

Article

Benchmarking and Improving Perceptual Straightening for Robust AI-Generated Video Detection

Xenofon Papougiannakis  and Ruben Tous 

Department of Computer Architecture, Universitat Politècnica de Catalunya, 08034 Barcelona, Spain;
xenofon.papougiannakis@estudiantat.upc.edu

* Correspondence: ruben.tous@upc.edu

Abstract

The proliferation of text-to-video generative models—including commercial systems such as Sora, Veo, Runway Gen-3, and Kling—makes automated detection of AI-generated videos an urgent problem in multimedia forensics. We present a study with two interleaved contributions: a cross-family benchmark and two improvements to the detection pipeline. For the benchmark, we evaluate three methodologically distinct detector families on three public datasets (VidProM, DeepAction, and DeCoF_T2V), spanning diverse open-source and commercial generators: RESTRAV, a geometry-supervised detector grounded in the perceptual straightening hypothesis; D3, a training-free detector based on second-order XCLIP temporal statistics; and DeMamba, a supervised Bidirectional Mamba module on frozen XCLIP features. RESTRAV and DeMamba achieve broadly comparable global performance on large datasets (VidProM AUC 0.985/0.989; DeCoF_T2V AUC 0.990/0.992), while D3 remains weak as a stand-alone detector under a fixed detection threshold calibrated on VidProM and transferred unchanged to the other benchmarks (AUC 0.534/0.410/0.454 on VidProM/DeepAction/DeCoF_T2V). Per-generator analysis reveals complementary failure modes: DeMamba recovers several hard commercial generators where RESTRAV struggles, whereas RESTRAV remains competitive in low-data settings at substantially lower computational cost. To bridge these weaknesses, we propose two improvements. First, Rich384 enriches RESTRAV's compact geometric descriptor with DINOv2 temporal embeddings, strengthening ranking quality on large benchmarks and recovering generators that geometry alone misses (AUC 0.995 on VidProM and 0.998 on DeCoF_T2V, with DeepAction AUC decreasing to 0.805). Second, we propose *GurAI*, our transparent logistic late-fusion method that combines Rich384 and DeMamba logits, raising threshold-free AUC to 0.998/0.999 on VidProM and DeCoF_T2V while preserving interpretability. On the small DeepAction cohort, fusion improves selected per-generator fake recalls at the cost of elevated REAL false positives (AUC 0.769; REAL accuracy 0.460). Our analysis suggests that transparent late fusion can exploit complementary detector strengths more effectively than architectural redesign alone when facing generator diversity.



Academic Editors: John Atanbori,
Farhan Riaz and Christos Frantzidis

Received: 12 June 2026

Revised: 17 July 2026

Accepted: 21 July 2026

Published: 23 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: AI-generated video detection; perceptual straightening; DINOv2; DeMamba; late-stage fusion; video forensics; text-to-video generation; benchmarking

1. Introduction

The past three years have produced a step change in the quality and accessibility of synthetic video. Text-to-video diffusion systems—from open-source models such as Text2Video-Zero [1] and OpenSora [2] to commercial state-of-the-art generators such as

OpenAI’s Sora [3], Google DeepMind’s Veo [4], Runway Gen-3 Alpha [5], and Kuaishou’s Kling [6]—can now produce videos of several seconds’ duration with photorealistic motion, consistent lighting, and plausible physical interactions. This progress raises profound concerns about synthetic media’s use in disinformation campaigns, political manipulation, fraud, and the broader erosion of trust in authentic video content. As the perceptual quality of generated videos approaches that of real footage, human viewers increasingly struggle to distinguish the two [7], underscoring the urgency of robust automated detection systems.

A promising approach to this problem was inspired by the *perceptual straightening* hypothesis from computational neuroscience [8], which was established in biological visual systems rather than in self-supervised ViT embeddings. Specifically, natural visual trajectories, when projected into a deep perceptual feature space, become characteristically straighter (lower curvature), whereas synthetically generated trajectories deviate from this expected straightening and remain comparatively curved. We test whether a similar straightening pattern appears in DINOv2 feature trajectories. RESTRAV [9], proposed at NeurIPS 2025, operationalises this insight by extracting a compact 21-dimensional geometric descriptor of inter-frame distances and curvatures from the DINOv2 [10] CLS token trajectory of a video and classifying it with a lightweight MLP. The method achieves strong detection performance on the VidProM benchmark [11], but its performance on *modern* commercial generators remains unclear, and no reproducible head-to-head comparison against methodologically distinct detector families has been published.

This paper addresses two interleaved problems: First, prior work, including RESTRAV [9], reports strong results by aggregating numbers from VidProM, GenVidBench, and task-specific DeMamba protocols but does not provide a unified cross-family evaluation on identical splits with a single reporting convention—which precludes direct comparison of geometry-based, training-free, and supervised temporal detectors. Second, the 21-D geometric compression that defines RESTRAV discards signals that are critical for distinguishing modern generators such as RunwayML, Veo, Kling, and Sora from real videos, as evidenced by our ablation studies (Section 5.1). We address both problems jointly.

Our contributions are as follows:

- C1. A cross-family benchmark.** We evaluate three methodologically distinct detector families—RESTRAV (geometry-supervised, DINOv2), D3 [12] (training-free, second-order XCLIP statistics, ICCV 2025), and DeMamba [13] (supervised Bidirectional Mamba on XCLIP)—on identical balanced splits of three public datasets (VidProM, DeepAction [7], and DeCoF_T2V [14]) under a unified REAL-positive reporting convention.
- C2. Rich384 feature enrichment.** We propose a 405-D feature vector that concatenates RESTRAV’s 21-D geometric descriptor with the full 384-D DINOv2 temporal mean, motivated by an ablation showing that these two representations capture complementary discriminative signals. On VidProM and DeCoF_T2V, Rich384 raises RESTRAV’s AUC to 0.995 and 0.998, respectively; on DeepAction, global AUC decreases from 0.856 to 0.805 but RunwayML fake recall improves from 0% to 61.5% (still modest in absolute terms).
- C3. GurAI transparent late fusion.** We introduce *GurAI*, a transparent logistic stacking of Rich384 probabilities and DeMamba logits, achieving an AUC of 0.998 on VidProM and 0.999 on DeCoF_T2V. On DeepAction, *GurAI* raises RunwayML and Veo fake recall to 0.889 and 1.000 at the fusion threshold, where the marginal towers are suboptimal (0.667 each for RunwayML; 0.750 each for Veo), but REAL accuracy falls to 0.460 and global AUC to 0.769 (Section 6.2).

- C4. Complementarity analysis.** Per-generator and cross-family analysis of where each detector family succeeds and fails demonstrates that late fusion is more effective than architectural redesign for handling the diversity of modern T2V generators.

2. Related Work

2.1. Text-to-Video Generation Architectures

Modern text-to-video generators span several architectural families, each leaving distinct temporal fingerprints that complicate universal detection. *Latent diffusion U-Net* pipelines such as Text2Video-Zero [1] operate in a compressed latent space with spatio-temporal attention blocks, producing characteristic artefacts in the inter-frame distance distribution. *Diffusion Transformer* (DiT) models [15] replace the U-Net backbone with a Vision Transformer operating over spatial and temporal tokens. Representative systems include OpenSora [2], CogVideo [16], Sora [3], and Veo [4]. *Autoregressive* models such as VideoPoet [17] tokenise video frames with a MAGVIT-v2 codebook and generate tokens with a decoder-only Transformer, producing a qualitatively different motion profile. *Commercial proprietary* systems, including Runway Gen-3 Alpha [5], Kling [6], and Pika [18], use diffusion-based architectures with large-scale proprietary training, yielding videos with exceptionally high visual fidelity and natural motion. The diversity of these architectures implies that a robust detector must generalise across generators whose temporal statistics, compression artefacts, and motion patterns may differ substantially.

