 0.90), (0.08, 1.00)\}$. The perturbed frames are processed without perturbation-specific retraining or recalibration, and AP and AUC are reported separately for each perturbation and severity level.

Temporal sampling configurations. To evaluate the robustness of the detector to temporal sampling, we independently varied the frame rate, temporal duration, and temporal stride while keeping the remaining inference components unchanged. The horizontal axis in Fig. 7 denotes the configuration index rather than a common physical quantity. For the FPS group, indices 1-4 correspond to frame rates of $\{2, 4, 8, 16\}$ FPS, respectively. For the duration group, indices 1-4 correspond to temporal windows of $\{1, 2, 3, 4\}$ seconds. For the temporal-stride group, indices 1-4 correspond to frame intervals of $\{1, 2, 3, 4\}$ frames. Thus, the same horizontal position represents the same ordinal severity level within each group, but not the same numerical parameter value across groups. For each configuration, the selected frames were processed by the same frozen DINO and VGGT encoders, and the resulting appearance and geometric scores were evaluated using the identical AP and AUC protocol.

Main parameter settings. We use the frozen DINOv3 ViT-L/16 checkpoint and the frozen VGGT-1B checkpoint. Unless otherwise stated, each video is represented by eight frames sampled according to the dataset index sequence. DINOv3 receives 224×224 images, while VGGT receives 518×518 padded inputs. We retain the non-CLS DINO patch tokens and map VGGT track coordinates to the patch grid using normalized coordinates followed by nearest-integer rounding and boundary clipping. All model checkpoints, preprocessing parameters, frame-index files, and software versions are released with the code. In the appearance branch, the CLS aggregation weight is set to $\lambda_1 = 0.8$. In the geometric branch, the entropy aggregation index is $\gamma = 0.05$. The motion weight is $\lambda_{mot} = 0.10$. These parameters are kept constant across all main experiments and are not modified across different datasets. The robustness and ablation experiments are conducted on the VideoFeedback dataset. In Fig. 3, we randomly select 50 real videos and 50 generated videos and compute their relative 3D motion using Eq. (11) to analyze the distributional difference between real and generated videos. All other ablation, robustness, and sensitivity studies are performed on VideoFeedback with the default configuration. The detailed algorithm for 3DMoC is shown in Algorithm 1.

Calibration Set. We adopted the calibration scheme proposed by STALL Hayun et al. (2026), following the same real-only calibration protocol. Specifically, it is constructed from 33K real videos of VATEX, which is completely disjoint from all evaluation benchmarks used in this work(VideoFeedback, GenVideo, and ComGenVid), thereby avoiding data overlap and information leakage. Only real videos are involved in calibration. For the global appearance statistics branch, frozen DINOv3 features extracted from the calibration videos are used to estimate the reference center μ_s, decorrelation transformation W_s, and the corresponding spatial score distribution. For the local appearance consistency branch, normalized inter-frame feature variations are used to estimate the temporal reference statistics μ_t and W_t as well as the corresponding reference score distribution. All calibration statistics are computed offline once and kept fixed across all benchmarks. No generated videos, task-specific optimization, or model fine-tuning are involved in this process. Results on calibration scale sensitivity and other ablation effects can be found in STALL Hayun et al. (2026). Our method is training-free in the sense that no detector parameters are optimized using real or generated training samples.

Algorithm 1 Overall inference procedure of our 3DMoC.

Require: Input videos V; frozen DINOv3 encoder $\mathcal{F}_D$; frozen VGGT model $\mathcal{F}_G$; spatial reference statistics $(\boldsymbol{\mu}_s, \mathbf{W}_s)$; temporal reference statistics $(\boldsymbol{\mu}_t, \mathbf{W}_t)$; motion aggregation exponent γ; fusion weight λ_{mot}
Ensure: Final detection score S_{final}

