DELTA: DETECTING AND LOCALIZING GENERATED VIDEO CONTENT THROUGH GENERATIVE LATENT DIFFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Partially manipulated videos pose a growing forensic challenge, as generated content can be inserted into short intervals while the rest of the video remains authentic. Existing approaches mainly derive temporal cues from frame-level discriminative representations. We argue that focusing on individual representation states can overlook how video content evolves over time. Based on this view, we introduce DELTA, a training-free, multi-scale descriptor that characterizes local transitions along generative latent trajectories. By modeling local evolution rather than isolated latent states, DELTA captures temporal transition structure beyond frame-level discriminative representations. For full-video detection, we map continuous spatial transitions into reusable discrete patterns for robustness. We further introduce a benchmark that evaluates temporal localization of partial forgeries using paired real sources and exact interval annotations. DELTA achieves strong performance across both temporal localization and full-video detection. In particular, on the unseen-generator evaluation, replacing CLIP features with DELTA under the same ActionFormer localizer improves mean AP from 77.68 to 86.31.

1 INTRODUCTION

Recent advances in video generation have enabled the creation of increasingly realistic synthetic content (Blattmann et al., 2023; Yang et al., 2025; Wan et al., 2025). However, real-world manipulations are not limited to fully generated videos; generated content can be inserted into short temporal intervals while the remaining video stays authentic (Liu et al., 2025; Bao et al., 2026). Such partially manipulated videos introduce a distinct forensic challenge: a reliable system must determine not only whether a video contains generated content, but also when the manipulation occurs. Therefore, beyond video-level detection, temporal localization of generated segments is essential for analyzing and verifying manipulated content.

Existing generated-video forensic methods have explored various spatial and temporal cues, including visual artifacts, semantic representations, and temporal inconsistencies (Chen et al., 2026; Zheng et al., 2025; Zhang et al., 2026). However, many approaches derive temporal evidence from frame-level discriminative representations, treating videos primarily as collections of individual observations. While these representations can capture the visual content of each frame, they provide limited understanding of how video content evolves over time. This limitation is particularly important for partial forgeries, where local evolution within and around manipulated intervals may contain informative forensic evidence. This motivates a fundamental question: *Can the temporal evolution of a generative representation provide a more effective signal for generated-video forensics?*

To address this question, we view the latent sequence of a pretrained video generative model as an evolving trajectory. Although real and generated videos may look similar, their local evolution can differ because they arise from different video formation processes. Figure 1 shows region-dependent differences in latent transition magnitudes. In these examples, manipulated segments exhibit lower typical magnitudes and less temporal variation than surrounding unmodified regions. These observations motivate characterizing local latent transitions through structured statistics at multiple temporal scales.

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.

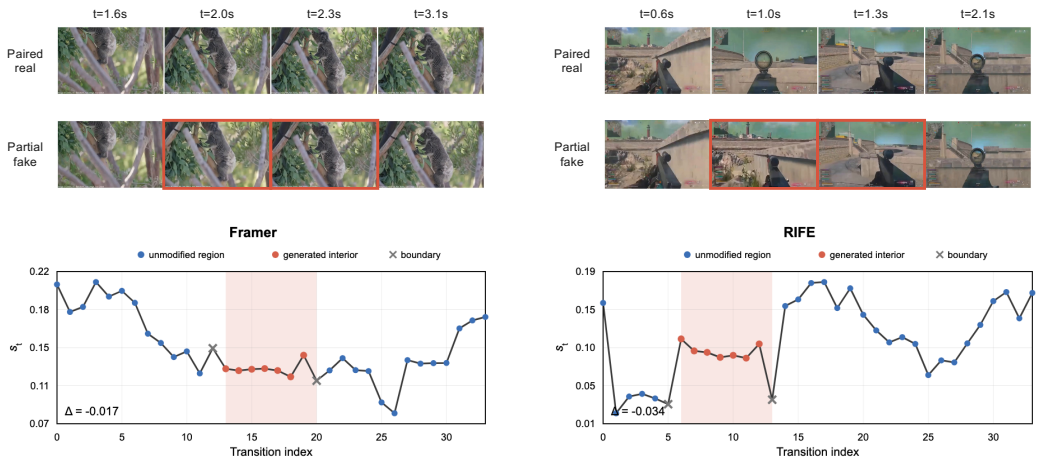


Figure 1: **Latent temporal changes around manipulated video segments.** Framer (left) and RIFE (right) examples pair real and forged frames at matched timestamps. Curves show the mean absolute adjacent-latent difference s_t in each partially forged video, with manipulated segments shaded. Δ is the median inside manipulated segments minus the median in unmodified regions, excluding boundary-crossing transitions.

We introduce DELTA, a training-free, multi-scale descriptor that characterizes local transitions along generative latent trajectories. DELTA preserves structured transition information for downstream learning. For temporal localization, these transitions are summarized into a temporal feature sequence, while for full-video detection, the spatial structure is retained and discretized by an FSQ quantizer (Mentzer et al., 2024) before classification. The quantizer suppresses fine-grained continuous variation while preserving reusable spatial transition patterns, so that the video-level detection relies more on transition patterns than on individual continuous transition values. This design provides a general representation of temporal transition structure that can support different forensic tasks with task-specific prediction heads.

To evaluate transition-based evidence across manipulation settings, we further introduce a controlled benchmark for partial-forgery temporal localization with paired sources, exact temporal annotations, and source-disjoint splits. In particular, RIFE and SVD-XT manipulate matched temporal intervals, enabling controlled comparison across manipulation mechanisms.

We evaluate DELTA for temporal localization on ActivityForensics (Bao et al., 2026) and our controlled benchmark, and for full-video detection on GenVideo (Chen et al., 2026) and FakeParts (Liu et al., 2025). DELTA improves mean AP over ActionFormer from 67.59 to 84.04 in the ActivityForensics intra-domain evaluation and from 77.68 to 86.31 in its open-world evaluation on the unseen Vidu generator. It also achieves 96.30 average AUROC on GenVideo and 74.51 on FakeParts.

This paper contributes:

- We introduce DELTA, a training-free, multi-scale descriptor that characterizes structured local transitions in generative latent space rather than isolated frame states.
- We introduce a partial-forgery temporal localization benchmark for evaluating transfer across manipulation methods. It pairs real sources with Framer, RIFE, and SVD-XT forgeries, provides exact interval annotations and source-disjoint splits, and matches interval placement between RIFE and SVD-XT.
- DELTA achieves strong results in temporal localization and full-video detection. Horizon ablations show that short-range transitions are strongest individually, while longer gaps add complementary evidence; paired analysis links transition deviations to manipulated segments.

