DOES AI-GENERATED VIDEO DETECTION REALLY NEED TEMPORAL MODELING?

Anonymous authors
Paper under double-blind review

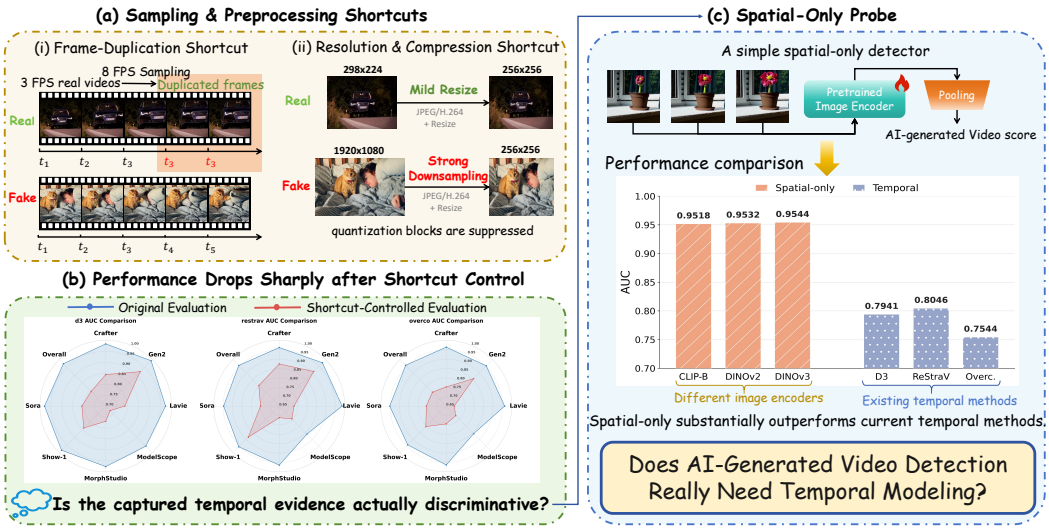


Figure 1: Motivation overview. (a) Examples of sampling- and preprocessing-induced shortcuts in existing temporal evaluation. (b) Performance of representative temporal detectors before and after shortcut control on GenVideo. (c) A simple spatial-only probe, which substantially outperforms existing temporal methods under the same setting, calling the role of temporal modeling into question.

ABSTRACT

Temporal information is a natural source of evidence for AI-generated video detection (AIGVD), yet under shortcut-controlled evaluation on GenVideo, a simple spatial-only detector reaches about 0.95 AUC, substantially outperforming existing temporal detectors at around 0.80. Rather than concluding that temporal modeling is unnecessary, we investigate why current methods fail to exploit temporal evidence effectively. Our analysis suggests a decomposition hypothesis: observed inter-frame variation entangles generation-related evidence with substantial non-generative variation from preprocessing, camera motion, and object motion. We examine two consequences of this entanglement. First, scalar summaries commonly used by current methods can map generation-related and non-generative changes to overlapping values, whereas element-wise differences preserve richer structure and substantially improve detection. Second, progressively reducing major non-generative components yields further consistent gains, making temporal evidence competitive with spatial evidence on GenVideo and stronger in AUC on the unseen GenVidBench benchmark. Moreover, the resulting debiased temporal evidence complements spatial evidence, with their fusion further improving performance. Across controlled representation analyses, progressive reduction of non-generative variation, and cross-benchmark evaluation, our results consistently show that properly modeled temporal evidence remains highly discriminative and valuable for AIGVD.

1 INTRODUCTION

With the rapid development of generative models (OpenAI, 2025; Wan et al., 2025), AI-generated videos have become increasingly realistic and difficult to distinguish from real ones. Their growing

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.

realism also raises concerns about misuse, including misinformation (Zhou et al., 2023; Xu et al., 2023a), fraud (Shetye et al., 2025; Martinson, 2026), and malicious manipulation (Easttom, 2025; Sophia, 2025), making AI-generated video detection (AIGVD) an essential research problem.

Unlike images, videos contain temporal information across frames, making temporal modeling a natural source of evidence for AIGVD. Beyond general sequence architectures such as recurrent networks (Liu et al., 2024) and Transformers (Kundu et al., 2025), recent methods tailored to AIGVD directly characterize inter-frame variation, based on the premise that generation-related artifacts manifest as abnormal temporal evolution (Zheng et al., 2025; Internò et al., 2025; Brokman et al., 2026; Ben Hayun et al., 2026). For example, D3 (Zheng et al., 2025) models second-order temporal variation, ReStraV (Internò et al., 2025) exploits changes in angular relationships between consecutive frame representations, and Over-Coherence (Brokman et al., 2026) measures temporal similarity across adjacent frames. These examples illustrate a common design choice in recent AIGVD temporal modeling: summarizing inter-frame variation, often using scalar statistics such as distances, angles, or similarities.

However, recent work (Michels et al., 2026) has shown that temporal AIGVD methods can exploit preprocessing-induced shortcuts, including frame duplication caused by temporal resampling and resolution- or compression-dependent artifacts, as illustrated in Figure 1(a). Once these shortcuts are controlled, the performance of representative temporal detectors drops substantially (Figure 1(b)). This raises an important question about what current temporal detectors are actually capturing. To establish a reference point, we remove temporal modeling entirely and construct a deliberately simple spatial-only detector that applies an image encoder independently to each frame and aggregates the resulting predictions. Surprisingly, under the same shortcut-controlled GenVideo (Chen et al., 2026) setting, several generic image encoders achieve around 0.95 AUC, substantially outperforming existing temporal methods, whose best performance is around 0.80 (Figure 1(c)). This striking reversal appears to challenge a basic premise of video detection and raises a fundamental question: **Does AI-generated video detection really need temporal modeling?**

Rather than taking this result as evidence that temporal information is unnecessary, we investigate why current AIGVD methods fail to exploit it effectively. Our analysis leads to a *decomposition hypothesis*: the observed inter-frame variation does not consist solely of generation-related evidence, but instead mixes multiple sources of variation arising from video preprocessing, camera motion, object motion, and other factors. Crucially, the non-generative components can be sufficiently strong to obscure the generation-related one. This entanglement not only introduces irrelevant variation, but may also encourage detectors to rely on spurious correlations between these non-generative factors and whether a video is real or AI-generated, ultimately harming generalization.

This hypothesis suggests two potential failure points in current temporal modeling: *how inter-frame variation is represented* and *what variation it contains*. We investigate both systematically. First, representative methods commonly compress inter-frame variation into scalar summaries, such as distances or angles. Under our decomposition, heterogeneous sources of variation can map to overlapping scalar values, making generation-related and non-generative changes difficult to separate. In a controlled comparison, replacing such scalar summaries with element-wise discrepancies preserves richer variation structure and substantially improves temporal detection.

Second, even with a richer representation, the underlying inter-frame signal still contains substantial non-generative variation. We therefore examine these sources progressively. Controlling sampling and compression improves temporal detection; compensating for global camera motion yields a further gain; and partially accounting for local object motion improves it again. The consistency of this progression supports our decomposition hypothesis: temporal evidence becomes increasingly reliable as the influence of non-generative variation is reduced.

After these factors are accounted for, the resulting debiased temporal evidence becomes competitive with spatial evidence on GenVideo and surpasses it in AUC on the unseen GenVidBench benchmark. More importantly, combining spatial evidence with debiased temporal evidence further improves performance without introducing additional learnable parameters, indicating that the two provide complementary information when temporal variation is modeled appropriately. Together, these results suggest that the weakness of existing temporal detectors does not stem from an inherent lack of useful temporal evidence, but from how inter-frame variation is represented and how strongly it remains entangled with ordinary video variation.