$\{I_t\}_{t=1}^{T} \leftarrow \text{Sample}(V)$ ▷ Sample video frames
for $t = 1$ **to** T **do**
  $(\mathbf{e}_t, \mathbf{F}_t) \leftarrow \mathcal{F}_D(I_t)$ ▷ Extract CLS and patch features
end for
$(\mathbf{Q}, \mathbf{P}, \mathbf{M}) \leftarrow \mathcal{F}_G(\{I_t\}_{t=1}^{T})$ ▷ Obtain correspondences and 3D points
$p^s \leftarrow \text{GlobalAppearance}(\{\mathbf{e}_t\}, \boldsymbol{\mu}_s, \mathbf{W}_s)$ ▷ Global appearance statistics
for $t = 1$ **to** $T - 1$ **do**
  $\Omega_t \leftarrow \text{ValidTrack}(\mathbf{Q}, \mathbf{M}, t)$
  for all $n \in \Omega_t$ **do**
    $\mathbf{x}_{t,n} = (u_{t,n}, v_{t,n})$
    $(\mathbf{g}_{t,n}, \mathbf{g}_{t+1,n}) \leftarrow \text{Align}(\mathbf{F}_t, \mathbf{F}_{t+1}, \mathbf{q}_{t,n})$
    $\Delta\mathbf{g}_{t,n} \leftarrow \mathbf{g}_{t+1,n} - \mathbf{g}_{t,n}$
    $\widehat{\Delta\mathbf{g}}_{t,n} \leftarrow \Delta\mathbf{g}_{t,n} / \|\Delta\mathbf{g}_{t,n}\|_2$
  end for
  $\ell_t^{\text{local}} \leftarrow \text{AppearanceDifference}(\{\widehat{\Delta\mathbf{g}}_{t,n}\}_{n\in\Omega_t}, \boldsymbol{\mu}_t, \mathbf{W}_t)$
end for
$p^t \leftarrow \text{TemporalNormalize}(\{\ell_t^{\text{local}}\}_{t=1}^{T-1})$ ▷ Tracked appearance consistency
$S_{\text{app}} \leftarrow \lambda_1 p^s + (1 - \lambda_1) p^t$
for $t = 1$ **to** $T - 2$ **do**
  for all $n \in \Omega_t$ **do**
    $\Delta\mathbf{p}_{t,n} \leftarrow \mathbf{p}_{t+1,n} - \mathbf{p}_{t,n}$
    $\Delta^2\mathbf{p}_{t,n} \leftarrow \Delta\mathbf{p}_{t+1,n} - \Delta\mathbf{p}_{t,n}$
    $m_{t,n} \leftarrow \frac{\|\Delta\mathbf{p}_{t,n}\|_2 + \|\Delta\mathbf{p}_{t+1,n}\|_2}{2}$
    $r_{t,n} \leftarrow \rho\left(\frac{\|\Delta^2\mathbf{p}_{t,n}\|_2}{\rho(m_{t,n}) + \epsilon}\right)$ ▷ Relative motion
  end for
end for
$\mathbf{q} \leftarrow \frac{r_{t,n}}{\sum_{(i,j)} r_{i,j} + \epsilon}$, ▷ Normalization of Relative-motion matrix
$H \leftarrow \text{Entropy}(\mathbf{q})$ ▷ Video-specific motion complexity
$S_{\text{mot}} \leftarrow \sum_{t,n} \left(\frac{\max(r_{t,n}, 0)}{H + \epsilon}\right)^{\gamma}$ ▷ 3D motion-continuity score
$\widetilde{S}_{\text{app}} \leftarrow \mathcal{N}(S_{\text{app}})$
$\widetilde{S}_{\text{mot}} \leftarrow \mathcal{N}(S_{\text{mot}})$ ▷ Robust normalization
$S_{\text{final}} \leftarrow \widetilde{S}_{\text{app}} + \lambda_{\text{mot}} \widetilde{S}_{\text{mot}}$ ▷ Score fusion
return S_{final}

C Additional Experiments

Ablation on Aggregation and Normalization. We further investigate the effects of the CLS-Patch aggregation weight, entropy aggregation parameter γ, and score normalization strategy. As shown in Table 5, increasing the CLS-Patch aggregation weight generally improves the AUC, while the AP remains relatively stable. The best AP is obtained at a weight of 0.8, whereas larger weights slightly favor AUC, indicating that a moderate combination of global CLS features and local patch consistency provides a balanced representation.

For the entropy aggregation parameter γ, Table 6 shows that the performance varies only marginally over a wide range of values. In particular, AP remains around 0.879 and AUC around 0.818, demonstrating that the proposed 3D motion characterization is insensitive to the choice of γ. We therefore set $\gamma = 0.05$ in our experiments.

Finally, we compare different score normalization strategies in Table 7. MAD normalization followed by the tanh mapping achieves the best overall performance, reaching 0.8814 AP and 0.8205

Table 5: CLS-Patch aggregation weight λ_1.

λ_1	AP	AUC
0.6	0.8760	0.8191
0.7	0.8812	0.8182
0.8	0.8814	0.8205
0.9	0.8814	0.8194
1.0	0.8764	0.8189

Table 6: Entropy aggregation parameter.

γ	AP	AUC
0.01	0.8787	0.8182
0.02	0.8794	0.8194
0.05	0.8814	0.8205
0.075	0.8793	0.8182
0.10	0.8792	0.8180

Table 7: Score normalization methods.

Norm	AP	AUC
tanh	0.8792	0.8180
MAD	**0.8814**	**0.8205**
Rank	0.8765	0.8129

Table 8: Ablation study on different 3D motion statistics.

Motion Statistic	AP	AUC
Raw 3D Displacement	0.8593	0.7859
Second-order Variation	0.8602	0.7903
Relative Motion	0.8717	0.7966
Relative Motion + Entropy	**0.8814**	**0.8205**

AUC, outperforming both tanh and Rank normalization. This suggests that standardizing the appearance and motion scores before fusion better aligns their distributions and enables more effective complementary score aggregation.

Ablation on Motion Statistics. We further investigate different formulations for characterizing 3D motion, as shown in Table 8. Using raw 3D displacement data directly already provides effective detection cues, while replacing first-order displacements with second-order motion variations results in a slight improvement in the AP, indicating that changes in motion evolution provide additional discriminative information beyond absolute displacement. Further normalizing the second-order changes by the local motion amplitude yields relative motion statistics, which achieve the highest AP. This suggests that scale-normalized motion changes better highlight anomalous local motion patterns while reducing the influence of absolute motion amplitude. These results demonstrate that entropy-based aggregation can further integrate the distribution of motion responses at the video level.

D LIMITATIONS.

Although our 3DMoC exhibits strong cross-generator generalization, its performance may degrade in a few challenging cases. For example, in nearly static videos, limited motion weakens 3D motion cues, or in scenes with persistent occlusions or large textureless regions, the reliability of correspondence and trajectory reconstruction decreases. However, since such videos typically contain limited dynamic information, the risk of conveying misleading motion semantics is low.