2.2. AI-Generated Video Detection

Early detection methods for synthetic media focused on *frame-level forensics*: frequency-domain anomalies [19] or face-specific manipulation detectors (e.g., FaceForensics++ [20]). These methods target face-swap deepfakes and do not generalise to fully generated text-to-video content, where the entire frame is synthetic.

Video-level supervised detectors learn temporal representations from labelled real/fake clips. DeMamba [13] pairs a frozen XCLIP-B/16 encoder [21] with a Bidirectional Detail-Mamba block, following the selective state-space design of the public DeMamba implementation, that captures long-range temporal dependencies in linear time, followed by a linear classification head. Trained on the million-scale GenVideo benchmark, DeMamba achieves strong cross-generator generalisation. Its main practical limitation is the CUDA-dependent mamba-ssm build and the GPU requirement for training.

Training-free detectors avoid labelled data entirely. D3 [12] (ICCV 2025) observes that the *second-order temporal differences* of XCLIP feature sequences (that is, the differences of consecutive first-order inter-frame distances) separate real from synthetic videos without any learned classifier. D3 reports a +10.4 mAP improvement over the prior training-free state of the art on the GenVideo benchmark. Its limitation is that a single threshold calibrated on one dataset may not transfer well to datasets with different generator distributions.

Geometry-supervised detectors, exemplified by RESTRAV [9], exploit the *perceptual straightening* phenomenon [8], i.e., the observation from computational neuroscience that real natural videos trace comparatively straight (low-curvature) trajectories in the feature space of visual systems while generated videos tend to trace more curved ones. RESTRAV computes a 21-D descriptor of inter-frame distances and curvatures from DINOv2 CLS token trajectories and trains a small MLP on the resulting features. This detector is lightweight (CPU-only inference) and sample-efficient and achieves competitive AUC on VidProM, but the aggressive geometric compression discards signals that are critical for modern commercial generators.

To our knowledge, no prior work has compared all three families on identical balanced splits of multiple datasets under a consistent positive-class reporting convention, which is the direct gap our benchmark fills.

2.3. Perceptual Straightening

The perceptual straightening hypothesis [8] holds that the visual cortex encodes natural video sequences as trajectories in a latent space where motion manifolds are approximately linear while arbitrary image sequences are curved. RESTRAV and our work do not claim to validate this neuroscientific hypothesis in DINOv2 space; rather, we reuse the *engineering hypothesis* that real and synthetic clips may trace different geometric paths in a vision foundation model (VFM) embedding space. Operationalising this with DINOv2 [10], a self-supervised ViT trained on 142M curated images with a combination of DINO and masked image modelling objectives, provides features that are both semantically rich and sensitive to subtle motion patterns, properties that are central to the success of RESTRAV.

2.4. Vision Foundation Model Feature Compression

Recent work on generative and forensic pipelines has increasingly treated VFM embeddings as an intermediate representation rather than operating on raw pixels. The ICML 2026 work *RePack then Refine* [22] compresses high-dimensional VFM features to reduce redundancy before downstream video synthesis, while our Rich384 variant concatenates a compact 21-D geometry descriptor with the full 384-D DINOv2 temporal mean. Both lines of work confront the same tension: VFMs are semantically rich but high-dimensional, whereas aggressive compression is efficient yet may discard generator-specific cues. We view Rich384 as a forensic-side point on this trade-off curve, enriching geometry rather than compressing the VFM representation away.

2.5. Classifier Fusion for Forensics

Combining the outputs of multiple classifiers through late-stage probability fusion is a well-established technique in ensemble learning. Logistic stacking over classifier logits (as opposed to probability averaging) handles score-scale mismatches between heterogeneous systems and has been shown to outperform simple averaging when the component classifiers carry complementary errors [19,23]. Our *GurAI* system applies this principle to video forensics, fusing a geometry-aware DINOv2 tower with a Mamba-based XCLIP tower via a single logistic layer trained on manifest-aligned logits.

3. Methods

3.1. Datasets and Evaluation Protocol

We evaluate on three publicly available benchmarks that collectively span 22 AI video generators and over 90,000 videos.

3.1.1. VidProM [11]

A large-scale video provenance dataset with 9859 real clips (DVSC2023 footage) and approximately 70,000 fake clips generated by seven text-to-video models: ModelScope, Pika, OpenSora, CogVideo, VideoCrafter2, Text2Video-Zero, and Stable-T2V. After class balancing ($\text{TARGET_PER_CLASS} = 9859$), the working dataset comprises 19,718 samples (9859 real and 9859 fake) with a 50/50 stratified train/test split.

3.1.2. DeepAction [7]

A smaller benchmark designed to test detectors against *modern, high-quality* generators. It contains 100 real clips from Pexels and 2500 fake clips generated by six state-of-the-art

systems: BDAAnimateDiffLightning, CogVideoX-5B, RunwayML Gen-3 Alpha, StableDiffusion Video, Veo, and VideoPoet. After class balancing ($\text{TARGET_PER_CLASS} = 100$), the working dataset comprises 200 samples (100 real and 100 fake), split 50/50 into training and test sets of 100 clips each. The small test fold ($n \approx 100$ per class) means that individual predictions have a disproportionate impact on per-generator metrics; results should be interpreted accordingly.

3.1.3. DeCoF_T2V [14]

A text-to-video subset of the DeCoF benchmark containing 4198 real clips and 4206 fake clips from nine generators: ModelScope, T2V, ZeroScope, Show-1, GEN2, Kling, Pika, Sora, and Veo. After class balancing ($\text{TARGET_PER_CLASS} = 4198$), the working dataset comprises 8396 samples (4198 real and 4198 fake) with a 50/50 stratified train/test split.

Table 1 summarises the three datasets.

Table 1. Summary of the three evaluation benchmarks. “Balanced set” refers to the class-balanced working subset used for training and testing.

	VidProM	DeepAction	DeCoF_T2V
Real videos	9859	100	4198
Fake videos	~70,000	2500	4206
Generators	7	6	9
Balanced set	19,718	200	8396
Test size	9859	100	4198
SOTA generators	OpenSora	RunwayML, Veo	Kling, Sora, Veo

Evaluation protocol.

For all supervised detectors, we use 50/50 stratified train/test splits. For D3 (training-free), we use the entire balanced set for evaluation, with a detection threshold calibrated on a held-out VidProM validation slice (1000 real + 1000 fake) and applied unchanged to all datasets. This reports cross-dataset *threshold transfer*, the deployment setting in which a training-free detector must operate without per-benchmark relabelling. Because ROC-AUC is threshold-free, it is unaffected by this choice. We additionally report D3 with a *per-dataset* threshold chosen to maximise F1 on each balanced evaluation set as an operating point upper bound (reported as a footnote to the global benchmark table in Section 5.2). We report global metrics with REAL as the positive class. Precision, recall, F1, average precision (AP), and ROC-AUC are computed with $\text{pos_label} = \text{REAL}$ and scores oriented as $\mathbb{P}(\text{REAL})$. Per-generator columns report *fake recall*, i.e., the fraction of test clips from a given generator that are correctly classified as fake. This convention is consistent with how RESTRAV’s training code reports metrics and allows direct comparison between all detectors.

DeepAction sample size caveat.

The DeepAction benchmark contains only 100 real and 100 fake clips with a 50/50 train/test split (50 clips per class for training). Per-generator test subsets range from 7 to 17 clips (see the per-generator analysis in Section 5.2), so a single misclassification can shift recall by more than 10 percentage points. We therefore interpret DeepAction as a *diagnostic* low-data regime useful for exposing per-generator failure modes but too small to support precise global accuracy estimates or deployment-ready operating points. The REAL-side trade-offs of Rich384 and GurAI on this cohort are analysed in Section 6.2.