Compared with previous work[1], our main contributions are summarized as follows:

- We revisit the role of temporal modeling in AIGVD and uncover a counterintuitive result: under shortcut-controlled evaluation, a simple spatial-only detector substantially outperforms representative temporal methods designed specifically for AIGVD.
- We propose a decomposition hypothesis to explain this failure and analyze temporal evidence along two dimensions: how inter-frame variation is represented and what sources of variation it contains. Controlled experiments show that preserving richer variation structure and progressively reducing non-generative components consistently improve temporal detection.
- We show that the resulting debiased temporal evidence becomes competitive with spatial evidence on GenVideo and surpasses it in AUC on the unseen GenVidBench benchmark. Its parameter-free fusion with spatial evidence further improves performance, demonstrating that appropriately modeled temporal evidence provides complementary information for AIGVD.

2 Why Do Current Temporal Methods Fail?

Analysis Setup. Given a video $\mathcal{X} = \{x_t\}_{t=1}^{T}$, we denote by $f(x_t)$ the representation of its t-th frame. Following the evaluation setting of D3 (Zheng et al., 2025), we conduct our analysis on the GenVideo benchmark (Chen et al., 2026), constructing the training set using real videos from Youku-mPLUG (Xu et al., 2023b) and AI-generated videos from Pika (Pika Art, 2022). To isolate the role of temporal modeling, our main analysis focuses on detectors whose predictions rely on temporal evidence alone (*i.e.,* D3 (Zheng et al., 2025), ReStraV (Internò et al., 2025), and OverCoherence (Brokman et al., 2026)), avoiding confounding from spatial evidence. Methods that jointly exploit spatial and temporal evidence are considered separately in Appendix D.2. All baselines are trained and evaluated under the same setting using their official implementations whenever available. Throughout the analysis, we adopt the shortcut-controlled preprocessing protocol of Michels et al. (2026); implementation details are provided in Appendix C.5. We report AUC throughout the GenVideo analyses because shortcut control leaves substantially different numbers of valid samples across generators. We retain the full real set rather than subsampling it to match each generator, providing a more comprehensive evaluation while avoiding arbitrary sampling bias. Since AP is sensitive to class prevalence, AUC is used as the primary metric.

2.1 Detection Without Temporal Modeling

Previous AIGVD temporal methods achieve strong performance by exploiting inter-frame dynamics, but recent studies (Michels et al., 2026) show that their performance drops substantially once preprocessing-induced shortcuts are controlled. This suggests that the temporal evidence captured by current methods may be far less discriminative than previously believed, leaving the actual contribution of temporal information unclear. Since temporal information is what distinguishes videos from a set of independent frames, we can assess its contribution directly by measuring how well AIGVD can be solved from individual frames alone. To this end, we construct a deliberately simple spatial-only baseline using a generic image encoder that is not specifically designed for AI-generated content detection.

Specifically, we treat each frame independently as an image and aggregate the top-k frame predictions into a video-level prediction. To ensure that this observation is not tied to a particular visual representation, we instantiate the baseline with several commonly used pretrained encoders, including CLIP-B (Radford et al., 2021), DINOv2 (Oquab et al., 2024), and DINOv3 (Siméoni et al., 2026), and apply lightweight LoRA adaptation (Hu et al., 2022) to each.

As shown in Figure 1, all three spatial-only baselines achieve substantially higher AUC than existing temporal methods under the same shortcut-controlled setting. In particular, even this deliberately simple frame-level approach reaches an AUC above 0.95, whereas the temporal methods remain around 0.75–0.80. Moreover, this advantage is consistent across the generators in Figure 2, indicating that the overall improvement is not driven by only a few subsets. It also holds when the encoder is frozen: the DINOv3 baseline drops only from 0.954 to 0.952, further showing that spatial information alone can support strong AIGVD performance (Appendix D.3). The strong performance of

[1] A detailed discussion of related work is provided in Appendix A.

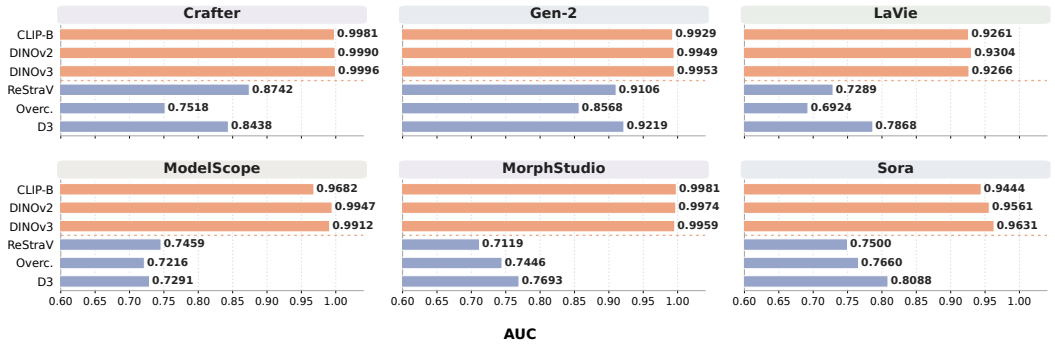


Figure 2: Per-generator comparison on a subset of GenVideo generators under the shortcut-controlled setting. The spatial-only detectors (orange) use DINOv2, DINOv3, and CLIP-B as frame-level backbones, while D3, ReStraV, and Over-Coherence (blue) are existing temporal methods.

the spatial-only baseline naturally raises a fundamental question: does AI-generated video detection really need temporal modeling? However, the poor performance of existing temporal methods alone cannot answer this question, as it may reflect limitations in how temporal information is currently exploited rather than the absence of useful temporal evidence. We therefore investigate the root causes of this failure to determine whether temporal modeling still holds potential for AIGVD.

2.2 A Decomposition Hypothesis for Temporal Evidence

Current AIGVD temporal methods characterize temporal evidence using different quantities derived from consecutive frame representations, such as second-order differences, angular changes, and cosine similarities. Despite their different forms, these quantities all reflect how frame representations change over time and can therefore be viewed as different characterizations of the observed inter-frame variation. Whether these quantities reliably indicate generation then depends on what that variation actually contains. Camera motion can induce global changes across frames, natural object motion produces spatially varying changes, and even sampling or compression can alter the observed pattern. Importantly, generation-related artifacts constitute only one of many such sources. Modeling the overall inter-frame variation directly therefore leaves subtle generation-related changes mixed with substantial non-generative ones, making them difficult to isolate.

Based on the above analysis, we conceptually decompose the observed inter-frame variation into generation-related and non-generative components:

$$\Delta z_t = \underbrace{\Delta z_t^{\text{object motion}} + \Delta z_t^{\text{camera motion}} + \Delta z_t^{\text{compression}} + \Delta z_t^{\text{sampling}} + \dots}_{\text{non-generative sources}} + \Delta z_t^{\text{gen}}, \quad (1)$$

where Δz_t broadly denotes the representation of inter-frame variation, which may be summarized using scalar measures such as distances, angles, or similarities, or retained as element-wise differences. $\Delta z_t^{\text{object motion}}$, $\Delta z_t^{\text{camera motion}}$, $\Delta z_t^{\text{compression}}$, $\Delta z_t^{\text{sampling}}$ denote variation induced by object motion, camera motion, compression, and frame sampling, respectively, while Δz_t^{gen} denotes the generation-related component. This decomposition leads to our central hypothesis:

> **Decomposition Hypothesis.** *Non-generative variation can obscure generation-related evidence within the observed inter-frame variation, and can further encourage detectors to exploit spurious correlations with whether a video is real or AI-generated. Reducing the influence of this non-generative variation should therefore yield more reliable temporal evidence.*

This hypothesis is also consistent with the surprising strength of the spatial-only baseline in Section 2.1. By treating each frame independently, frame-level detection never forms Δz_t at all, and therefore avoids the non-generative temporal variation in Eq. (1). However, it also discards temporal evidence entirely, leaving open the question of how such evidence can be modeled reliably. Under our decomposition hypothesis, this depends on two aspects: how the mixed variation Δz_t is represented, and how strongly its non-generative components affect the observed signal.