2 Related Work

2.1 Generated Video Detection

Generated-video detectors combine spatial, semantic, and temporal cues (Wang et al., 2020; Radford et al., 2021; Ojha et al., 2023; Tan et al., 2024; 2025). DeMamba introduces GenVideo and models spatial details and temporal inconsistencies with Detail Mamba (Chen et al., 2026). D3 (Zheng et al., 2025) summarizes changes in adjacent-frame feature distances into a clip-level score; distance reduction discards signed channel responses. Generative-model-based detectors include diffusion-reconstruction methods DIRE and DRCT (Wang et al., 2023; Chen et al., 2024), and NSG-VD, which estimates spatiotemporal gradients from diffusion-model scores (Zhang et al., 2026). DBINDS (Wu et al., 2026) combines multi-domain, multi-scale statistics of inversion-derived noise differences for video classification. DELTA operates directly on generative latent maps, rather than semantic frame embeddings or inversion-derived noise. It summarizes each channel at multiple temporal gaps, retaining signed responses, magnitudes, and temporal locations for interval prediction (Appendix H.1).

2.2 Temporal Forgery Localization

Temporal forgery localization extends binary detection to identifying manipulated intervals within a video. Existing benchmarks cover several manipulation settings. ForgeryNet studies temporal localization of facial forgeries (He et al., 2021), while LAV-DF (Cai et al., 2022), AV-Deepfake1M (Cai et al., 2024), and AV-Deepfake1M++ (Cai et al., 2025) focus on temporally localized audio-visual manipulations in talking-face videos. FakeParts covers partial visual edits across multiple manipulation types (Liu et al., 2025), and ActivityForensics focuses on manipulated human activities under intra-domain, cross-domain, and open-world settings (Bao et al., 2026). Concurrent work TASLE studies sparse AI-generated segments embedded in long-form videos (Feng et al., 2026). Together, these benchmarks span facial, audio-visual, activity-centric, and general partial-forgery settings.

Existing localization methods model temporal forgery cues in different ways. ActionFormer provides a general temporal localization architecture based on local self-attention and multi-scale temporal feature modeling (Zhang et al., 2022). UMMAFormer (Zhang et al., 2023) introduces temporal anomaly modeling for video-inpainting localization, while TADiff (Bao et al., 2026) builds on ActionFormer and uses diffusion-based feature regularization to localize manipulated activities. More recent work combines structural and semantic evidence with diffusion-guided refinement for temporal forgery localization. Generic temporal detection architectures such as DiGIT (Kim et al., 2025) also provide related mechanisms for modeling and predicting temporal segments.

3 Task Settings and Controlled Benchmark

3.1 Two Forensic Tasks

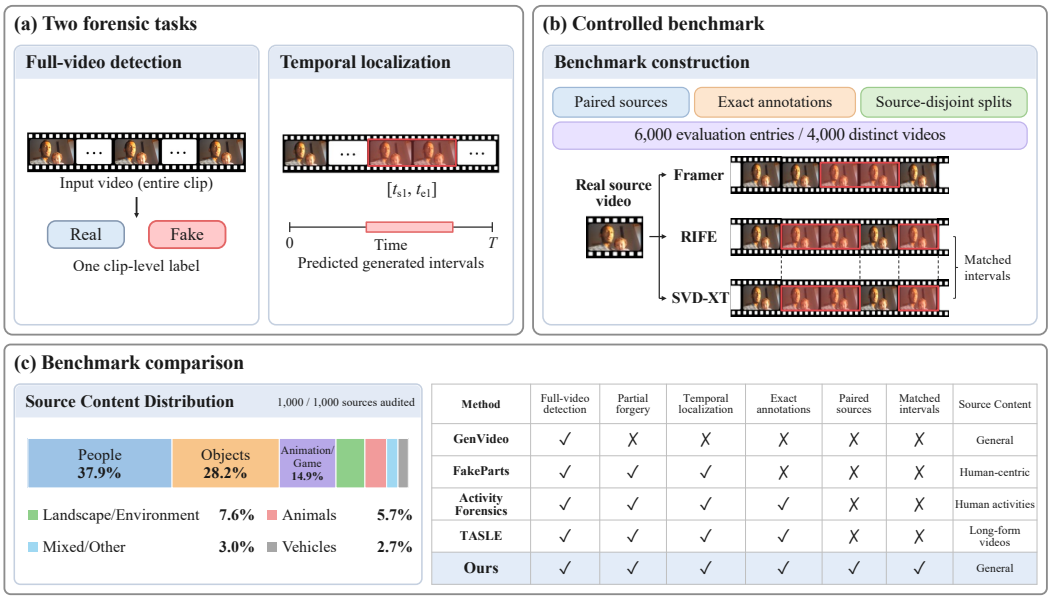
Let $V = (x_1, \ldots, x_T)$ denote a video of T frames, where x_t is the tth frame. Full-video detection predicts a label $y \in \{\text{real}, \text{fake}\}$, whereas temporal forgery localization predicts a ranked set

$$\mathcal{P}(V) = \left\{ \left([t_s^{(k)}, t_e^{(k)}], c_k \right) \right\}_{k=1}^{K},$$

where K is the number of predicted intervals, $t_s^{(k)}$ and $t_e^{(k)}$ are the start and end frame indices of the kth interval, satisfying $1 \le t_s^{(k)} \le t_e^{(k)} \le T$, and c_k is its confidence score. We evaluate video-level decisions with AUROC and ranked interval predictions with temporal AP.

3.2 Controlled Partial-Forgery Benchmark

To complement established benchmarks with controlled evaluation of temporal localization, we construct a partial-forgery benchmark in which one or more contiguous intervals of a real source video are replaced with generated content. The benchmark spans three manipulation domains—Framer (Wang et al., 2025), RIFE (Huang et al., 2022), and SVD-XT (Blattmann et al., 2023)—and



Method	Full-video detection	Partial forgery	Temporal localization	Exact annotations	Paired sources	Matched intervals	Source Content
GenVideo	✓	✗	✗	✗	✗	✗	General
FakeParts	✓	✓	✓	✗	✗	✗	Human-centric
Activity Forensics	✓	✓	✓	✓	✗	✗	Human activities
TASLE	✓	✓	✓	✓	✗	✗	Long-form videos
Ours	✓	✓	✓	✓	✓	✓	General

Figure 2: **Task setting and controlled partial-forgery benchmark.** (a) Full-video detection predicts one clip-level label, whereas temporal localization predicts one or more generated intervals. (b) The benchmark constructs Framer, RIFE, and SVD-XT partial forgeries from paired real sources with exact annotations. RIFE and SVD-XT share the same interval locations, and the lower panel summarizes the content distribution of the 1,000 source videos. (c) Comparison with representative generated-video and temporal-forgery benchmarks.

pairs every partial forgery with its unmodified real source, allowing generated and real intervals to be compared while holding source content fixed.

Across the three domains, the benchmark contains 6,000 evaluation entries corresponding to 4,000 distinct videos, with exact temporal annotations in frame indices and seconds. Framer contains one fixed-length interval, whereas RIFE contains one to three variable-length intervals. SVD-XT reuses the sampled locations and durations from RIFE, holding both source content and interval placement fixed when comparing these two manipulation pipelines.