3.2. RESTRAV: Baseline Geometry-Supervised Detector

RESTRAV [9] detects AI-generated video by exploiting the perceptual straightening phenomenon. We follow its three-stage pipeline:

1. **Feature extraction.** Twenty-four frames are uniformly sampled over an approximately two-second clip, resized to 224×224 , and passed through a frozen DINOv2 ViT-S/14 backbone [10] to obtain per-frame 384-D CLS token embeddings $\{z_1, z_2, \dots, z_{24}\} \subset \mathbb{R}^{384}$. We keep 24 frames as a fixed budget to match the original RESTRAV implementation and to maintain comparable temporal coverage across benchmarks with different native frame rates; increasing this budget raises memory and preprocessing cost nearly linearly without changing the downstream classifier architecture.
2. **Geometric descriptor.** From the temporal embedding sequence, two signals are computed: inter-frame distances $d_t = \|z_{t+1} - z_t\|$ and curvature angles $\kappa_t = \arccos\left(\frac{(z_{t+1} - z_t) \cdot (z_t - z_{t-1})}{\|z_{t+1} - z_t\| \cdot \|z_t - z_{t-1}\|}\right)$. For each sequence, four summary statistics are computed (mean, min., max., variance). These eight pooled scalars are concatenated with the first seven leading distances and the first six leading curvatures, yielding a **21-dimensional feature vector** per video. This construction follows the original RESTRAV design [9]: with 24 uniformly sampled frames, there are 23 inter-frame distances and 22 curvature angles; retaining the leading seven distances and six curvatures captures the earliest temporal dynamics together with four pooled statistics per signal type (mean, min., max., variance), totalling $7 + 6 + 8 = 21$ dimensions.

$$\mathbf{x}_{\text{geo}} = [d_1, \dots, d_7, \kappa_1, \dots, \kappa_6, \mu_d, \min_d, \max_d, \sigma_d^2, \mu_\kappa, \min_\kappa, \max_\kappa, \sigma_\kappa^2] \in \mathbb{R}^{21}. \quad (1)$$

3. **Classification.** The 21-D vector is z-score normalised using training set statistics, then fed to an MLP ($21 \rightarrow 64 \rightarrow 32 \rightarrow 1$ with ReLU activations and sigmoid output). Training uses BCEWithLogitsLoss and Adam (lr = 10^{-3}) for 10 epochs. The decision threshold τ^* is chosen on the training set by grid search to maximise F1.

3.3. Competing Detector Families

3.3.1. D3: Training-Free Second-Order Detector

D3 [12] requires no labelled training data. Given a video, it samples frames at 8 fps, passes them through a frozen XCLIP-B/16 encoder [21] to obtain feature vectors $\{f^t\}$, and computes first-order differences $\Delta f^t = \|f^{t+1} - f^t\|_2$ and then second-order differences $\Delta^2 f^t = \Delta f^{t+1} - \Delta f^t$. The detection score is the standard deviation of the second-order signal $\text{std}(\Delta^2 f)$; lower values indicate synthetic content. No model weights are learned on target data; the same XCLIP checkpoint is used everywhere.

3.3.2. DeMamba: Supervised Bidirectional Mamba Detector

DeMamba [13] pairs a frozen XCLIP-B/16 encoder with a learnable Bidirectional Detail-Mamba block, following the selective state-space design of the public DeMamba implementation. The Mamba block scans the spatio-temporal patch token sequence twice (forward and backward) and computes selective state updates that encode both local fine detail and long-range temporal context. Pooled tokens feed a linear head that produces a fake/real logit. We train DeMamba end-to-end on each dataset's 50/50 balanced split with the same BCEWithLogitsLoss and F1-maximising threshold protocol as RESTRAV, keeping the XCLIP backbone frozen.

3.4. Rich384: Feature Enrichment

We introduce Rich384 as the first proposed improvement to the RESTRAV pipeline. An ablation study (Section 5.1) shows that the 21-D geometric descriptor and the temporal mean of DINOv2 CLS embeddings capture complementary information across generators. The geometry descriptor ranks strongly on older models such as ModelScope and ZeroScope

(AUC ≈ 0.985) but performs poorly on Kling (14.3%) and RunwayML (0%). The 384-D DINOv2 temporal mean alone recovers Kling (85.7%) and partially recovers RunwayML (30.8%) yet loses ground on the easier generators. Neither representation is sufficient on its own.

Rich384 concatenates both signals before classification. We join the 21-D geometric vector $\mathbf{x}_{\text{geo}}$ (Equation (1)) with the temporal mean of the 24 DINOv2 CLS tokens:

$$\mathbf{x}_{\text{rich}} = \left[\mathbf{x}_{\text{geo}}, \frac{1}{24} \sum_{t=1}^{24} \mathbf{z}_t \right] \in \mathbb{R}^{405}. \quad (2)$$

Although this temporal mean discards frame order, it preserves the generator-specific scene and appearance statistics accumulated across the clip. In our setting, these global VFM cues complement the explicitly dynamic distance/curvature descriptors, which already encode first-order temporal structure. The resulting 405-D vector is z-score normalised on training statistics and fed to the same MLP architecture ($405 \rightarrow 64 \rightarrow 32 \rightarrow 1$) used by baseline RESTRAV, with all other training choices unchanged (class-balanced batches, frozen DINOv2 backbone, F1-maximising threshold on the training fold).

3.5. GurAI: Transparent Logistic Late Fusion

We introduce *GurAI* as the second proposed improvement. Cross-family benchmarking shows that DeMamba and Rich384 succeed on different generators: DeMamba recovers several hard commercial models where RESTRAV struggles, while Rich384 remains competitive at substantially lower computational cost in low-data settings. Rather than redesigning either tower, *GurAI* combines their outputs through a shallow logistic layer while keeping both scorers frozen at inference.

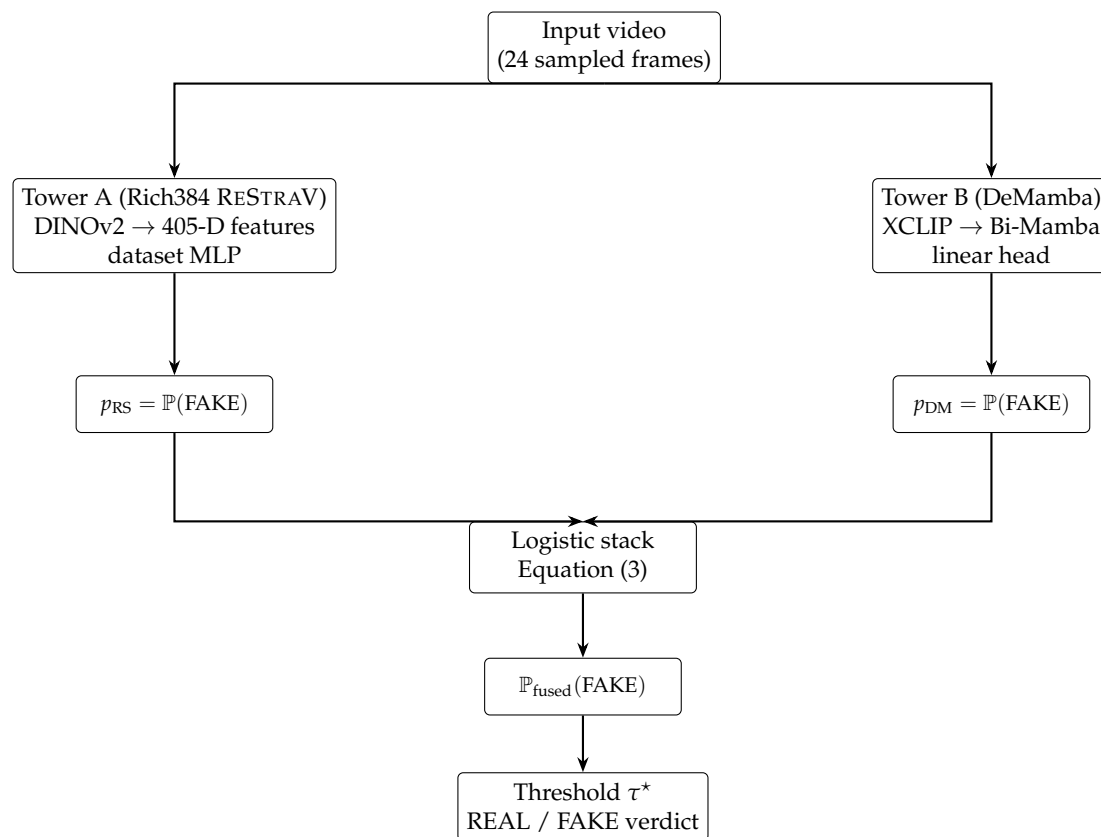
Each clip is scored independently by two towers: Tower A is the dataset-specific Rich384 head described above; Tower B is the DeMamba checkpoint trained on the same benchmark split. Both towers output fake probabilities, denoted by p_{RS} and p_{DM} . We map each probability to a logit $\ell(p) = \log \frac{p}{1-p}$ (clipped for numerical stability) and fit a single fusion rule on the aligned training intersection:

$$z = b + w_{\text{DM}} \ell(p_{\text{DM}}) + w_{\text{RS}} \ell(p_{\text{RS}}), \quad \mathbb{P}_{\text{fused}}(\text{FAKE}) = \sigma(z). \quad (3)$$