3 DEBIASING TEMPORAL EVIDENCE

Guided by the decomposition hypothesis above, we now investigate how temporal evidence can be made more reliable. We first revisit how inter-frame variation is represented, examining whether scalar summaries make the resulting temporal evidence more susceptible to being dominated by non-generative variation. We then progressively reduce several major non-generative components, including sampling and compression, camera motion, and object motion. These analyses lead to a debiased temporal detector, which we finally compare with spatial evidence to examine whether temporal evidence remains discriminative and provides information beyond individual frames.

3.1 REPRESENTATION OF INTER-FRAME VARIATION

A common design among representative AIGVD temporal methods is to summarize inter-frame change using scalar measures, such as distances, angles, or similarities, and distinguish real from AI-generated videos based on the resulting scalar values (Zheng et al., 2025; Internò et al., 2025; Brokman et al., 2026). Under our decomposition hypothesis, such a scalar aggregates every source in Eq. (1) into a single number. Non-generative sources alone already span a wide range of this number: a nearly static video produces a very small value, whereas fast object or camera motion produces a much larger one, even though both videos are real. Generation-related and non-generative variation can therefore map to overlapping values, which a single scalar cannot separate.

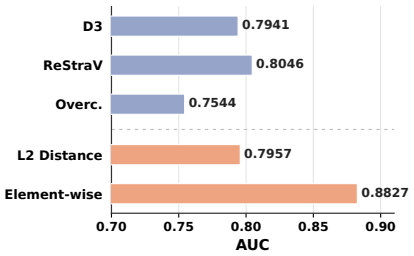


Figure 3: Comparison between scalar and element-wise representations on GenVideo over all generators.

However, existing temporal methods differ in encoder, architecture, and training, making direct cross-method comparisons insufficient to isolate the effect of representation. We therefore construct a controlled comparison using a frozen DINOv3 as the frame encoder. Given the representations of two consecutive frames, we first compute their ℓ_2 distance as a scalar temporal measure. We then replace this scalar measure with the element-wise difference between the two representations, while keeping all remaining settings unchanged. As shown in Figure 3, the generic ℓ_2 variant reaches an AUC of 0.796, already comparable to purpose-built scalar temporal methods, suggesting that a more specialized scalar statistic alone does not resolve the limitation. The element-wise variant raises AUC to 0.883 (per-generator results are reported in Table 3). Since only the representation differs, this controlled comparison isolates the effect of representation form. As the decomposition predicts, compressing mixed inter-frame variation into a single scalar makes temporal evidence more susceptible to non-generative variation, whereas retaining element-wise discrepancies preserves richer variation structure and yields substantially more discriminative evidence. Yet, improving the representation of Δz_t does not remove the non-generative components themselves. We therefore next examine how these components affect temporal evidence by progressively reducing their influence.

3.2 PREPROCESSING-INDUCED VARIATION

We begin with sampling and compression, as they provide the most directly controllable non-generative sources in our decomposition and have been identified as preprocessing-induced shortcuts in prior work (Michels et al., 2026). That work uses such controls to show that detectors exploit these shortcuts, inflating reported performance. Our decomposition raises a different question: does reducing these components in training make the learned temporal evidence more reliable? To isolate this effect, we keep the shortcut-controlled test set fixed and compare the base temporal detector of Section 3.1, trained on videos with these variations retained, against an S&C-controlled variant trained on the same source videos. Frame duplication and resolution-dependent compression are pipeline-induced rather than intrinsic to the videos and can be avoided through alternative preprocessing choices, so the controlled test set reflects the videos rather than artifacts of the preprocessing pipeline. Conceptually, reducing $\Delta z_t^{\text{sampling}}$ and $\Delta z_t^{\text{compression}}$ in Eq. (1) yields:

$$\Delta z_t \approx \Delta z_t^{\text{camera motion}} + \Delta z_t^{\text{object motion}} + \cdots + \Delta z_t^{\text{gen}}, \tag{2}$$

Table 1: Effect of progressively accounting for non-generative sources on temporal detection in GenVideo. The base temporal detector is the element-wise detector of Section 3.1, here trained without sampling and compression control. AUC is computed jointly over all generators; per-generator results are in Appendix D.1.

Method / Setting	Sampling / Compression	Camera Motion	Object Motion	Overall AUC	Performance
Existing Temporal Methods					
ReStraV (NeurIPS'25)	–	–	–	0.7557	
ReStraV (NeurIPS'25) + S&C Control	✓	–	–	0.8046	
Controlled Progressive Analysis					
Base Temporal Detector	–	–	–	0.8523	
+ S&C Control	✓	–	–	0.8827	
+ Camera Motion Compensation	✓	✓	–	0.9301	
+ Object Motion Compensation	✓	✓	✓	**0.9485**	
Spatial Evidence and Fusion					
Spatial-Only (DINOv3)	–	–	–	0.9544	
Spatial + Debiased Temporal	✓	✓	✓	**0.9770**	

where the sampling- and compression-related components have been substantially reduced. As shown in Table 1, controlling these factors yields a clear improvement. Notably, the same sampling/compression control also improves ReStraV, providing cross-method evidence for this effect.

3.3 Natural Motion-Induced Variation

Sampling and compression can be controlled during preprocessing, whereas motion-induced variation is inherent to the video content and requires a different treatment. Camera and object motion can both alter the observed inter-frame variation without being directly related to the generation process, potentially introducing noise or spurious cues into Δz_t. We therefore examine whether reducing their influence further improves temporal detection. Both interventions share the same strategy: rather than comparing two frames directly, we first warp the earlier frame to account for the motion between them, and then measure what remains unexplained.

Camera Motion. Among the motion-related non-generative sources in Eq. (1), camera-induced variation has a particularly interpretable physical structure: much of its global effect, such as translation, rotation, zoom, and perspective change, can be approximated by a low-dimensional geometric transformation estimated directly from two adjacent frames. We therefore use homography (Dubrofsky, 2009) to estimate this transformation and align the two frames before computing Δz_t. Specifically, given x_t and x_{t+1}, we estimate a homography matrix H_t and warp x_t toward x_{t+1}, giving the aligned frame $\tilde{x}_{t+1}^{\text{cam}} = \mathcal{W}(x_t; H_t)$, where $\mathcal{W}$ denotes homography warping. Further implementation details are provided in Appendix B.2. The camera-motion-adjusted inter-frame variation can then be written as:

$$\Delta z_t : \quad |f(x_{t+1}) - f(x_t)| \;\longrightarrow\; \left|f(x_{t+1}) - f(\tilde{x}_{t+1}^{\text{cam}})\right|, \tag{3}$$

where $|\cdot|$ denotes the element-wise absolute value. Conceptually, based on Eq. (2), reducing the camera-motion component further simplifies the decomposition to

$$\Delta z_t \approx \Delta z_t^{\text{object motion}} + \cdots + \Delta z_t^{\text{gen}}. \tag{4}$$

Reducing this physically interpretable variation removes part of the normal geometric change caused by camera motion. As a result, the residual inter-frame variation is less dominated by naturally occurring global motion, making finer unexplained discrepancies easier to observe.

Object Motion. We next consider object motion. Unlike camera motion, which can often be approximated by a low-dimensional global transformation, natural object motion varies across regions and objects, resulting in a high-dimensional motion field that is much harder to model completely. We therefore estimate it densely rather than parametrically, using a pretrained RAFT (Teed & Deng, 2020) model to predict local motion between adjacent frames after global camera alignment.