Source identities are separated across train, validation, and test splits. Models are trained and selected on Framer and then evaluated unchanged on Framer, RIFE, and SVD-XT, providing both in-domain and cross-pipeline evaluation.

Source and content coverage. The 1,000 real source videos are sampled from TenKReal, the real-video set used by FakeParts (Liu et al., 2025), and cover seven content categories: people, objects, animation/game footage, landscapes and environments, animals, vehicles, and mixed or other visual content.

Relation to existing benchmarks. Figure 2 contrasts our benchmark with full-video generation, partial visual editing, activity manipulation, and long-form segment-forgery datasets. Its distinguishing control is matched interval placement across RIFE and SVD-XT, allowing manipulation methods to be compared on the same sources and temporal locations.

4 DELTA

Figure 3 summarizes the method. DELTA first constructs multi-scale transitions from frozen generative latent maps and then preserves the form of evidence needed by each task: compact temporal descriptors for localization and spatial transition maps for full-video detection.

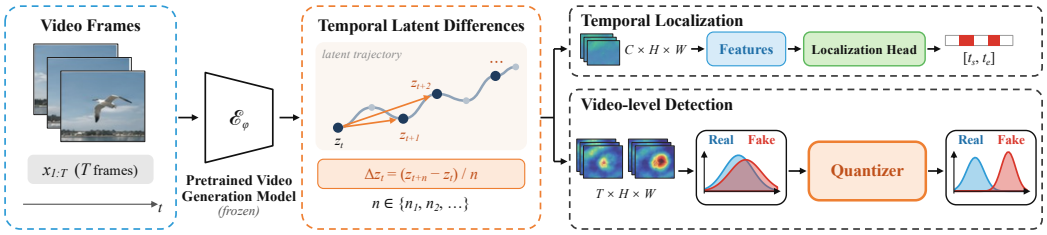


Figure 3: **Overview of DELTA.** Temporally separated latent maps from a frozen pretrained video generative model yield local transition features. Spatial summaries support temporal localization, while spatial difference maps support video-level detection through an FSQ quantizer.

4.1 Multi-Scale Latent Transitions

Let E denote a frozen representation encoder from a pretrained video generative model. We encode each frame independently as

$$z_t = E(x_t), \qquad z_t \in \mathbb{R}^{C \times H' \times W'}. \tag{1}$$

We treat the resulting latent sequence $\{z_t\}$ as a trajectory in generative latent space. Temporal context is introduced by the differences and downstream head, not by cross-frame encoding.

To characterize local temporal changes at different scales, we choose an ordered set of temporal gaps $\mathcal{H} = \{h_1, \ldots, h_M\}$, with $1 \le h_1 < \cdots < h_M$, and compute horizon-normalized latent differences

$$\Delta_h z_t = \frac{z_{t+h} - z_t}{h}, \qquad h \in \mathcal{H}. \tag{2}$$

The normalization by h makes differences computed over different temporal gaps comparable.

For temporal localization, $\phi(u)$ summarizes a spatial map u by its signed mean, absolute mean, root mean square, and standard deviation. With $u_{t,c}^{(h)} = \Delta_h z_{t,c}$, we concatenate these statistics across channels and temporal gaps:

$$d_t = \bigoplus_{h \in \mathcal{H}} \bigoplus_{c=1}^{C} \phi\left(u_{t,c}^{(h)}\right). \tag{3}$$

The descriptor has $4CM$ dimensions. Section 5.2 specifies the experimental setting; Appendix D gives the statistic definitions and temporal alignment.

DELTA extraction introduces no learned parameters. For temporal localization, the descriptor sequence $\{d_t\}$ is passed to the temporal localizer. For video-level detection, we retain the spatial latent-difference maps instead of spatially summarizing them.

4.2 Temporal Localization

The DELTA descriptors are passed to ActionFormer (Zhang et al., 2022) to predict forged intervals. Each d_t is a temporally aligned token, and the localizer converts the descriptor sequence into ranked interval proposals. For comparison, CLIP frame features use the same ActionFormer architecture, training procedure, and evaluator, isolating the effect of the representation.

4.3 Quantized Full-Video Detection

Full-video detection aggregates spatial difference maps into one prediction. Because exact latent-difference values can vary with content, motion, and video processing, we apply FSQ (Mentzer et al., 2024) to emphasize recurring transition patterns over fine-grained continuous variation. A spatial encoder forms low-dimensional feature groups, FSQ discretizes them, and a CNN with temporal pooling produces the video representation. The encoder and classifier learn from video labels, while FSQ uses no learned codebook. Architecture and training details appear in Appendix D.

5 EXPERIMENTS

We evaluate DELTA along three axes: full-video detection across unseen generators and manipulation tasks, temporal localization with an established localizer, and transfer across the manipulation domains of the controlled benchmark.

5.1 EXTERNAL BENCHMARKS, METRICS, AND PROTOCOLS

ActivityForensics. ActivityForensics (Bao et al., 2026) inserts generated or edited human activities into otherwise authentic videos. Its official protocol uses 3,029 training, 795 validation, and 1,297 intra-domain test videos; open-world evaluation uses 265 videos from the unseen Vidu generator. We report temporal AP at tIoU 0.75, 0.85, and 0.95 and their mean (mAP). Further details are in Appendix C.

Controlled benchmark. The benchmark in Section 3 contains 1,000 source videos, each paired with three manipulation variants. We split source identities 600/200/200 for training, validation, and testing. We report temporal mAP at the same three tIoU thresholds, alongside frame AUROC.

GenVideo. Following GenVideo (Chen et al., 2026), we train on its Pika-generated videos and Kinetics-400 (Kay et al., 2017) real videos. The protocol uses 19,976 training, 797 validation, and 5,051 test videos, covering full-video synthesis from ten held-out generators at test time. We follow the official GenVideo split and real-reference construction for all ten unseen generators and report their equal-weight mean AUROC. Recent work has shown that GenVideo results can be sensitive to preprocessing order (Michels et al., 2026). We therefore report the original pipelines and additionally evaluate all methods under the same post-downscale JPEG preprocessing, where frames are resized before JPEG encoding. Exact resize, filtering, and encoding settings are given in the appendix.

FakeParts. FakeParts (Liu et al., 2025) covers eight tasks: generation conditioned on text or images, interpolation, extrapolation, inpainting, outpainting, face swaps, and style changes. Our evaluation pool contains approximately 81K videos. We transfer the Pika/Kinetics-400-trained detector without target training. Following the benchmark, each task is evaluated against its designated real distribution: seven use TenKReal, while Inpainting uses paired ROVI. We report per-task AUROC and their equal-weight mean.