Coefficients $\{b, w_{\text{DM}}, w_{\text{RS}}\}$ are estimated by logistic regression on the aligned training intersection. Both towers are trained on the same split before fusion coefficients are fit from their *in-sample* training outputs. This standard stacking procedure can introduce optimistic calibration on very small training folds (50 clips per class on DeepAction) but does not use test set labels. All reported test metrics are computed exclusively on held-out clips. The operating threshold on $\mathbb{P}_{\text{fused}}(\text{REAL}) = 1 - \mathbb{P}_{\text{fused}}(\text{FAKE})$ is chosen on the same merged training table to maximise REAL-positive F1, matching the convention used by both towers. Figure 1 summarises the end-to-end workflow.

Fusion metrics are reported on the *aligned test cohort*: clips for which both towers emitted scores after their respective preprocessing (VidProM $n_{\text{test}} = 9810$; DeepAction $n_{\text{test}} = 100$; DeCoF_T2V $n_{\text{test}} = 2962$). These counts differ slightly from full split sizes when one pipeline drops a clip during manifest construction. Because Equation (3) exposes scalar weights w_{DM} and w_{RS} together with a bias b , the fusion layer remains interpretable; the learned weights indicate how much each tower's evidence is trusted on each training distribution. By “transparent”, we therefore mean parameter-level interpretability of a three-parameter fusion rule rather than full explanation of internal decisions inside each deep tower. Typical values are VidProM $w_{\text{DM}} \approx 0.50$, $w_{\text{RS}} \approx 0.35$; DeCoF_T2V weights

remain comparable; DeepAction $w_{DM} \approx 1.17$, $w_{RS} \approx 0.19$, reflecting that DeMamba logits dominate on the 100-sample training fold.



Stack coefficients fit on aligned training clips. Both towers remain frozen at inference.

Figure 1. GurAI detection workflow. Each clip is scored by a frozen Rich384 RESTRAV tower and a frozen DeMamba tower. A logistic layer fuses the two fake probabilities before a single thresholded decision, where τ^* denotes the F1-optimal decision threshold selected on the training split.

4. Experimental Setup

All experiments were run under fixed random seeds (seed 42) on the balanced train/test splits described in Section 3.1. RESTRAV and Rich384 run on CPU only (Apple M-series, 16 GB RAM); feature extraction takes approximately 8 ms per video. DeMamba requires a GPU; we used Google Colab Pro+ instances with NVIDIA G4 and T4 accelerators. D3 is CPU-only and needs no learned weights on the target benchmarks.

We built on three public codebases with only non-functional patches. RESTRAV follows the community implementation of Internò et al. [9]: <https://github.com/ChristianInternò/ReStraV> (accessed on 12 June 2026). DeMamba comes from the official repository of Chen et al. [13]: <https://github.com/chenhaoxing/DeMamba> (accessed on 12 June 2026). D3 comes from the official repository of Zheng et al. [12]: <https://github.com/Zig-HS/D3> (accessed on 12 June 2026). Across these projects, we adjusted CUDA/CPU device flags, applied an ffmpeg pipe fix in D3's video reader, and pinned library versions; the detection logic itself is unchanged. For reproducibility, we used the exact backbones facebook/dinov2-small (ViT-S/14) for RESTRAV and Rich384 and microsoft/xclip-base-patch16 for D3/DeMamba, both frozen. Our local patches were limited to (i) device selection guards for CPU/GPU execution, (ii) ffmpeg reader compatibility in D3 manifest loading, and (iii) deterministic seed propagation in dataloaders and split scripts.

Supervised detectors share a common training protocol. RESTRAV and Rich384 use an MLP ($21 \rightarrow 64 \rightarrow 32 \rightarrow 1$ or $405 \rightarrow 64 \rightarrow 32 \rightarrow 1$), ReLU activations, sigmoid output, BCEWithLogitsLoss, Adam (learning rate 10^{-3}), 10 epochs, batch size 64, z-score normalisation with training set statistics, and a decision threshold τ^* chosen on the training fold to maximise F1. DeMamba follows the same BCEWithLogitsLoss and F1-maximising threshold protocol with batch size 16 and the standard hyperparameters from the official repository. Specifically, we used AdamW with learning rate 10^{-4} , weight decay 10^{-2} , and 10 training epochs on each balanced split while keeping the XCLIP encoder frozen. For *GurAI*, we fitted logistic regression with the default L2 penalty ($C = 1.0$, scikit-learn) on aligned training logits and chose the operating threshold to maximise REAL-positive F1.

Computational profile.

Our comparison was intentionally pragmatic rather than hardware-optimal: RESTRAV and Rich384 run on CPU with lightweight MLP heads; D3 is also training-free/CPU-friendly, while DeMamba requires GPU execution due to its Bidirectional Mamba block. We therefore report complementary strengths in Section 6: RESTRAV and Rich384 offer lower deployment cost, whereas DeMamba contributes robustness on hard commercial generators. Fine-grained wall-clock benchmarking depends strongly on I/O and frame decoding pipelines and is left for future work.

To support reproducibility, we saved balanced CSV manifests for every experiment, together with `predictions_test.csv` files, so that all metrics can be recomputed without rerunning inference. RESTRAV metrics are reported from the REAL-positive-oriented probability column $\mathbb{P}(\text{REAL})$; DeMamba metrics are recomputed from saved probabilities with `pos_label = 0` (REAL) and $1 - \hat{p}_{\text{fake}}$ as the ranking score.

5. Results

5.1. Geometry vs. DINO-Only: Complementarity Ablation

Before presenting the full benchmark, we motivate Rich384 with a focused ablation comparing the 21-D geometry descriptor against a 384-D DINOv2-only feature (temporal mean of CLS tokens). Table 2 summarises the global results; Figure 2 shows the ROC curves.