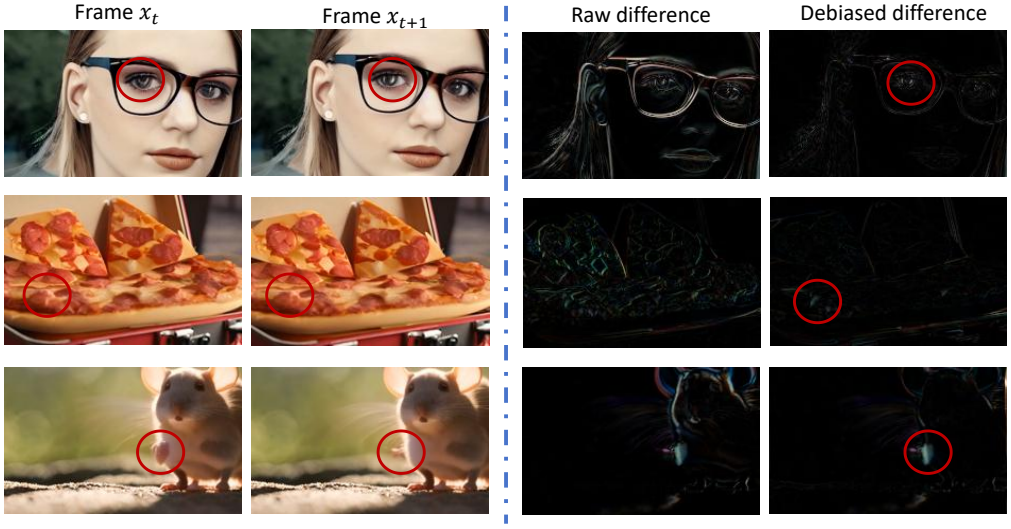


Figure 4: Case study of raw and debiased inter-frame differences on representative AI-generated videos. Red circles mark high-response regions in the debiased difference, which correspond to visually implausible changes when compared with the original frame pair.

Similarly, we estimate the flow field F_t from the camera-motion-aligned frame $\tilde{x}_{t+1}^{\text{cam}}$ to x_{t+1}, and further warp the former according to this local motion: $\tilde{x}_{t+1}^{\text{cam+obj}} = \mathcal{W}_{\text{flow}}\left(\tilde{x}_{t+1}^{\text{cam}}; F_t\right)$. This compensation mainly reduces changes that can be explained by stable cross-frame correspondences. Natural object motion can often be compensated through the estimated flow, whereas changes not explained by spatial displacement, such as abnormal color changes, inconsistent texture evolution, and other appearance irregularities, tend to remain in the residual after warping. Building on the camera-motion adjustment above, the inter-frame variation evolves as:

$$\Delta z_t : \quad |f(x_{t+1}) - f(x_t)| \;\longrightarrow\; \left|f(x_{t+1}) - f(\tilde{x}_{t+1}^{\text{cam}})\right| \;\longrightarrow\; \left|f(x_{t+1}) - f(\tilde{x}_{t+1}^{\text{cam+obj}})\right|. \tag{5}$$

Conceptually, this intervention further reduces $\Delta z_t^{\text{object motion}}$ component in Eq. (4), giving

$$\Delta z_t \approx \cdots + \Delta z_t^{\text{gen}}. \tag{6}$$

This simplified decomposition should not be interpreted as isolating a pure generation-related signal. The residual variation may still contain unmodeled non-generative factors, as well as camera- and object-motion effects that are only partially reduced by the approximations above. We refer to the resulting configuration as the *debiased temporal detector*.

The empirical results in Table 1 reveal a clear progressive trend. Starting from the element-wise inter-frame differences, controlling sampling and compression improves the AUC to 0.8827. Reducing global camera motion further raises it to 0.9301, which is the largest single gain among the interventions considered, indicating that camera dynamics are a particularly influential non-generative component. Further accounting for object motion improves the AUC to 0.9485. The smaller gain is consistent with local object motion being harder to approximate reliably than low-dimensional global camera motion. Importantly, the debiased temporal detector approaches the performance of the spatial-only baseline, despite relying only on inter-frame evidence. This progressive improvement supports our decomposition hypothesis: temporal evidence itself is not inherently ineffective; rather, its reliability depends strongly on how much non-generative variation remains entangled in the observed inter-frame signal.

Case Study. To better understand what remains in the inter-frame difference after reducing non-generative variation, we visualize several representative AI-generated examples in Figure 4. In the raw difference, the response is dominated by strong edges and object contours that simply track how the scene moves, obscuring finer irregularities. After debiasing, this explainable structure is largely suppressed and the remaining response concentrates in a few localized regions. For example, the highlighted regions correspond to an abnormal change in eye color in the first case, the disappearance

of fine pizza texture in the second, and an abrupt structural change around the mouse's paw in the third. This gives a direct visual counterpart to our hypothesis: the generation-related irregularities are already present in the raw difference but buried under non-generative motion structure, and debiasing does not create new evidence so much as expose evidence that these sources had obscured.

3.4 Why Temporal Modeling Still Matters

The analyses above show that temporal evidence becomes increasingly reliable as non-generative variation is reduced. Yet Section 2.1 found that spatial cues alone already support strong detection, leaving open whether temporal evidence provides information beyond individual frames. To probe this, we simply average the logits of the spatial-only and debiased temporal detectors. Because this fusion introduces no learnable parameters, its gain cannot be attributed to fitting an additional module. As shown in the last block of Table 1, the two detectors achieve comparable AUCs on GenVideo (0.954 and 0.949), while their fusion reaches 0.977.

These results provide a direct answer to our motivating question: temporal modeling remains valuable for AIGVD. Once non-generative variation is reduced, temporal evidence becomes highly discriminative and contributes beyond what spatial evidence alone recovers, providing distinct rather than redundant information for robust AIGVD. Moreover, temporal evidence proves even more valuable under benchmark shift (Section 4), where temporal evidence even surpasses spatial evidence in AUC while exhibiting different prediction patterns.

4 Cross-Benchmark Evaluation Beyond GenVideo

The preceding analyses show that temporal evidence becomes highly discriminative on GenVideo after debiasing, with per-generator results provided in Appendix D.1. We next evaluate cross-benchmark generalization by testing the GenVideo-trained models on GenVidBench (Ni et al., 2026) without retraining, comparing existing temporal methods, the spatial-only detector, the debiased temporal detector, and their parameter-free fusion. Detailed experimental settings are provided in Appendix C, with additional results in Appendix D. Our code is released at `https://anonymous.4open.science/r/Anonym_Code_1199`

4.1 Experimental Setup

Cross-benchmark setup. To test generalization beyond GenVideo, we evaluate the GenVideo-trained models on GenVidBench (Ni et al., 2026), using MuseV, SVD, CogVideo, and Mora as unseen generators and HD-VG-130M (Wang et al., 2025) as the real source. This setting changes not only the generators and real source, but also the benchmark from which both classes are collected. We report both AP and AUC, and break results down by action category: Active Engagement (AE), Presenting and Displaying (PD), Holding Objects (HO), Observing and Perceiving (OP), Static Postures (SP), and Others (Oth). Videos within each category exhibit similar motion, so comparisons within a category are not confounded by differences in motion magnitude.

Baselines. We perform comparisons with three representative temporal detectors specifically designed for AIGVD (*i.e.,* D3 (Zheng et al., 2025), ReStraV (Internò et al., 2025), OverCoherence (Brokman et al., 2026)), with further details provided in Appendix C.1.

4.2 Results

We evaluate generalization on GenVidBench using models trained only on the GenVideo training set, without retraining. We report AP and AUC per action category in Table 2. The fused detector achieves the best results on both metrics and in every action category.

First, debiased temporal evidence outperforms all three existing temporal methods designed for AIGVD (*i.e.,* D3, ReStraV, and Over-Coherence) in every action category on both metrics. Against ReStraV, the strongest prior method, it reaches 0.877 AP and 0.791 AUC overall, compared with 0.808 and 0.758. This indicates that the limitation is not the amount of generation-related evidence in inter-frame variation, but how much of it is obscured by non-generative variation: reducing that variation recovers evidence that was already present.