5.2 BASELINES AND IMPLEMENTATION

Temporal localization baselines. On ActivityForensics, all methods follow the official split: training uses the five seen generators, intra-domain testing uses those same generators, and open-world testing uses Vidu without further tuning. ActionFormer (Zhang et al., 2022) uses CLIP (Radford et al., 2021) frame features as its input representation. TADiff (Bao et al., 2026) regularizes temporal features through diffusion, DiGIT (Kim et al., 2025) uses multi-dilated temporal encoding, and UMMAFormer (Zhang et al., 2023) models temporal anomalies. On our benchmark, UFD (Ojha et al., 2023) and C2P-CLIP (Tan et al., 2025) add CLIP-based forgery cues, while RAFT (Teed & Deng, 2020) supplies optical-flow cues. On our controlled benchmark, localizers train on the Framer training split; frozen baselines are calibrated using only Framer data. All methods use the same evaluation protocol on the Framer, RIFE, and SVD-XT test splits.

Video-level detection baselines. On GenVideo, we compare with DeMamba (Chen et al., 2026), NPR (Tan et al., 2024), TALL (Xu et al., 2023), STIL (Gu et al., 2021), and NSG-VD (Zhang et al., 2026). On FakeParts, we evaluate NPR, CNNDet (Wang et al., 2020), DeMamba, C2P-CLIP, and D3 (Zheng et al., 2025). Retrained baselines and DELTA use 9,997 Pika fake videos and 9,979 Kinetics-400 real videos. D3 is training-free.

Implementation. We instantiate E with the Wan2.1 VAE encoder (Wan et al., 2025), keeping it fixed while training the task-specific components. Localization uses $\mathcal{H} = \{1, 2, 4, 8\}$ and $C = 16$, producing 256-dimensional descriptors. Video-level detection uses adjacent differences of eight uni-

formly sampled frames. Model selection uses only validation data. Appendix D gives preprocessing, alignment, head architectures, and training and inference settings.

5.3 Temporal Forgery Localization

ActivityForensics. On ActivityForensics (Table 1), DELTA raises mean AP over ActionFormer from 67.59 to 84.04 in the intra-domain setting and from 77.68 to 86.31 in the open-world setting, corresponding to relative improvements of 24.3% and 11.1%, respectively. The gains are largest at tIoU 0.95, suggesting more precise boundary cues under the same localizer. The open-world improvement further supports transfer of these cues to the unseen Vidu generator.

Table 1: Temporal localization under the official ActivityForensics intra-domain and open-world protocols. Values are temporal AP (%). mAP averages the three thresholds. Bold/underline indicate best/second-best. Gains are over ActionFormer, in percentage points.

	Intra-domain				Open-World			
Method	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	**mAP** ↑	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	**mAP** ↑
UMMAFormer (Zhang et al., 2023)	<u>87.70</u>	80.54	52.66	73.63	91.03	87.58	60.50	79.70
DiGIT (Kim et al., 2025)	81.47	73.30	50.19	68.32	89.65	83.81	58.72	77.39
TADiff (Bao et al., 2026)	87.21	<u>81.41</u>	<u>55.24</u>	<u>74.62</u>	<u>91.93</u>	**90.10**	<u>68.06</u>	<u>83.36</u>
ActionFormer (Zhang et al., 2022)	82.56	75.18	45.03	67.59	88.61	83.45	60.98	77.68
DELTA	**91.37**	**87.69**	**73.06**	**84.04**	**92.13**	<u>88.72</u>	**78.08**	**86.31**
Gain over ActionFormer	+8.80	+12.52	+28.03	+16.45	+3.52	+5.27	+17.10	+8.63

Controlled partial-forgery benchmark. Table 2 reports strict temporal mAP and frame AUROC. Models trained or calibrated on Framer are transferred to RIFE and SVD-XT without further tuning. DELTA is strongest on Framer and RIFE, exceeding ActionFormer by 50.00 and 8.53 mAP points; ActionFormer performs better on SVD-XT. DELTA also lowers the real-video false-alarm rate from 9.67% to 3.83%. Against ActionFormer, the small frame-AUROC gap but large mAP gap on Framer suggests better interval delineation. The RIFE gain further supports transfer of transition cues across manipulation pipelines.

Table 2: Localization on the controlled partial-forgery benchmark (%). mAP averages temporal AP at tIoU 0.75, 0.85, and 0.95; AUROC is frame-level. Bold/underline indicate best/second-best.

	SVD-XT		RIFE		Framer	
Method	mAP ↑	AUROC ↑	mAP ↑	AUROC ↑	mAP ↑	AUROC ↑
UMMAFormer (Zhang et al., 2023)	6.50	83.63	2.27	75.16	8.34	84.95
UFD (Ojha et al., 2023)	0.10	63.78	0.01	57.15	0.05	59.13
C2P-CLIP (Tan et al., 2025)	0.23	64.76	0.02	53.95	0.08	57.55
RAFT (Teed & Deng, 2020)	0.13	45.52	0.01	37.25	0.06	38.85
TADiff (Bao et al., 2026)	<u>18.64</u>	86.65	3.49	71.08	30.58	89.52
ActionFormer (Zhang et al., 2022)	**21.79**	**89.31**	<u>3.52</u>	<u>75.87</u>	<u>38.52</u>	<u>93.60</u>
DELTA	17.13	<u>87.37</u>	**12.05**	**79.73**	**88.52**	**97.21**

5.4 Video-Level Detection on GenVideo and FakeParts

Tables 3 and 4 report the per-generator GenVideo results and the per-task FakeParts results, respectively. On GenVideo (Table 3), DELTA averages 96.30 AUROC under the original pipelines. Under identical post-downscale JPEG preprocessing, NSG-VD drops from 95.50 to 84.41, whereas DELTA retains 95.28. This smaller drop indicates that DELTA's temporal evidence survives a common encoding pipeline and is not explained solely by preprocessing cues. On FakeParts, DELTA reaches 74.51 average AUROC versus 73.24 for NPR. The per-task results show complementary strengths across manipulation types, while the aggregate gain indicates that training on fully generated videos can transfer useful transition cues to partial edits.

Table 3: **Video-level detection on GenVideo and post-downscale JPEG robustness.** Values are AUROC (%). Upper: original method-specific preprocessing; lower: identical post-downscale JPEG preprocessing. Superscripts show the change from each method's original Avg., in percentage points.