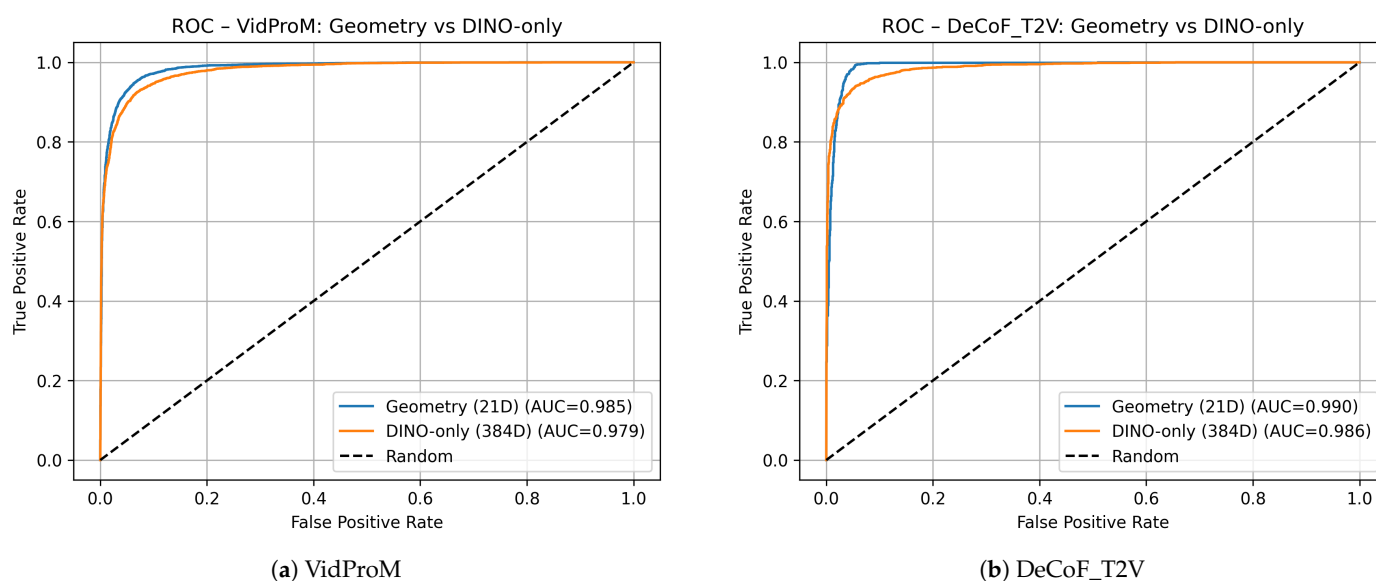


Figure 2. ROC curves: geometry (21-D) vs. DINO-only (384-D) features. While geometry leads globally, DINO-only substantially recovers hard generators on DeCoF_T2V (e.g., Kling, Pika), confirming complementarity.

Table 2. Geometry (21-D) vs. DINO-only (384-D) feature representation. The geometry descriptor leads globally on all three datasets, yet DINO-only recovers hard generators that the geometry descriptor misses (Kling: 85.7% vs. 14.3%; RunwayML: 30.8% vs. 0% on DeepAction), motivating concatenation.

Dataset	Features	Acc	P	R	F1	AUC
VidProM	Geometry (21-D)	0.941	0.928	0.957	0.942	0.985
	DINO-only (384-D)	0.928	0.926	0.930	0.928	0.979
DeepAction	Geometry (21-D)	0.790	0.710	0.980	0.824	0.856
	DINO-only (384-D)	0.700	0.667	0.800	0.727	0.784
DeCoF_T2V	Geometry (21-D)	0.967	0.952	0.983	0.967	0.990
	DINO-only (384-D)	0.939	0.927	0.953	0.940	0.986

5.2. Three-Family Benchmark

Table 3 reports global REAL-positive metrics for all three detector families across the three datasets.

Table 3. Global benchmark results under unified REAL-positive convention. **RESTRAV** and **DeMamba**: metrics use REAL as positive class. **D3**: AP/AUC follow the official fake-positive convention. D3 binary columns use a VidProM validation threshold transferred unchanged. Per-dataset F1-maximising thresholds yield similar accuracy (≈ 0.50) on DeepAction and DeCoF_T2V because ranking AUC remains near chance (0.41/0.45), confirming that poor transfer is not solely an artefact of a single fixed τ .

Dataset	Detector	AP	AUC	Acc	P	R	F1
VidProM	RESTRAV (DINOv2 + 21-D)	0.984	0.985	0.941	0.928	0.957	0.942
	DeMamba (XCLIP + Bi-Mamba)	0.989	0.989	0.950	0.955	0.944	0.949
	D3 (XCLIP + 2nd-order, TF)	0.602	0.534	0.500	0.500	1.000	0.667
DeepAction [†]	RESTRAV (DINOv2 + 21-D)	0.739	0.829	0.800	0.750	0.900	0.818
	DeMamba (XCLIP + Bi-Mamba)	0.751	0.772	0.690	0.880	0.440	0.587
	D3 (XCLIP + 2nd-order, TF)	0.481	0.410	0.500	0.500	1.000	0.667
DeCoF_T2V	RESTRAV (DINOv2 + 21-D)	0.988	0.990	0.967	0.952	0.983	0.967
	DeMamba (XCLIP + Bi-Mamba)	0.992	0.992	0.959	0.952	0.966	0.959
	D3 (XCLIP + 2nd-order, TF)	0.514	0.454	0.499	0.499	0.997	0.666

TF = training-free; [†] matched split: RESTRAV retrained on DeMamba partition.

On the two large benchmarks, VidProM and DeCoF_T2V, RESTRAV and DeMamba operate in the same performance band: REAL-positive AUC differs by ≤ 0.004 on VidProM and ≤ 0.002 on DeCoF_T2V, well within the noise of a single stratified split. D3 scores near chance globally (AUC ≈ 0.53 on VidProM, 0.41 on DeepAction, and 0.45 on DeCoF_T2V), and its fixed transferred threshold produces trivial “label everything fake” behaviour at the operating point. Per-dataset threshold tuning does not recover informative ranking on the two smaller benchmarks because AUC itself remains below 0.5.

On DeepAction, the pattern diverges: the benchmark is smaller and comprises predominantly modern commercial generators. On the matched split, RESTRAV leads on global AUC (0.829 vs. 0.772) and F1 (0.818 vs. 0.587), reflecting its sample efficiency. DeMamba sets an aggressive $\tau^* = 0.14$ that achieves high REAL precision (0.880) but low REAL recall (0.440). D3 collapses below chance (AUC = 0.410) because the XCLIP second-order statistics on DeepAction sit in a different score range than VidProM, where the threshold was calibrated. The score distribution shift across datasets is characteristic of training-free detectors that rely on a fixed operating point. Unsupervised per-dataset threshold adjustment can improve binary metrics but does not change the threshold-free AUC values reported above.

The global tables hide complementary per-generator behaviour, which Tables 4–6 make explicit. On VidProM, RESTRV achieves near-perfect fake recall on ModelScope, VideoCrafter2, and Text2Video-Zero but struggles with OpenSora (65.3%). DeMamba lifts OpenSora to 95.5% and Pika from 86.8% to 99.6%. On DeepAction, RESTRV almost entirely misses RunwayML (22.2%) and Veo (12.5%), whereas DeMamba recovers RunwayML to 88.9% and Veo to 100%. On DeCoF_T2V, RESTRV dominates bulk generators (ModelScope, T2V, ZeroScope: 100%) but fails on Kling (12.2%), Sora (17.2%), and Pika (46.7%). DeMamba recovers Kling (85.1%), Sora (58.3%), and Pika (93.8%).

Table 4. Per-generator fake recall: RESTRV vs. DeMamba on VidProM.

Generator	RESTRV	DeMamba
ModelScope	1.000	0.946
VideoCrafter2	1.000	0.980
Text2Video-Zero	0.999	0.931
Stable-T2V	0.968	0.996
CogVideo	0.978	0.885
Pika	0.868	0.996
OpenSora	0.653	0.955

Table 5. Per-generator fake recall: RESTRV vs. DeMamba on DeepAction (matched split, n = test fakes per generator).

Generator	n	RESTRV	DeMamba
BD-AnimateDiff	9	0.889	1.000
CogVideoX-5B	8	1.000	1.000
RunwayML	9	0.222	0.889
StableDiff. Vid.	8	1.000	1.000
Veo	8	0.125	1.000
VideoPoet	8	1.000	0.750

Table 6. Per-generator fake recall: RESTRV vs. DeMamba on DeCoF_T2V.

Generator	RESTRV	DeMamba
ModelScope	1.000	0.969
T2V	1.000	0.988
ZeroScope	1.000	0.921
Show-1	0.998	0.961
GEN2	0.849	0.980
Pika	0.467	0.938
Kling	0.122	0.851
Sora	0.172	0.583
Veo	0.400	0.429

Because DeepAction per-generator subsets are small ($n = 8$ – 9), we report 95% exact binomial (Clopper–Pearson) confidence intervals for all DeepAction per-generator recalls in Table 7.