Table 2: Results on GenVidBench per action category. Bold marks the best value per column and underline marks the second-best value.

Method	AE		PD		HO		OP		SP		Oth		ALL	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
D3 (ICCV'25)	0.613	0.675	0.603	0.653	0.627	0.696	0.634	0.688	0.620	0.680	0.614	0.658	0.618	0.674
ReStraV (NeurIPS'25)	0.810	0.765	0.800	0.741	0.825	0.773	0.809	0.768	0.816	0.768	0.796	0.730	0.808	0.758
Over-Coherence (WACV'26)	0.667	0.699	0.589	0.624	0.601	0.650	0.647	0.654	0.606	0.626	0.573	0.698	0.605	0.669
Spatial-only	0.900	0.778	0.889	0.751	0.892	0.758	0.906	0.788	0.899	0.777	0.881	0.735	0.896	0.767
Debiased temporal	0.888	0.813	0.860	0.765	0.868	0.780	0.890	0.812	0.884	0.802	0.859	0.756	0.877	0.791
Fusion	**0.930**	**0.857**	**0.917**	**0.826**	**0.918**	**0.829**	**0.932**	**0.861**	**0.926**	**0.849**	**0.909**	**0.813**	**0.924**	**0.842**

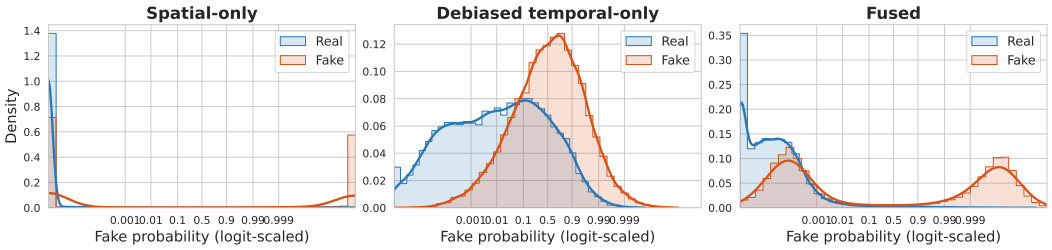


Figure 5: Score distributions of spatial, de-biased temporal, and their fused model on GenVidBench.

Second, comparing the two forms of evidence reveals that they behave differently rather than one consistently dominating the other. Spatial-only achieves higher AP in all six categories (0.896 vs. 0.877 overall), whereas debiased temporal achieves higher AUC in all six (0.791 vs. 0.767 overall). The AUC ordering reverses what we observe on GenVideo, where spatial evidence is slightly ahead: as the decomposition predicts, the components reduced here are tied to how videos are captured and processed, leaving temporal evidence less dependent on any particular benchmark.

Third, their parameter-free fusion substantially improves over either single source, indicating that spatial and temporal evidence capture distinct and complementary information. To better understand this complementarity, we further visualize their prediction distributions in Figure 5. Spatial predictions are highly saturated, with many samples concentrated near the probability extremes. This produces strong confidence for a large fraction of generated videos, but also leaves some samples confidently misranked, explaining its high AP but comparatively lower AUC. Debiased temporal predictions are more graded, with the two classes exhibiting overlapping but systematically shifted distributions, leading to more consistent ranking across the population. Their different error patterns make the two sources complementary, and simple fusion combines the strong separation of spatial evidence with the smoother ordering of temporal evidence.

5 Conclusion

In this paper, we asked whether AI-generated video detection really needs temporal modeling. We find that a simple spatial-only detector outperforms temporal detectors designed for this task, once preprocessing shortcuts are controlled. By taking a closer look at current AIGVD temporal methods, we propose a decomposition hypothesis: they all model the observed inter-frame variation, which mixes generation-related evidence with substantial variation from non-generative sources such as video preprocessing, camera motion, and object motion. Beyond adding noise, these sources can correlate with whether a video is real or AI-generated, so detectors may learn them instead of generation-related evidence. To test this, we first examined how inter-frame variation is represented, and found that the scalar summaries commonly used by existing methods discard much of its structure. We then reduced its non-generative components one source at a time. Each step improved temporal detection, raising it from 0.85 to 0.95 AUC on GenVideo, bringing it level with spatial evidence. The resulting temporal evidence becomes comparable to spatial evidence on GenVideo and even surpasses it in AUC under cross-benchmark evaluation on GenVidBench. Temporal evidence was therefore not weak, only entangled. Temporal modeling thus remains valuable for AIGVD: once inter-frame variation is represented more faithfully, and its non-generative components are properly accounted for, temporal evidence becomes highly discriminative for AIGVD.

AI USE STATEMENT

Generative AI tools were used for language polishing and providing suggestions or edits to software code. All AI-assisted text and code were reviewed and verified by the authors. All research ideas, methodological design, experiments, analyses, and conclusions were developed and validated by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

We have made extensive efforts to ensure the reproducibility of our analyses. Details of the models used in our analyses are provided in Appendix B. The experimental setup, model configurations, training details, and hyperparameter settings are provided in Appendix C. The preprocessing and shortcut-control protocols, as well as the implementations of the spatial and temporal analyses, are described in detail in the corresponding appendix sections. We also provide an anonymous link to our source code and evaluation scripts (`https://anonymous.4open.science/r/Anonym_Code_1199`).

REFERENCES

Omer Ben Hayun, Roy Betser, Meir Yossef Levi, Levi Kassel, and Guy Gilboa. Training-free detection of generated videos via spatial-temporal likelihoods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16299–16310, 2026.

Jonathan Brokman, Oren Rachmil, Omer Hofman, Roy Betser, Amit Giloni, Roman Vainshtein, and Hisashi Kojima. Training-free detection of text-to-video generations via over-coherence. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3993–4003, 2026.

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. *Science China Information Sciences*, 69(6):162103, 2026.

Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Joint pattern recognition symposium*, pp. 236–243. Springer, 2003.

Elan Dubrofsky. Homography estimation. *Diplomová práce. Vancouver: Univerzita Britské Kolumbie*, 5, 2009.

Chuck Easttom. Malicious use of artificial intelligence. In *IEEE 15th Annual Computing and Communication Workshop and Conference*, pp. 00499–00507, 2025.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. Ai-generated video detection via perceptual straightening. *Advances in neural information processing systems*, 38:20672–20705, 2025.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28050–28060, 2025.

Zhengcen Li, Chenyang Jiang, Hang Zhao, Shiyang Zhou, Yunyang Mo, Feng Gao, Fan Yang, Qiben Shan, Shaocong Wu, and Jingyong Su. Preserving forgery artifacts: Ai-generated video detection at native scale. In *International Conference on Learning Representations*, 2026.

Tony Lindeberg. Scale invariant feature transform. *Scholarpedia*, 7(5):10491, 2012.

Qingyuan Liu, Pengyuan Shi, Yun-Yun Tsai, Chengzhi Mao, and Junfeng Yang. Turns out i'm not real: Towards robust detection of ai-generated videos. *arXiv preprint arXiv:2406.09601*, 2024.

Francis Martinson. The marketing-fraud convergence: When legitimate ai tools enable financial crime. *International Journal of Computer Applications*, 187(98):1–5, 2026.

Joren Michels, Lode Jorissen, and Nick Michiels. Dataset biases and shortcut learning in motion-based ai-generated video detection. *arXiv preprint arXiv:2607.00948*, 2026.

Zhenliang Ni, Qiangyu Yan, Mouxiao Huang, Tianning Yuan, Yehui Tang, Hailin Hu, Xinghao Chen, and Yunhe Wang. Genvidbench: A 6-million benchmark for ai-generated video detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15582–15590, 2026.

OpenAI. Sora 2. `https://openai.com/index/sora-2/`, September 2025. Accessed: 2026-08-09.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

Kyoungjun Park, Yifan Yang, Juheon Yi, Shicheng Zheng, Muhammad Muaz, Yifei Shen, Dongqi Han, Caihua Shan, and Lili Qiu. Vidguard-r1: Ai-generated video detection and explanation via reasoning mllms and rl. In *International Conference on Learning Representations*, volume 2026, pp. 108034–108058, 2026.