	AUROC (%) ↑										
Method	MS	Morph	Moon	HotShot	Show-1	Gen-2	Crafter	LaVie	Sora	Wild	Avg. ↑
Original method-specific preprocessing											
DeMamba (Chen et al., 2026)	**97.79**	**98.90**	99.82	84.89	88.27	99.34	97.71	92.58	82.04	89.37	93.07
NPR (Tan et al., 2024)	87.52	95.80	99.36	77.58	85.54	98.77	96.62	83.24	82.88	87.00	89.43
TALL (Xu et al., 2023)	94.53	96.05	99.29	91.90	95.31	99.33	98.44	93.70	86.48	92.85	94.79
STIL (Gu et al., 2021)	96.52	97.50	97.43	86.26	89.33	97.02	94.74	89.70	84.07	89.38	92.19
NSG-VD (Zhang et al., 2026)	93.15	97.34	**99.88**	88.37	95.64	97.63	**99.03**	92.64	**96.74**	94.60	95.50
DELTA	93.72	96.33	99.21	**98.24**	**97.05**	**99.53**	98.60	**95.73**	89.67	**94.91**	**96.30**
Post-downscale JPEG preprocessing											
DeMamba (Chen et al., 2026)	96.72	97.24	98.94	81.84	87.12	97.68	94.89	91.20	**89.18**	85.51	92.03 $^{-1.04}$
NPR (Tan et al., 2024)	91.77	94.41	97.43	69.66	70.66	95.40	92.75	80.74	80.25	80.19	85.33 $^{-4.10}$
TALL (Xu et al., 2023)	93.77	95.09	98.94	93.03	95.68	98.73	97.99	93.75	81.76	93.22	94.20 $^{-0.59}$
STIL (Gu et al., 2021)	**96.92**	**98.33**	98.35	91.83	92.51	98.41	96.90	92.97	86.73	89.99	94.29 $^{+2.10}$
NSG-VD (Zhang et al., 2026)	82.09	82.85	97.55	70.76	84.38	86.35	93.80	78.12	86.83	81.41	84.41 $^{-11.09}$
DELTA	93.06	94.38	**99.43**	**97.05**	**96.48**	**98.98**	**98.84**	**96.61**	84.21	**93.79**	**95.28** $^{-1.02}$

Table 4: **Video-level detection on FakeParts.** Values are AUROC (%); Avg. is the equal-weight mean over the eight tasks.

Method	TI2V	T2V	Interp.	Extrap.	Inpaint.	Outpaint.	Face	Style	Avg. ↑
NPR (Tan et al., 2024)	**97.58**	**91.35**	70.36	55.75	60.13	56.78	73.17	80.83	73.24
CNNDet (Wang et al., 2020)	93.96	88.66	53.92	**57.70**	55.98	52.74	81.66	86.38	71.37
DeMamba (Chen et al., 2026)	88.90	87.26	55.44	46.95	54.18	59.22	**84.10**	**89.25**	70.66
C2P-CLIP (Tan et al., 2025)	67.66	58.11	73.77	56.63	**61.91**	82.22	55.43	74.78	66.31
D3 (Zheng et al., 2025)	65.42	73.81	30.22	34.26	58.28	72.50	57.01	54.51	55.75
DELTA	77.24	84.43	**77.63**	53.31	57.44	**94.93**	69.15	81.92	**74.51**

6 Analysis

Table 5: State–transition ablation (Activity-Forensics mAP, %).

	Intra-domain ↑		Open-World ↑	
Representation	State	Transition	State	Transition
RGB	21.73	65.23	6.44	44.74
CLIP (Radford et al., 2021)	18.99	44.78	13.31	36.93
Wan (Wan et al., 2025)	**56.11**	**84.04**	**64.39**	**86.31**

Table 6: Temporal-horizon ablation (Activity-Forensics mAP, %).

Horizons	Intra-domain ↑	Open-World ↑
$h = 1$	82.56	81.47
$h = 2$	79.53	78.00
$h = 4$	75.62	73.32
$h = 8$	69.88	68.40
$h \in \{1, 2, 4, 8\}$	**84.04**	**86.31**

Representation and temporal differencing. Table 5 shows that transitions outperform states for every representation, while Wan is strongest in both forms. Differencing Wan raises intra-domain mAP from 56.11 to 84.04 and open-world mAP from 64.39 to 86.31. Thus both the generative latent space and its temporal changes contribute to localization. RGB and CLIP also benefit from differencing but remain below Wan, showing that the gain depends on both the temporal construction and the representation being differenced.

Multi-scale first-order transitions. The adjacent gap $h = 1$ is strongest alone, consistent with precise short-range boundary cues. Combining $h \in \{1, 2, 4, 8\}$ nevertheless improves over $h = 1$ by 1.48 mAP under intra-domain evaluation and 4.84 mAP under open-world evaluation (Table 6). Performance decreases as the individual horizon grows, but the combined descriptor exceeds every single-horizon result. The improvement is larger in the open-world setting.

Spatial statistics. DELTA combines signed mean, absolute mean, RMS, and spatial standard deviation for each latent channel. Together they outperform every statistic alone; Appendix H.3 reports

the single-statistic and leave-one-out results. No individual statistic reproduces the full descriptor, and removing signed mean causes the largest leave-one-out drop.

Boundary-aligned latent transition changes. Across 600 validation pairs, standardized fake–real transition distances are around 4 times larger inside manipulated segments than outside (Figure 4). Unmodified regions remain closer to their paired real sources, so the difference is localized within the video.

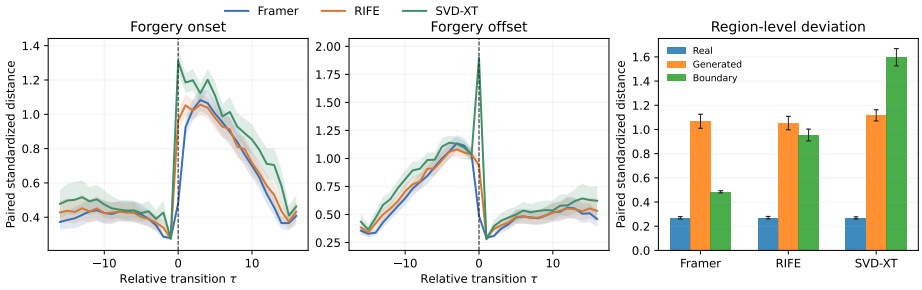


Figure 4: **Boundary-aligned paired transition analysis.** Fake–real transition distances around manipulation boundaries; bands show 95% bootstrap intervals.

Robustness to input degradation and encoder replacement. Table 7 evaluates fixed checkpoints under 360p and H.264 degradation. DELTA loses 7.82 and 15.80 mAP, versus 28–30 and 40–41 for the baselines, indicating that latent transitions retain more localization evidence under input degradation. Consequently, it has less dependence on fine appearance cues. Table 8 replaces Wan2.1 with CogVideoX1.5 while retaining the localization pipeline. CogVideoX1.5 reaches 78.72/79.77 mAP, above ActionFormer in both evaluation settings. Wan2.1 improves this to 84.04/86.31, with the largest gap at AP@0.95, where accurate boundary placement matters most.

Table 7: Robustness on ActivityForensics intra-domain (mAP, %). Colored arrows show absolute drops from Original.

Method	Original ↑	360p ↑	H.264 CRF32 ↑
TADiff (Bao et al., 2026)	74.62	46.76 ↓27.86	33.13 ↓41.49
ActionFormer (Zhang et al., 2022)	67.59	37.57 ↓30.02	27.30 ↓40.29
DELTA	**84.04**	**76.22** ↓ 7.82	**68.24** ↓15.80

Table 8: Encoder replacement under the official ActivityForensics intra-domain and open-world protocols. Values are temporal AP and mAP (%).