Table 7. DeepAction per-generator fake recall with 95% exact binomial confidence intervals (Clopper–Pearson).

Generator	<i>n</i>	RESTRAV	DeMamba	Rich384	GurAI
BDAAnimateDiffLightning	9	0.889 [0.518, 0.997]	1.000 [0.664, 1.000]	0.889 [0.518, 0.997]	1.000 [0.664, 1.000]
CogVideoX-5B	8	1.000 [0.631, 1.000]	1.000 [0.631, 1.000]	0.875 [0.473, 0.997]	1.000 [0.631, 1.000]
RunwayML	9	0.222 [0.028, 0.600]	0.889 [0.518, 0.997]	0.667 [0.299, 0.925]	0.889 [0.518, 0.997]
StableDiffusion Video	8	1.000 [0.631, 1.000]	1.000 [0.631, 1.000]	0.875 [0.473, 0.997]	1.000 [0.631, 1.000]
Veo	8	0.125 [0.003, 0.527]	1.000 [0.631, 1.000]	0.750 [0.349, 0.968]	1.000 [0.631, 1.000]
VideoPoet	8	1.000 [0.631, 1.000]	0.750 [0.349, 0.968]	0.875 [0.473, 0.997]	0.750 [0.349, 0.968]

5.3. Rich384 Feature Enrichment Results

Table 8 reports global metrics for the baseline and the final Rich384 configuration. Figure 3 additionally includes Rich128 (149-D) as an intermediate concatenation variant (21-D geometry plus the first 128 dimensions of the DINOv2 temporal mean).

Table 8. Rich384 (405-D) vs. RESTRAV baseline (21-D). Rich384 raises AUC on both large datasets and recovers hard generators. The AUC drop on DeepAction reflects a threshold trade-off (Section 6.2).

Dataset	Features	Acc	P	R	F1	AUC
VidProM	Baseline (21-D)	0.941	0.928	0.957	0.942	0.985
	Rich384 (405-D)	0.965	0.972	0.957	0.964	0.995
DeepAction	Baseline (21-D)	0.790	0.710	0.980	0.824	0.856
	Rich384 (405-D)	0.760	0.810	0.680	0.739	0.805
DeCoF_T2V	Baseline (21-D)	0.967	0.952	0.983	0.967	0.990
	Rich384 (405-D)	0.977	0.969	0.986	0.977	0.998

Bold indicates the better value between the two feature configurations for each dataset and metric.

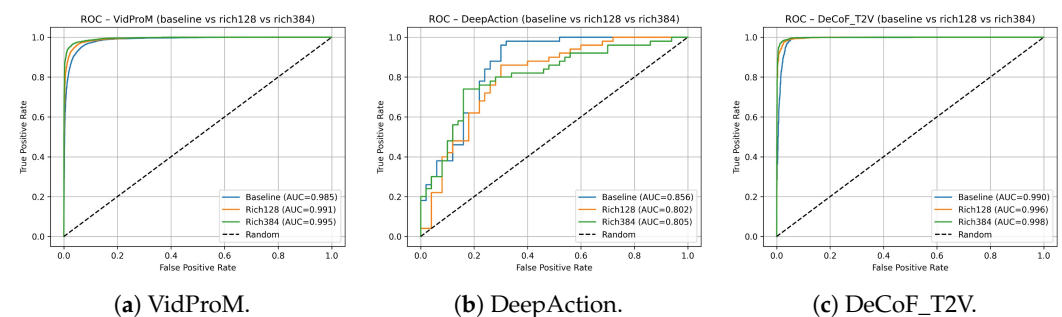


Figure 3. ROC curves for the RESTRAV baseline (21-D), Rich128 (149-D), and Rich384 (405-D). Rich128 concatenates geometry with a truncated DINOv2 temporal mean. Rich384 uses the full 384-D mean and is retained as the final configuration because it achieves the highest AUC on all three datasets. VidProM and DeCoF_T2V improve monotonically with richer features. On DeepAction, the curves cross as Rich384 trades REAL accuracy for improved synthetic recall on hard generators.

On VidProM, Rich384 raises the AUC to 0.995 (+0.010), with striking per-generator gains: OpenSora 67.9% → 92.3%, Pika 85.4% → 95.3%. On DeCoF_T2V, the AUC reaches 0.998 (+0.008), with Kling rising from 14.3% to 73.5% and Pika from 46.7% to 93.3%. On DeepAction, global AUC decreases (0.856 → 0.805), but RunwayML accuracy rises from 0% to 61.5%—a critical qualitative improvement. The global metric decrease is driven by a shift in the optimal threshold that trades REAL recall for improved synthetic recall. A focused RunwayML-only diagnostic experiment (real vs. RunwayML only) confirms that Rich384 achieves AUC = 0.913 on this generator pair versus 0.724 for the baseline, confirming that the improvement is genuine and not an artefact of multi-generator class rebalancing.

5.4. GurAI Fusion Results

Table 9 compares threshold-free ranking quality for DeMamba, Rich384, and GurAI on the aligned test cohorts.

Table 9. REAL-positive ranking metrics on the fusion-aligned test cohorts. GurAI consistently lifts both AP and AUC on VidProM and DeCoF_T2V. DeepAction ranking dips slightly under fusion because the logistic layer inherits DeMamba’s aggressive bias on the 100-sample training fold.

Dataset	Model	AP	AUC
VidProM	DeMamba	0.989	0.989
	Rich384	0.996	0.996
	GurAI	0.998	0.998
DeepAction	DeMamba	0.777	0.752
	Rich384	0.900	0.858
	GurAI	0.796	0.769
DeCoF_T2V	DeMamba	0.984	0.992
	Rich384	0.997	0.998
	GurAI	0.998	0.999

Bold indicates the best value per dataset and metric.

The per-generator analysis (Tables 10–12) illustrates where fusion adds value. On VidProM, GurAI lifts OpenSora recall from 0.915 (Rich384) to 0.969 and closes small gaps on other generators without collapsing REAL accuracy. On DeCoF_T2V, Kling recall improves from 0.872 to 0.915, and REAL accuracy is maintained at 0.990 vs. 0.986 (Rich384). On DeepAction, the per-generator picture is most dramatic: RunwayML recall rises from 0.667 (marginal towers at 0.5 threshold) to 0.889, and Veo from 0.750 to 1.000. This improvement comes at the cost of REAL accuracy (0.820 $\rightarrow$ 0.460), a trade-off that is discussed in Section 6.2.

Table 10. Per-generator fake recall (marginal towers at $\mathbb{P}(\text{FAKE}) \geq 0.5$; GurAI at fusion threshold) and REAL accuracy on VidProM fusion-aligned test cohort.

Generator	n	DeMamba	Rich384	GurAI
ModelScope	707	0.939	0.979	0.984
Text2Video-Zero	706	0.924	0.984	0.982
VideoCrafter2	706	0.977	0.993	0.997
Stable-T2V	706	0.996	0.983	1.000
CogVideo	705	0.867	0.966	0.955
Pika	706	0.996	0.935	0.993
OpenSora	706	0.936	0.915	0.969
REAL	4868	0.953	0.984	0.981

Table 11. Per-generator fake recall and REAL accuracy on DeepAction fusion-aligned test cohort (fusion threshold $\tau_{\text{fus}} = 0.06$). FPR@TPR95 and EER summarise the REAL-side error rate on the 50 test real clips (REAL-positive score convention).

Generator	n	DeMamba	Rich384	GurAI
BDAAnimateDiffLightning	9	0.889	0.889	1.000
CogVideoX-5B	8	0.750	0.875	1.000
RunwayML	9	0.667	0.667	0.889
StableDiffusion Video	8	0.750	0.875	1.000
Veo	8	0.750	0.750	1.000
VideoPoet	8	0.750	0.875	0.750

Table 11. Cont.