Pika Art. Pika. `https://pika.art/`, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763, 2021.

Akshay Shetye, Nilakshi Jain, Shwetambari Borade, and Vineet Kumar. Deepfake technologies in financial fraud: Generation, detection, and mitigation strategies. In *Global Conference in Emerging Technology*, pp. 1–6, 2025.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael 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, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

LI Sophia. The social harms of ai-generated fake news: Addressing deepfake and ai political manipulation. *Digital Society & Virtual Governance*, 1(1):72–88, 2025.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419, 2020.

Danial Samadi Vahdati, Tai D Nguyen, Aref Azizpour, and Matthew C Stamm. Beyond deepfake images: Detecting ai-generated videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 4397–4408, 2024.

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

Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. Swap attention in spatiotemporal diffusions for text-to-video generation. *International Journal of Computer Vision*, 133(7):4177–4195, 2025.

Haiquan Wen, Yiwei He, Zhenglin Huang, Tianxiao Li, Zihan Yu, Xingru Huang, Lu Qi, Baoyuan Wu, Xiangtai Li, and Guangliang Cheng. Busterx: Mllm-powered ai-generated video forgery detection and explanation. *arXiv preprint arXiv:2505.12620*, 2025.

Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating misinformation in the era of generative ai models. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 9291–9298, 2023a.

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*, 2023b.

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*, pp. 5288–5296, 2016.

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. *Advances in neural information processing systems*, 38:174683–174730, 2025.

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.

Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pp. 1–20, 2023.

A Related Work

This section briefly reviews the topic related to this work, including AI-Generated Video Detection methods and temporal Modeling for AI-Generated Video Detection.

A.1 AI-Generated Video Detection

With the rapid development of video generation models, AI-generated video detection has received increasing attention. Existing methods can be roughly divided into two categories, *i.e.,* visual forensic detectors and MLLM-based detectors. Visual forensic detectors aim to model low- or mid-level visual cues, such as spatial artifacts, frequency traces, and representation discrepancies. For example, Vahdati et al. (2024) show that AI-generated videos exhibit forensic traces distinct from those of generated images and learn these video-specific traces for detection and source attribution. Qwen2.5-ViT (Li et al., 2026) preserves fine-grained forgery artifacts by processing videos at their native spatial resolution and temporal duration with a Qwen2.5-VL Vision Transformer. Recent MLLM-based approaches leverage high-level semantic and physical reasoning to identify implausible content or generation artifacts. For example, BusterX (Wen et al., 2025) formulates video forgery detection as a visual reasoning task and uses reinforcement learning to produce both authenticity predictions and interpretable rationales. VidGuard-R1 (Park et al., 2026) further employs reinforcement learning to encourage physics-grounded reasoning over temporal consistency, lighting, texture, and physical plausibility. Beyond these largely spatial or semantic cues, videos also provide temporal information across frames, which has motivated a growing line of AIGVD methods specifically designed to model temporal evidence.

A.2 Temporal modeling for AI-Generated Video Detection

Temporal information provides a key distinction between videos and individual images, and has therefore attracted increasing attention in AIGVD. Recent studies have shown that real and AI-generated videos exhibit different temporal patterns, motivating a growing line of methods that explicitly model cross-frame dynamics for detection.

One recent line of work focuses primarily on temporal evidence, showing that real and AI-generated videos can be distinguished from their inter-frame dynamics alone. For example, D3 (Zheng et al., 2025) observes that AI-generated videos tend to exhibit smaller and more stable second-order temporal variations than real videos, and exploits this discrepancy for detection. ReStraV (Internò et al., 2025) identifies differences in inter-frame angular changes, with real videos generally exhibiting smaller angles between consecutive frame representations than AI-generated videos, and uses these angular variations as discriminative cues. Over Coherence (Brokman et al., 2026) finds that AI-generated videos often exhibit abnormally high temporal coherence between adjacent frames, and detects them by measuring similarity across consecutive frame representations.

Another line jointly models spatial and temporal evidence. STALL (Ben Hayun et al., 2026) combines frame-level spatial likelihoods with temporal likelihoods computed from differences between consecutive frame representations. NSG-VD (Zhang et al., 2025) models spatial probability gradients together with temporal density changes under physical conservation principles. Despite incorporating spatial information, their temporal components are still derived from variation across frames.

Despite these different designs, both lines ultimately derive their temporal evidence from inter-frame variation. Recent work (Michels et al., 2026) further shows that such temporal evidence can be strongly affected by preprocessing-induced shortcuts, raising concerns about whether the captured variation faithfully reflects generation-related cues.

B Model Details

B.1 DINOv3 Backbone and Representation Details

Spatial-only detector. We use DINOv3 ViT-L/16 as the backbone, whose feature dimension is $D = 1024$. For each frame, we concatenate the normalized class token from the final transformer

block with the average-pooled patch tokens, yielding a $2D = 2048$ dimensional representation. A linear classifier produces frame-level logits. The pretrained DINOv3 parameters are frozen, while rank-32 LoRA adapters are inserted into the query-key-value projections of all self-attention blocks.

For video-level prediction, we uniformly sample 16 frames from each two-second clip at 8 FPS and apply confidence-based top-K aggregation. Specifically, we rank the frames by their maximum softmax probability, select the top $K_s = 4$ frames, and average their logits to obtain the final video-level prediction. In the frozen-backbone ablation of Appendix D.3, the LoRA adapters are removed while all other settings remain unchanged.

Temporal detector. The temporal detector uses the same DINOv3 ViT-L/16 backbone, which remains fully frozen. For each adjacent frame pair, we first compensate for global camera motion using homography and optionally refine the alignment using RAFT-Small. The aligned frame and target frame are then encoded by DINOv3. We extract patch tokens from transformer blocks $\{12, 18, 23, 24\}$ using one-based indexing, while discarding class and register tokens. After ℓ_2 normalization, we compute the element-wise absolute difference between corresponding patch representations. Mean and max pooling over patches are applied at each layer, and the four layer-wise representations are concatenated into an 8192-dimensional representation for each frame pair.

For video-level prediction, a lightweight confidence head scores all 15 adjacent frame pairs in a 16-frame clip. We select the top $K_t = \lfloor (T-1)/3 \rfloor = 5$ pairs and aggregate their representations using temporal mean and max pooling. The resulting 16,384-dimensional video representation is fed into a two-layer MLP classifier with hidden dimension 512, GELU activation, and dropout rate 0.1.

B.2 Implementation Details for Camera-Motion Analysis

For each pair of adjacent frames x_t and x_{t+1}, we first detect SIFT (Lindeberg, 2012) keypoints and extract their local descriptors in both frames. Feature correspondences are then established between the two frames, from which a homography matrix H_t is estimated using RANSAC (Chum et al., 2003) to suppress mismatched correspondences. The estimated transformation is used to warp the preceding frame x_t toward the coordinate system of x_{t+1}:

$$\tilde{x}_{t+1}^{\mathrm{cam}} = \mathcal{W}(x_t; H_t). \tag{7}$$

This alignment mainly accounts for global geometric changes such as camera translation, rotation, zooming, and perspective variation that can be approximated by a projective transformation. Local motion caused by independently moving objects is not explicitly modeled at this stage.

Both x_{t+1} and $\tilde{x}_{t+1}^{\mathrm{cam}}$ are then passed through the same frozen visual encoder $f(\cdot)$. The validity mask is mapped to the feature resolution, and inter-frame residuals are computed only over valid spatial locations. In this way, the camera-motion-adjusted representation primarily suppresses global motion that can be explained by the estimated geometric transformation while retaining residual variations that cannot be explained by this global alignment. Note that warping introduces invalid regions for which no valid correspondence exists. We construct a validity mask to identify these regions and map it to the spatial resolution of the image encoder outputs. Patch features falling into invalid regions are excluded from the residual computation, so that only valid aligned regions contribute to the inter-frame difference.