	Intra-domain				Open-World			
Encoder	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	mAP ↑	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	mAP ↑
CogVideoX1.5 (Yang et al., 2025)	89.80	84.91	61.46	78.72	89.65	84.96	64.70	79.77
Wan2.1 (Wan et al., 2025)	**91.37**	**87.69**	**73.06**	**84.04**	**92.13**	**88.72**	**78.08**	**86.31**

7 Conclusion

We presented DELTA, a multi-scale generative-latent transition representation for temporal localization and video-level detection, together with a controlled partial-forgery benchmark. DELTA summarizes multi-scale differences for interval localization and quantizes spatial difference maps for video-level classification. Across ActivityForensics, our benchmark, GenVideo, and FakeParts, DELTA generalizes to unseen generators and manipulation pipelines. Ablations show that generative latent features, temporal differencing, complementary horizons, and spatial statistics all contribute. These results show that temporal changes in generative latents provide useful cues for both localization and video-level detection.

Reproducibility Statement

Sections 4–5 describe the method, evaluation protocols, and experimental settings. Appendix C documents the datasets and splits, while Appendix D provides preprocessing, architecture, optimization, and model-selection details. We will release configurations, dataset manifests, checkpoints, per-video outputs, and scripts for reconstructing the reported tables.

Ethics Statement

This work aims to support media verification by detecting and localizing generated video content. It uses existing video datasets and constructs partial forgeries for research evaluation. Potential harms include false attribution, uneven accuracy across video domains, and misuse of the generated examples. Dataset use and release should respect source licenses and privacy restrictions. Detector outputs should inform human review alongside provenance evidence, rather than determine a video's origin or an individual's intent.

AI Use Statement

Generative AI tools were used only to polish the language and improve the clarity of the manuscript. The authors reviewed all AI-assisted revisions and take responsibility for the final content of this work.

References

Peijun Bao, Anwei Luo, Gang Pan, Alex C Kot, and Xudong Jiang. Activityforensics: A comprehensive benchmark for localizing manipulated activity in videos. *arXiv preprint arXiv:2604.03819*, 2026.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Zhixi Cai, Kalin Stefanov, Abhinav Dhall, and Munawar Hayat. Do you really mean that? content driven audio-visual deepfake dataset and multimodal method for temporal forgery localization. In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–10. IEEE, 2022.

Zhixi Cai, Shreya Ghosh, Aman Pankaj Adatia, Munawar Hayat, Abhinav Dhall, Tom Gedeon, and Kalin Stefanov. Av-deepfake1m: A large-scale llm-driven audio-visual deepfake dataset. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 7414–7423, 2024.

Zhixi Cai, Kartik Kuckreja, Shreya Ghosh, Akanksha Chuchra, Muhammad Haris Khan, Usman Tariq, Tom Gedeon, and Abhinav Dhall. Av-deepfake1m++: A large-scale audio-visual deepfake benchmark with real-world perturbations. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 13686–13691, 2025.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *ICML*, volume 2, pp. 6, 2024.

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.

Yue Feng, Jingjing Li, Qijia Lu, Wei Ji, Jingrou Zhang, Fei Shen, Xiao Li, Yizhen Jia, Qiang Chen, Limin Wang, et al. Explainable forensics of manipulated segments in untrimmed long videos. *arXiv preprint arXiv:2606.02402*, 2026.

Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pp. 3473–3481, 2021.

Yinan He, Bei Gan, Siyu Chen, Yichun Zhou, Guojun Yin, Luchuan Song, Lu Sheng, Jing Shao, and Ziwei Liu. Forgerynet: A versatile benchmark for comprehensive forgery analysis. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4358–4367. IEEE, 2021.

Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European conference on computer vision*, pp. 624–642. Springer, 2022.

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

Ho-Joong Kim, Yearang Lee, Jung-Ho Hong, and Seong-Whan Lee. Digit: Multi-dilated gated encoder and central-adjacent region integrated decoder for temporal action detection transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24286–24296. IEEE, 2025.

Ziyi Liu, Firas Gabetni, Awais Hussain Sani, Xi Wang, Soobash Daiboo, Gaetan Brison, Gianni Franchi, and Vicky Kalogeiton. Fakeparts: a new family of ai-generated deepfakes. *arXiv preprint arXiv:2508.21052*, 2025.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

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.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24480–24489. IEEE, 2023.

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. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

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

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.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8692–8701. IEEE, 2020.

Wen Wang, Qiuyu Wang, Kecheng Zheng, Hao Ouyang, Zhekai Chen, Biao Gong, Hao Chen, Yujun Shen, and Chunhua Shen. Framer: Interactive frame interpolation. In *International Conference on Learning Representations*, volume 2025, pp. 51601–51628, 2025.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22388–22398. IEEE, 2023.

Yanlin Wu, Xiaogang Yuan, and Dezhi An. Dbinds: detection based on initial noise difference sequence from diffusion model inversion for ai-generated videos. *Journal of King Saud University Computer and Information Sciences*, 38(7):744, 2026.

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.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22601–22611. IEEE, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European conference on computer vision*, pp. 492–510. Springer, 2022.

Rui Zhang, Hongxia Wang, Mingshan Du, Hanqing Liu, Yang Zhou, and Qiang Zeng. Ummaformer: A universal multimodal-adaptive transformer framework for temporal forgery localization. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 8749–8759, 2023.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. *Advances in neural information processing systems*, 38:174683–174730, 2026.

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 *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1979–1989. IEEE, 2025.

A Limitations and Future Directions

DELTA relies on frozen pretrained generative encoders to construct its latent transition representation. Although feature extraction itself requires no additional training, large video generative encoders such as Wan2.1 and CogVideoX1.5 introduce non-trivial computation and storage overhead compared with lightweight visual backbones. Improving encoder efficiency or distilling the transition representation would therefore be useful for large-scale or real-time forensic deployment.

The quality of DELTA also depends on the representation space provided by the pretrained generative encoder. As video generation models continue to evolve, the most suitable encoder may vary across domains and manipulation families. Exploring lighter and more broadly transferable generative representations is an important direction.

Finally, the current temporal horizons are defined in sampled-frame units. Their effective duration therefore changes with frame rate. Time-normalized or adaptive temporal scales could provide a more consistent representation for videos with heterogeneous temporal sampling.

B Broader Implications

Temporal localization can help media reviewers identify which parts of a video warrant closer inspection, rather than treating the entire clip as suspect. The paired benchmark also provides a way to study localized manipulation while controlling source content and interval placement. In practical use, false alarms can cast doubt on authentic footage, and detection accuracy may vary across domains. Outputs should therefore complement provenance information and human review, with dataset use and release respecting source licenses and privacy.