Generator	<i>n</i>	DeMamba	Rich384	GurAI
REAL (Acc)	50	0.660	0.820	0.460
REAL (FPR@TPR95)	50	0.740	0.640	0.740
REAL (EER)	50	0.280	0.200	0.280

Table 12. Per-generator fake recall and REAL accuracy on DeCoF_T2V fusion-aligned test cohort (fusion threshold $\tau_{\text{fus}} \approx 0.56$).

Generator	<i>n</i>	DeMamba	Rich384	GurAI
GEN2	49	0.980	1.000	1.000
Kling	47	0.872	0.872	0.915
ModelScope	481	0.977	1.000	0.998
Pika	48	0.938	0.979	1.000
Show-1	481	0.965	0.996	0.998
Sora	24	0.583	0.667	0.625
T2V	480	0.990	1.000	0.998
Veo	7	0.429	0.857	0.714
ZeroScope	480	0.942	1.000	0.998
REAL	865	0.958	0.986	0.990

5.5. Summary: All Configurations

Table 13 consolidates AUC across all configurations.

Table 13. AUC summary across all detector configurations and datasets. Starred (†) entries are on the fusion-aligned cohort, which is slightly smaller than the full test split.

Configuration	VidProM	DeepAction	DeCoF_T2V
D3 (training-free, XCLIP)	0.534	0.410	0.454
RESTRAY baseline (21-D + MLP)	0.985	0.829 [†]	0.990
DeMamba (XCLIP + Bi-Mamba)	0.989	0.772	0.992
Rich384 (405-D + MLP)	0.995	0.805	0.998
GurAI (Rich384 + DeMamba fusion) [†]	0.998	0.769	0.999

[†] Reported on fusion-aligned cohort. Bold indicates the best AUC per dataset.

6. Discussion

6.1. When Does Fusion Help? A Complementarity Analysis

The three-family benchmark reveals a consistent and interpretable pattern across two large datasets: RESTRAY/Rich384 achieves near-perfect recall on generators whose temporal statistics are captured by first-order distance and curvature summaries (ModelScope, ZeroScope, T2V, VideoCrafter2), while DeMamba recovers the generators where these geometric features fail, principally, modern DiT-based and commercial proprietary systems (OpenSora, Pika, Kling, RunwayML, Veo). The failure modes are complementary rather than correlated.

This complementarity has a structural explanation: Rich384 operates on the DINOv2 CLS token trajectory, which encodes high-level semantic content and motion direction. Its discriminative signal derives from *first-order* geometry, namely, whether consecutive frames move at consistent speeds and angles. Older generators, which often produce artefactual motion patterns, are exposed by exactly these first-order statistics. DeMamba's Bidirectional Mamba block, by contrast, scans the full spatio-temporal patch token sequence and selectively gates information across all time steps. This gives it access to subtle long-range dependencies and fine-grained spatial detail that summary statistics discard. Modern

commercial generators, trained on internet-scale video data with explicit human motion and physical plausibility objectives, produce trajectories that defeat the first-order signal but are still detectable through their fine-grained patch-level temporal structure.

Logistic stacking over logits (Equation (3)) translates this complementarity into additive probability mass: when both towers agree that a clip is fake, the fused probability is amplified; when they disagree, the stacking layer learns which tower to trust more on each training distribution. On VidProM and DeCoF_T2V, where the training fold is large enough to reliably estimate three scalar parameters, this consistently lifts threshold-free AUC above either individual tower.

6.2. The DeepAction REAL-Side Trade-Off

The DeepAction benchmark (100 real, 100 fake, six generators) presents an important and honest qualification. Under *GurAI*, REAL-positive AUC on the aligned cohort is 0.769, below Rich384 alone (0.858). The operational REAL accuracy drops to 0.460 at the fusion threshold $\tau_{\text{fus}} = 0.06$.

This behaviour is a direct consequence of the tiny training fold: with only 50 training clips per class, logistic regression for *GurAI* is fit on fewer than 100 rows. The learned DeMamba weight ($w_{\text{DM}} \approx 1.17$) heavily favours DeMamba, which, at its own training-optimal threshold, already sacrifices REAL recall for synthetic precision on DeepAction. The fused system therefore inherits DeMamba's aggressive synthetic detection bias and expresses it through a very low τ_{fus} . On the per-generator side, this is desirable—RunwayML recall rises from 0.667 to 0.889 and Veo from 0.750 to 1.000—but, on the REAL side, it causes excessive false positives.

Two observations place this trade-off in context: First, on the two large datasets, the issue does not arise: VidProM REAL accuracy is 0.981 and DeCoF_T2V is 0.990 under *GurAI*, with REAL-positive AUC of 0.998 and 0.999, respectively. Second, the DeepAction benchmark itself has 100-vs-100 balanced samples with per-generator subsets as small as seven clips, making any point estimate on it statistically fragile. Practitioners should treat DeepAction results as evidence of *which generators* expose detector weaknesses, rather than as definitive accuracy estimates.

6.3. Practical Forensic Implications

From an operational perspective, DeepAction exposes an important constraint: improving fake recall on hard generators is not sufficient when REAL-side false positives are high. In practical moderation or evidentiary pipelines, false accusations on real videos can be more damaging than missed synthetic clips. Our current recommendation is therefore conservative deployment: use RESTRAV/Rich384 as a low-cost first-pass filter and reserve DeMamba/*GurAI* for triage scenarios where human review is available.

6.4. Backbone Caveat

A limitation of the current experimental design is that RESTRAV and Rich384 use DINOv2, while D3 and DeMamba use XCLIP. We cannot fully disentangle head and backbone effects. A backbone swap experiment (replacing DINOv2 ViT-S/14 with DINOv3 ViT-S/16) yielded only +0.004 AUC on VidProM, suggesting that the geometric head is the bottleneck rather than the backbone for the 21-D configuration. Whether a geometry descriptor applied on top of XCLIP features would improve or degrade performance remains an open question for future work. A strict common-backbone ablation would require retraining all three detector families with harmonised feature extractors and preprocessing. Because this revision cycle prioritises reviewer-driven clarifications over new large-scale retraining, we report this as a planned follow-up experiment rather than an additional partial comparison.

6.5. Limitations

- **Benchmark coverage.** Additional recent benchmarks such as SynthForensics will be incorporated once publicly released datasets and stable access URLs are available.
- **Postprocessing robustness.** All experiments use raw extracted clips. In operational multimedia forensics, detectors must survive re-encoding, resizing, and platform compression (e.g., H.264/H.265). Geometric descriptors computed from DINOv2 trajectories may be sensitive to such perturbations. Evaluating AUC under FFmpeg compression at multiple CRF values is an important next step that we leave to future work because it requires a separate robustness study beyond the scope of this benchmark revision.
- **Qualitative failure visualisation.** This revision focuses on quantitative reproducibility; we do not yet include frame-level galleries of false positives/false negatives. We plan to add a curated failure-case panel in a camera-ready follow-up.
- **Scene/motion-conditioned errors.** The public datasets used here do not provide unified scene motion tags, so we cannot report stratified error rates by motion regime or scene type without additional annotation. This analysis is left for future work.
- **Fine-grained compute benchmarking.** We report qualitative compute characteristics (CPU-friendly RESTRAY/Rich384, GPU-oriented DeMamba) but not tightly controlled wall-clock and memory benchmarks across hardware stacks, since dataloading and decoding dominate timing variance in our mixed environments.
- **Fusion cohort alignment.** *GurAI* metrics are reported on the intersection of clips that survived both towers' preprocessing, which is slightly smaller than the full test split. Numbers in Tables 9–12 should not be compared directly against Table 3 without accounting for this.
- **Generator diversity.** The 22 generators across three datasets are not uniformly represented; some (Ve0 on DeCoF_T2V, $n = 14$ test clips) have too few samples for reliable per-generator estimates.