A single homography cannot fully account for camera-induced variation in scenes with significant depth and translation, where parallax may remain after alignment. We therefore treat it as an approximate intervention rather than a complete motion-compensation model. Its purpose is to reduce a physically interpretable source of global non-generative variation and examine its effect on temporal detection, while spatially varying object motion is considered separately in the subsequent analysis.

B.3 Implementation Details for Object-Motion Analysis

Object motion is spatially varying and cannot be well approximated by a single low-dimensional transformation. We therefore model it using dense optical flow. Specifically, after the global camera alignment described in Appendix B.2, we estimate the optical flow field F_t from the camera-aligned frame $\tilde{x}_{t+1}^{\mathrm{cam}}$ to the target frame x_{t+1} using a pretrained RAFT (Teed & Deng, 2020) model. The aligned frame is then further warped according to the estimated flow:

$$\tilde{x}_{t+1}^{\mathrm{cam+obj}} = \mathcal{W}_{\mathrm{flow}}\left(\tilde{x}_{t+1}^{\mathrm{cam}}; F_t\right). \tag{8}$$

As a pretrained generic estimator, RAFT is trained on conventional optical-flow benchmarks to model correspondence-based motion across frames, not generation-specific artifacts. Natural motion that admits reliable cross-frame correspondences can therefore be largely explained by the estimated flow. Irregular changes such as sudden appearance or disappearance, abrupt structural deformation, or inconsistent texture evolution often admit no reliable correspondence, so warping by the estimated flow does not align them, and they survive in the residual. This is consistent with Figure 4, where the response after debiasing concentrates on precisely such regions.

Both x_{t+1} and $\tilde{x}_{t+1}^{\mathrm{cam+obj}}$ are then passed through the same frozen visual encoder $f(\cdot)$, and the object-motion-adjusted residual is computed only over the remaining valid patches. Compared with the camera-only adjustment, this step further reduces local inter-frame variation that can be explained by estimated dense motion.

Dense optical flow does not perfectly separate object motion from generation-related temporal irregularities, especially around occlusions, motion boundaries, and regions with inaccurate correspondence. We therefore use it as an approximate intervention rather than a complete object-motion compensation model.

C Experiment Details

C.1 Baselines

We compare against three representative temporal detectors specifically designed for AIGVD:

- **D3.** D3 is a training-free AIGVD method that characterizes second-order temporal variation in frame representations. It is motivated by the observation that AI-generated videos tend to exhibit smaller and more stable second-order changes than real videos.
- **ReStraV.** ReStraV detects AI-generated videos by modeling inter-frame representation variation, particularly changes in feature distances and angular relationships between consecutive frames. These statistics are used to distinguish the temporal dynamics of real and generated videos.
- **Over-Coherence.** Over-Coherence is a training-free detector based on the temporal coherence of consecutive frame representations. It measures cosine similarity across adjacent frames and exploits the tendency of generated videos to exhibit abnormal levels of temporal coherence.

For all baselines, we follow the official implementations and evaluation protocols whenever available, while applying the same preprocessing and shortcut-control setting used throughout our experiments to ensure a fair comparison.

C.2 Implementation Details

All videos are uniformly sampled at 8 FPS for 2 seconds, resulting in 16 frames per video. Each frame is resized such that its shorter side is 224 pixels and then center-cropped to 224×224.

For each pair of adjacent frames, we estimate a homography using SIFT matching and RANSAC, and warp the preceding frame toward the subsequent frame. Invalid regions introduced by warping are identified using a validity mask and filled with the corresponding pixels from the target frame before feature extraction.

We use a frozen DINOv3 ViT-L/16 backbone to extract multi-layer patch features. Feature residuals are computed only over valid patches and aggregated using mean and max pooling. The most informative temporal pairs are selected by a learned confidence-based top-k module and aggregated for video-level classification.

Only the confidence and classification heads are optimized using AdamW with a learning rate of 1×10^{-4}, weight decay of 1×10^{-2}, and a batch size of 8 videos for 10 epochs. Class-balanced sampling is applied in each epoch. All experiments are conducted on a server equipped with eight NVIDIA RTX PRO 6000 GPUs with 96GB memory each.

C.3 Evaluation Metric

We report average precision (AP) and area under the ROC curve (AUC), with generated videos as the positive class. AUC is invariant to the class prior and reflects global separability, whereas AP is dominated by the high-confidence end of the ranking and depends on the ratio of generated to real videos; AP is therefore compared only within a benchmark. On GenVideo, the protocol of Appendix C.5 leaves some generators with only a few dozen clips. Rather than subsampling real videos to match, which would make the result depend on which real clips happen to be drawn, we evaluate each generator against the full set of real videos. AUC is unaffected by the resulting imbalance, while AP is difficult to interpret at this ratio, so we report AUC alone on GenVideo and both metrics on GenVidBench.

C.4 Dataset and protocol

Following D3 (Zheng et al., 2025), we train on GenVideo (Chen et al., 2026) using real videos from Youku-mPLUG (Xu et al., 2023b) and generated videos from Pika. All videos, in both training and evaluation, are processed with the shortcut-controlled protocol in Appendix C.5, unless otherwise stated, and we report AUC throughout. Under the GenVideo cross-generator protocol, we test on the remaining generators, including ModelScope (MSE), MorphStudio (MSO), Show-1, Gen-2, Crafter, LaVie, and Sora, with real videos from MSR-VTT (Xu et al., 2016). Thus, neither the generators nor the real source is seen during training. We further evaluate the same trained models on GenVidBench (Ni et al., 2026), using MuseV, SVD, CogVideo, and Mora as unseen generators and HD-VG-130M (Wang et al., 2025) as the real source. This setting changes not only the generators and real source, but also the benchmark from which both classes are collected. For GenVidBench, we report performance by action category: Active Engagement (AE), Presenting and Displaying (PD), Holding Objects (HO), Observing and Perceiving (OP), Static Postures (SP), and Others (Oth). Videos within each group exhibit similar motion characteristics, reducing the influence of motion-related shortcuts and enabling a fairer comparison across motion regimes.

C.5 Preprocessing Shortcut Control

Recent studies have revealed that existing temporal detectors are strongly affected by data-preprocessing shortcuts introduced during video sampling and compression (Michels et al., 2026). One prominent example is frame duplication. When videos have insufficient duration or frame rate, some evaluation pipelines repeatedly sample or duplicate frames to satisfy a fixed temporal input. Since such cases can be highly imbalanced between real and AI-generated videos, the resulting abnormally small inter-frame variation becomes an easy class-correlated cue for temporal detectors. In addition, preprocessing differences related to video resolution and JPEG/H.264 compression can introduce further class-specific artifacts. To reduce these shortcuts, Michels et al. (2026) adopt a unified preprocessing protocol that eliminates frame duplication and resolution-dependent compression.

To reduce these shortcuts, Michels et al. (2026) adopt a unified preprocessing protocol. Videos are sampled at 8 FPS over a 2-second interval, and videos shorter than 2 seconds or below 8 FPS are discarded; for real video datasets MSR-VTT, low-frame-rate clips are instead replaced by their higher-frame-rate source videos, so that real videos are not disproportionately removed. The sampled frames are saved losslessly in PNG format, resized such that the shorter side is 224, and center-cropped to 224×224. This controlled protocol reduces obvious sampling- and preprocessing-induced shortcuts, providing a cleaner setting for evaluating temporal detectors.

D Additional Results

D.1 Full Results on GenVideo Analysis

We report the complete per-generator results on GenVideo in Table 3, covering all settings analyzed in Section 3. All models are trained on the same GenVideo training set under the shortcut-controlled protocol and evaluated with AUC.