C Detailed Dataset and Protocol Definitions

GenVideo training uses 9,997 Pika videos as the generated class and 9,979 Kinetics-400 videos as the real class. Its 797-video validation split covers 9 held-out generators. The test set contains 4,551 fake videos from 10 generators and a shared set of 500 MSR-VTT real videos (Xu et al., 2016), for 5,051 videos in total. Each generator AUROC uses this same real set, and the reported Avg. is the unweighted mean over the ten generator-wise AUROCs. FakeParts contains eight manipulation tasks; seven use TenKReal as the real distribution, while Inpainting uses its paired ROVI real videos. ActivityForensics uses 3,029 training and 795 validation videos from five seen generators. Its official intra-domain evaluation uses 1,297 test videos from those generators, while open-world evaluation uses 265 videos from the unseen Vidu generator.

The controlled benchmark contains paired real and partial-forgery videos from Framer, RIFE, and SVD-XT. Its 1,000 source identities are split 600/200/200 for training, validation, and test, with no source identity crossing partitions. For every manipulation pipeline, this yields 1,000 partial forgeries paired with 1,000 real source videos. Each partial forgery records its source identity, manipulation pipeline, interval boundaries in frames and seconds, frame labels, and split assignment. Framer is used for training and model selection; RIFE and SVD-XT are evaluated without additional tuning. Machine-readable manifests and all exclusions will accompany the released benchmark.

D Detailed Implementation

This section records the configuration used for the reported DELTA systems. The video-level and localization heads are trained independently; they share the frozen latent representation but are not score-fusion variants.

D.1 Frozen Encoder and Input Construction

Wan latent maps. Frames are decoded as RGB, resized to 224×224 by bilinear interpolation, and normalized from $[0, 1]$ to $[-1, 1]$. Each frame is encoded independently with the frozen Wan2.1 VAE in evaluation mode and bfloat16 precision. The resulting state is

$$z_t \in \mathbb{R}^{16\times 28\times 28}. \tag{4}$$

Encoding frames independently ($T = 1$ per VAE call) prevents the encoder from mixing information across the temporal boundary being measured. Stored latent differences use float16; all downstream normalization and learning use float32.

Video-level sampling. For GenVideo and FakeParts, eight frames are sampled uniformly over the full decoded video. The detector therefore receives seven signed adjacent latent differences,

$$\delta_t = z_{t+1} - z_t, \qquad \boldsymbol{\delta} \in \mathbb{R}^{7\times 16\times 28\times 28}. \tag{5}$$

The same frozen GenVideo-selected checkpoint is transferred to FakeParts; no FakeParts labels are used for checkpoint or threshold selection.

Localization sampling. ActivityForensics uses the fixed 193-frame sequence specified by its protocol, which gives 192 transition tokens. The controlled benchmark uses the first 35 aligned frames at 7 FPS, giving 34 transition tokens. All methods compared through the same ActionFormer receive the same timestamps and temporal mask.

D.2 Exact Multi-Horizon DELTA Descriptor

Starting from adjacent differences δ_t, the exact signed displacement over horizon h is obtained by a prefix sum,

$$\delta_t^{(h)} = \sum_{j=0}^{h-1} \delta_{t+j} = z_{t+h} - z_t, \qquad \Delta_h z_t = \frac{\delta_t^{(h)}}{h}. \tag{6}$$

We use $h \in \{1, 2, 4, 8\}$. No shortened horizon or zero-padded latent is used. To retain one descriptor for every adjacent-frame timestamp, a horizon-h window is centered at each token and clipped to the nearest valid complete window. With N adjacent differences, its start index is

$$a_h(t) = \mathrm{clip}\left(t - \left\lfloor \frac{h-1}{2} \right\rfloor,\, 0,\, N - h\right), \tag{7}$$

and the feature at token t is $\Delta_h z_{a_h(t)}$. Thus only the nearest complete window is repeated at the two temporal edges.

For a spatial map $u \in \mathbb{R}^{H'\times W'}$, DELTA records

$$m(u) = \frac{1}{H'W'} \sum_p u_p, \qquad a(u) = \frac{1}{H'W'} \sum_p |u_p|, \tag{8}$$

$$r(u) = \sqrt{\frac{1}{H'W'} \sum_p u_p^2}, \qquad s(u) = \sqrt{\frac{1}{H'W'} \sum_p (u_p - m(u))^2}. \tag{9}$$

Concatenating signed mean, absolute mean, RMS, and spatial standard deviation over 16 channels and four horizons produces the reported $4 \times 16 \times 4 = 256$ dimensional descriptor at each timestamp.

D.3 FSQ–CNN Video Detector

The seven input difference maps are standardized channelwise using means and standard deviations estimated from at most 2,048 randomly selected training videos. No validation or test feature contributes to these statistics. A spatial encoder maps each $16 \times 28 \times 28$ difference map through

$$\begin{aligned} &\mathrm{Conv}_{3\times 3}(16, 96) \to \mathrm{GN}_8 \to \mathrm{GELU} \to \\ &\mathrm{Conv}_{3\times 3, s=2}(96, 96) \to \mathrm{GN}_8 \to \mathrm{GELU} \to \mathrm{Conv}_{1\times 1}(96, 16). \end{aligned} \tag{10}$$

The stride-2 encoder produces a $16 \times 14 \times 14$ map whose channels are split into four 4-D groups. We apply FSQ (Mentzer et al., 2024) with levels $L = (8, 5, 5, 5)$, corresponding to 1,000 implicit codes. FSQ uses no learned codebook or commitment loss. The finite cells provide a piecewise-constant representation of learned transition features. A straight-through gradient estimator enables the encoder to learn through quantization under the video classification loss. The resulting quantized representation is compact and supports discrimination between real and generated videos.

The quantized map is processed by a 3×3 convolution from 16 to 96 channels, eight-group normalization, GELU, global average pooling, and a $96 \to 160$ linear layer followed by layer normalization and GELU. Let $f_t \in \mathbb{R}^{160}$ be the resulting feature. We concatenate its temporal mean, population standard deviation, and maximum,

$$f_V = [\mathrm{mean}_t\, f_t; \mathrm{std}_t\, f_t; \max_t f_t] \in \mathbb{R}^{480}, \tag{11}$$

and classify it with $480 \to 160 \to 1$ linear layers, GELU, and dropout 0.25. The complete FSQ–CNN head contains 205,969 trainable parameters; the FSQ operation itself has no trainable parameter.

Training uses AdamW for at most 80 epochs with batch size 64, learning rate 3×10^{-4}, and weight decay 10^{-4}. The binary cross-entropy logit loss uses class weight $n_{\mathrm{real}}/n_{\mathrm{fake}}$ for the positive class and label smoothing 0.02. We retain the checkpoint with highest validation AUROC and stop after 15 validation epochs without improvement. AUROC is computed directly from logits; a validation-set Youden threshold is stored only for auxiliary accuracy reporting.