7. Conclusions

We presented a study on AI-generated video detection that combines a cross-family benchmark with two improvements to the detection pipeline.

The benchmark places three methodologically distinct detector families, namely, a training-free second-order method (D3), a geometry-supervised method (RESTRAY), and a supervised Bidirectional Mamba method (DeMamba), on identical balanced splits of three public datasets under a unified REAL-positive reporting convention for the first time. The central finding is that RESTRAY and DeMamba are competitive at the global level on large datasets (AUC difference ≤ 0.004) but exhibit strongly complementary per-generator failure modes: RESTRAY dominates on open-source bulk generators whose first-order geometric statistics are distinctive, while DeMamba recovers modern commercial generators (RunwayML, Ve0, Kling, OpenSora, Pika) where geometry compresses away the discriminative signal. D3, despite achieving state-of-the-art results on a subset of generators in its original evaluation, performs at near chance level globally in our unified protocol, primarily under a VidProM-fixed threshold transferred to the other benchmarks (Table 3; Section 5.2), demonstrating the limitations of training-free detectors when a fixed threshold must transfer across diverse generator distributions.

Building on the complementarity analysis, we proposed Rich384, a 405-dimensional feature vector that concatenates the 21-D geometric descriptor with the 384-D DINOv2 temporal mean. Rich384 raises the AUC to 0.995 on VidProM and 0.998 on DeCoF_T2V, and lifts RunwayML fake recall from 0% to 61.5% on DeepAction (still modest in absolute terms). On that small benchmark, global AUC decreases from 0.856 to 0.805. We then

introduced *GurAI*, a transparent logistic late fusion of Rich384 and DeMamba logits trained on manifest-aligned clips. *GurAI* achieves an AUC of 0.998/0.999 on VidProM/DeCoF_T2V, and, on DeepAction, improves selected per-generator fake recalls (RunwayML 0.889; Veo 1.000) relative to the marginal towers at the fusion threshold, with an explicit disclosure that REAL accuracy drops to 0.460 and global AUC to 0.769 on that cohort (Section 6.2).

The broader message is that generator diversity is the core challenge in AI-generated video detection: no single architecture or feature representation dominates all regimes. Transparent, modular late-stage fusion, rather than monolithic architectural redesign, is a practical and effective strategy for combining complementary strengths without sacrificing interpretability.

Future work will extend evaluation to the SynthForensics benchmark once it is publicly available, explore temporal sequence models (LSTM, temporal Transformer) to replace summary statistics, investigate domain adaptation to new generators without full retraining, and evaluate robustness to compression and watermarking perturbations.

Author Contributions: Conceptualisation, X.P. and R.T.; methodology, X.P.; software, X.P.; validation, X.P. and R.T.; formal analysis, X.P.; writing—original draft preparation, X.P.; writing—review and editing, R.T.; supervision, R.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been financed by Project PID2024-160996OB-I00 under grant agreement MICIU/AEI/10.13039/501100011033 and FEDER, UE, and by Generalitat de Catalunya (AGAUR) under grant agreement 2021-SGR-00478.

Data Availability Statement: VidProM is available online: <https://huggingface.co/datasets/WenhaoWang/VidProM> (accessed on 12 June 2026); DeepAction is available online: https://huggingface.co/datasets/faridlab/deepaction_v1 (accessed on 12 June 2026); DeCoF is available online: <https://github.com/wuwuwuyue/DeCoF> (accessed on 12 June 2026). Prediction CSVs and balanced manifests for all experiments are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Khachatryan, L.; Movsisyan, A.; Tadevosyan, V.; Henschel, R.; Wang, Z.; Navasardyan, S.; Shi, H. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 15954–15964. [CrossRef]
2. Open-Sora Team. Open-Sora: Democratizing Efficient Video Production for All. Available online: <https://github.com/hpcaitech/Open-Sora> (accessed on 12 June 2026).
3. Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; et al. Video Generation Models as World Simulators. Available online: <https://openai.com/index/video-generation-models-as-world-simulators> (accessed on 12 June 2026).
4. Google DeepMind. Veo: High-Fidelity Video Generation. Available online: <https://deepmind.google/technologies/veo/> (accessed on 12 June 2026).
5. Runway. Gen-3 Alpha. Available online: <https://runwayml.com/research/gen-3-alpha> (accessed on 12 June 2026).
6. Kuaishou Technology. Kling: AI Video Generation. Available online: <https://klingai.com/> (accessed on 12 June 2026).
7. Bohacek, M.; Farid, H. Human Action CLIPs: Detecting AI-Generated Human Motion. Available online: https://huggingface.co/datasets/faridlab/deepaction_v1 (accessed on 12 June 2026).
8. Hénaff, O.J.; Goris, R.L.; Simoncelli, E.P. Perceptual Straightening of Natural Videos. *Nat. Neurosci.* **2019**, *22*, 984–991. [CrossRef] [PubMed]
9. Internò, C.; Geirhos, R.; Olhofer, M.; Liu, S.; Hammer, B.; Klindt, D. AI-Generated Video Detection via Perceptual Straightening. In Proceedings of the Advances in Neural Information Processing Systems, San Diego, CA, USA, 2–7 December 2025; Volume 38.
10. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.* **2024**. Available online: <https://openreview.net/forum?id=a68SUt6zFt> (accessed on 17 July 2026). [CrossRef]

11. Wang, W.; Yang, Y. VidProM: A Million-scale Real Prompt-Gallery Dataset for Text-to-Video Diffusion Models. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024.
12. Zheng, C.; Lin, C.; Zhao, Z.; Yang, L.; Liu, S.; Yang, M.; Wang, C.; Shen, C. D3: Training-Free AI-Generated Video Detection Using Second-Order Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Honolulu, HI, USA, 19–23 October 2025; pp. 12852–12862. Available online: https://openaccess.thecvf.com/content/ICCV2025/html/Zheng_D3_Training-Free_AI-Generated_Video_Detection_Using_Second-Order_Features_ICCV_2025_paper.html (accessed on 11 June 2026).
13. Chen, H.; Hong, Y.; Huang, Z.; Xu, Z.; Gu, Z.; Li, Y.; Lan, J.; Zhu, H.; Zhang, J.; Wang, W.; et al. DeMamba: AI-Generated Video Detection on Million-Scale GenVideo Benchmark. Available online: <https://github.com/chenhaoxing/DeMamba> (accessed on 12 June 2026).
14. Ma, L.; Yan, Z.; Guo, Q.; Liao, Y.; Yu, H.; Zhou, P. Detecting AI-Generated Video via Frame Consistency (DeCoF). Available online: <https://github.com/wuwuwuyue/DeCoF> (accessed on 12 June 2026).
15. Peebles, W.; Xie, S. Scalable Diffusion Models with Transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 4195–4205. [CrossRef]
16. Hong, W.; Ding, M.; Zheng, W.; Liu, X.; Tang, J. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. In Proceedings of the International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023.
17. Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Hornung, R.; Adam, H.; Akbari, H.; Alon, Y.; Biber, V.; et al. VideoPoet: A Large Language Model for Zero-Shot Video Generation. In Proceedings of the International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
18. Pika Labs. Pika: AI Video Generation Platform. Available online: <https://pika.art/> (accessed on 12 June 2026).
19. Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; Holz, T. Leveraging Frequency Analysis for Deep Fake Image Recognition. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020; pp. 3247–3258.
20. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11. [CrossRef]
21. Ni, B.; Peng, H.; Chen, M.; Zhang, S.; Meng, G.; Fu, J.; Xiang, S.; Ling, H. Expanding Language-Image Pretrained Models for General Video Recognition. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022. [CrossRef]
22. Dong, G.; Schultz, L.; Hassanpour, N.; Gao, C. RePack then Refine: Efficient Diffusion Transformers with Vision Foundation Models. In Proceedings of the Forty-Third International Conference on Machine Learning, Seoul, Republic of Korea, 6–11 July 2026.
23. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual Event, 18–24 July 2021; pp. 8748–8763.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.