Table 3: Complete per-generator results on the GenVideo benchmark under the shortcut-controlled setting, reported as AUC. *Overall* is computed jointly over all generators and is therefore not the average of the per-generator columns. ReStraV* denotes the original ReStraV without sampling and compression control. The three indented rows apply motion compensation to the element-wise difference either individually or jointly. Bold marks the best value in each column.

Method	Crafter	Gen-2	LaVie	MSE.	MS.	Show-1	Sora	Overall
Existing Temporal Methods								
D3 (ICCV'25)	0.8438	0.9219	0.7868	0.7291	0.7693	0.8415	0.8088	0.7941
ReStraV* (NeurIPS'25)	0.8078	0.9212	0.6572	0.6936	0.6377	0.7971	0.7010	0.7557
ReStraV (NeurIPS'25)	0.8742	0.9106	0.7289	0.7459	0.7119	0.8843	0.7500	0.8046
Over-Coherence (WACV'26)	0.7518	0.8568	0.6924	0.7216	0.7446	0.8043	0.7660	0.7544
Existing Spatio-Temporal Methods								
NSG-VD (NeurIPS'25)	0.7434	0.7962	0.6515	0.6713	0.6741	0.7748	0.651	0.7113
STALL (CVPR'26)	0.7943	0.8676	0.8508	0.7841	0.8169	0.816	0.792	0.8431
Spatial-Only Detectors								
CLIP-B	0.9981	0.9929	0.9261	0.9682	**0.9981**	0.8653	0.9444	0.9518
DINOv2	0.9990	0.9949	0.9304	**0.9947**	0.9974	0.8597	0.9561	0.9532
DINOv3	**0.9996**	0.9953	0.9266	0.9912	0.9959	0.8387	**0.9631**	0.9544
Controlled Progressive Analysis + S&C Control								
L2 Distance	0.8392	0.9295	0.7313	0.7100	0.6901	0.8509	0.7645	0.7957
Element-wise Difference	0.9095	0.9705	0.8165	0.8339	0.8375	0.9102	0.7965	0.8827
+ Homography (camera motion)	0.9506	0.9918	0.8964	0.9195	0.9099	0.9037	0.9017	0.9301
+ Flow (camera & object motion)	0.9422	0.9928	0.8862	0.9201	0.8966	0.9163	0.8919	0.9272
+ Homography then optical flow	0.9632	0.9956	0.9253	0.9496	0.9322	0.9153	0.9227	0.9485
Fusion	0.9985	**0.9994**	**0.9618**	0.9939	0.9844	**0.9349**	0.9553	**0.9770**

First, the progressive trend reported in Section 3 is not an aggregate artifact but holds for individual generators. Moving from the element-wise difference to full motion compensation improves AUC on all seven generators, with gains ranging from 0.5 points on Show-1 to 12.6 points on Sora. The same holds for the sampling and compression control: it improves ReStraV on six of the seven generators (for example, 0.7971 to 0.8843 on Show-1), confirming that the effect is not specific to our analysis pipeline. Note also that optical flow estimates the total apparent motion field, so the flow-only row compensates for camera and object motion together rather than object motion alone.

Second, the two motion compensations support the two-stage design used in Section 3. Applied individually, homography and optical flow reach 0.9301 and 0.9272 overall, so flow alone does not surpass homography despite also compensating object motion. Camera motion is a global effect that a low-dimensional transformation can estimate from the whole frame at once, whereas a dense field estimates it independently at every pixel and accumulates local error, so the additional object-motion compensation it provides is offset by a less accurate global alignment. Applying the two in sequence avoids this trade-off and reaches 0.9485: the global component is removed parametrically, and dense flow is left only with the residual local motion.

Third, the strength of spatial-only detection is stable across both backbones and generators, with one clear exception. The three frame encoders differ by at most 0.26 points overall, and each exceeds 0.92 AUC on six of the seven generators. On Show-1, however, all three fall to 0.8387–0.8653, while every element-wise temporal configuration reaches 0.90 or above. This is the clearest per-generator case in which temporal evidence carries information that spatial cues do not.

Fourth, the fused detector attains the best overall AUC but is not uniformly best per generator. It ranks first on Gen-2, LaVie, Show-1, and overall, and falls slightly behind the strongest spatial-only variant on Crafter, ModelScope, MorphStudio, and Sora. In all four cases, spatial-only is already above 0.96, leaving little headroom and making the averaged logit slightly costly. The largest fusion gains instead appear where neither source is saturated, such as LaVie (0.9304 to 0.9618) and Show-1 (0.9153 to 0.9349), consistent with the saturation pattern visualized in Figure 5.

Table 4: Effect of encoder adaptation on the DINOv3 spatial-only detector under the shortcut-controlled GenVideo setting. AUC is reported for each generator and jointly over all generators.

Setting	Crafter	Gen2	LaVie	ModelScope	MorphStudio	Show-1	Sora	Overall
DINOv3 + LoRA	**0.9996**	0.9953	0.9266	0.9912	**0.9959**	0.8387	**0.9631**	**0.9544**
DINOv3 (Frozen)	0.9963	**0.9992**	**0.9316**	**0.9970**	0.9913	**0.8452**	0.9495	0.9515

D.2 Existing Spatio-Temporal Methods Analysis

Our main analysis focuses on temporal-only detectors so that the contribution of temporal evidence can be isolated from spatial cues. For completeness, we further evaluate NSG-VD and STALL, which jointly exploit spatial and temporal evidence. Importantly, their temporal modeling still falls within our decomposition framework: both derive temporal evidence from observed cross-frame variation. STALL models inter-frame motion naturalness in a pretrained feature space, while NSG-VD estimates temporal derivatives from inter-frame changes as part of its normalized spatiotemporal gradient. Thus, although their final predictions also use spatial information, the temporal evidence they rely on is still formed from the same mixed inter-frame variation considered in Eq. (1).

These methods also clarify the distinction between the two aspects studied in our analysis. Their temporal modeling is not restricted to the scalar summaries examined in Section 3.1, yet richer representations alone do not remove the non-generative variation contained in the underlying inter-frame signal. This is particularly relevant to NSG-VD, whose temporal derivative is approximated directly from inter-frame changes without explicitly separating camera or object motion. Its temporal representation can therefore still contain the non-generative components described by our decomposition. In other words, preserving richer inter-frame structure addresses how temporal variation is represented, whereas our subsequent analysis concerns what that variation contains.

As shown in Table 3, NSG-VD and STALL achieve overall AUCs of 0.711 and 0.843, respectively. Both remain below the spatial-only baselines (0.952–0.954), while STALL also remains below the element-wise temporal detector (0.883). The contrast is particularly clear on Show-1, where spatial-only detection is relatively weak (0.839–0.865) but the element-wise temporal variants exceed 0.90 AUC. NSG-VD and STALL reach only 0.775 and 0.816, respectively. This suggests that combining spatial evidence with a richer temporal representation does not by itself resolve the non-generative variation contained in inter-frame evidence.

Our fusion instead combines spatial evidence with the debiased temporal detector and reaches 0.93 AUC on Show-1 and 0.977 overall. These results are consistent with our central view: the limitation of temporal modeling is not simply whether inter-frame variation is represented by a scalar or a richer form, but whether the observed variation remains entangled with non-generative sources.

D.3 Spatial-Only Detection without Encoder Adaptation

To examine whether the strong spatial-only performance depends on task-specific adaptation of the image encoder, we repeat the DINOv3 spatial-only experiment with the encoder fully frozen, removing LoRA while keeping all other settings unchanged. As shown in Table 4, freezing the encoder has little effect on the overall performance: the AUC only decreases from 0.9544 to 0.9515. The per-generator results also show no consistent advantage from LoRA adaptation; the frozen variant performs better on several generators while remaining comparable on the others. These results further show that strong AIGVD performance can already be achieved from spatial information without adapting the pretrained encoder.