D.4 ActionFormer Localization Head

We use the public ActionFormer localizer (Zhang et al., 2022). On ActivityForensics, DELTA uses 256 input channels and the matched CLIP-L/14 control uses its native 768 channels; both use maximum sequence length 192, and shorter sequences are upsampled onto this canvas. The convolutional Transformer backbone has architecture $(2, 2, 5)$ and six temporal pyramid levels with regression ranges

$$[0, 4], [4, 8], [8, 16], [16, 32], [32, 64], [64, 10^4]. \tag{12}$$

The six local-attention windows are $(7, 7, 7, 7, 7, -1)$ with four attention heads; the final level uses global attention. Embedding, pyramid, and prediction head widths are 256. Classification and regression heads contain three 3-tap temporal layers. We use absolute positional encoding, identity FPN, pyramid scale factor 2, drop-path 0.1, and no auxiliary boundary or quality head.

The ActivityForensics localizer uses AdamW with learning rate 10^{-3}, weight decay 0.05, and batch size 32. Optimization runs for 30 cosine-schedule epochs after five linear warm-up epochs (35 total passes through the training loop). The gradient norm is clipped to 1.0; label smoothing is 0.1; the classification prior is 0.01; center sampling uses radius 1.5; the initial loss normalizer is 200; and regression uses distance-IoU loss with loss weight 2.0. DELTA and the CLIP control use the same localizer, optimizer, post-processing, and evaluator. Source-video bootstrap intervals use 1,000 resamples.

At inference, candidates below 10^{-3} are removed before soft NMS. We keep at most 2,000 pre-NMS candidates and 100 final segments, using soft-NMS IoU threshold 0.1, Gaussian sigma 0.75, minimum duration 10^{-3}, and voting threshold 0.9. ActivityForensics intra-domain AP is evaluated at tIoU $\{0.75, 0.85, 0.95\}$; open-world evaluation on Vidu uses the same frozen checkpoint and proposal rules.

For the controlled benchmark, the input dimension is 256, maximum sequence length is 576, maximum buffer factor is 6, and local-attention window is 19. The public default widths of 512 are used for the embedding, pyramid, and prediction heads. Training uses AdamW with batch size 2, learning rate 10^{-4}, weight decay 0.05, and 35 cosine epochs after five warm-up epochs. Its initial loss normalizer is 100; gradient clipping, classification prior, and center-sampling radius remain 1.0, 0.01, and 1.5. Soft NMS retains at most 2,000 pre-NMS and 200 final proposals, with score threshold 10^{-3}, IoU threshold 0.1, sigma 0.5, minimum duration 0.05, multiclass NMS, and voting threshold 0.7. We report frame AUROC and interval AP at tIoU $\{0.75, 0.85, 0.95\}$, whose equal-weight mean forms strict temporal mAP. DELTA and CLIP use the same schedule, proposal settings, and evaluator, and reported results are averaged over three independent runs.

D.5 Selection and Metrics

All model choices are made on the source validation split. Video-level AUROC is computed independently for each fake domain against its prescribed real pool, followed by an equal-weight macro average. Localization results are produced from saved proposals and the same evaluator at tIoU 0.75, 0.85, and 0.95.

E BOUNDARY-ALIGNED TRANSITION ANALYSIS

For the paired mechanism analysis, we define the standardized transition distance

$$D_t = \sqrt{\frac{1}{CH'W'} \sum_{c,h,w} \left(\frac{\Delta z^{\text{fake}}_{t,c,h,w} - \Delta z^{\text{real}}_{t,c,h,w}}{\sigma_c} \right)^2}, \tag{13}$$

where σ_c is estimated only from paired-real transitions. The analysis uses 600 fake/real validation pairs, with 200 pairs from each of Framer, RIFE, and SVD-XT. Generated intervals use inclusive frame boundaries; onset is the transition at $\mathrm{start}-1$, and offset is the transition at end. Curves cover ± 16 transitions around each boundary. Multiple boundaries are first averaged within a video, followed by 2,000 source-video bootstrap resamples. Raw L2 distance, cosine distance, and a Ledoit–Wolf channel-level Mahalanobis distance provide complementary checks.

F ENCODER REPLACEMENT

We replace Wan2.1 with the frozen official `zai-org/CogVideoX1.5-5B` VAE (Yang et al., 2025) while retaining the task-specific DELTA pipelines. Frames are again encoded independently with posterior mode at 224×224. For video-level detection, adjacent latent differences are processed by the FSQ quantizer and a CNN under uniform eight-frame sampling; the GenVideo-selected detector is evaluated on GenVideo (Table 9) and FakeParts (Table 10). For ActivityForensics, the localization pipeline retains the $h \in \{1, 2, 4, 8\}$ horizon-normalized differences and ActionFormer setup used in the main experiment; localization results are reported in Section H.2.

Table 9: Encoder comparison on GenVideo. Values are video-level AUROC (%) on GenVideo validation; Avg. is the equal-weight mean over the ten generators.

Encoder	MS	Morph	Moon	HotShot	Show-1	Gen-2	Crafter	LaVie	Sora	Wild	Avg. ↑
Wan2.1 (Wan et al., 2025)	93.72	**96.33**	**99.21**	98.24	97.05	**99.53**	98.60	95.73	**89.67**	**94.91**	**96.30**
CogVideoX1.5 (Yang et al., 2025)	**94.11**	96.20	99.08	**98.69**	**98.07**	98.67	**98.98**	**96.44**	82.90	94.66	95.78

Table 10: Encoder comparison on FakeParts. Each GenVideo-selected detector is evaluated without additional training. Values are video-level AUROC (%); Avg. is the equal-weight mean over the eight tasks.

Encoder	TI2V	T2V	Interp.	Extrap.	Inpaint.	Outpaint.	Face	Style	Avg. ↑
Wan2.1 (Wan et al., 2025)	77.24	84.43	77.63	53.31	**57.44**	94.93	**69.15**	**81.92**	74.51
CogVideoX1.5 (Yang et al., 2025)	**80.13**	**85.41**	**79.02**	**53.74**	56.67	**97.93**	61.82	81.52	**74.53**

G ROBUSTNESS DETAILS

The controlled GenVideo comparison in Table 3 decodes the source videos, resizes frames to 298×224 with Lanczos filtering, and encodes them as RGB JPEG-Q95 with 4:2:0 subsampling. All methods use the same video list, sampled frames, and real-reference subset.

H MECHANISM AND NEGATIVE RESULTS

H.1 RELATION TO EXISTING TEMPORAL-DIFFERENCE METHODS

D3. D3 (Zheng et al., 2025) operates on frame-level visual embeddings from pretrained encoders such as XCLIP. Given features f_t, it first reduces each adjacent-frame change to a scalar distance,

$$a_t = \|f_{t+1} - f_t\|_2, \tag{14}$$

and then analyzes the temporal variation of these distances. This early norm reduction discards signed, channel-specific changes. DELTA instead operates on generative encoder latents and retains signed transitions at each timestamp and across multiple temporal horizons.

DBINDS. DBINDS (Wu et al., 2026) uses Stable Diffusion v1.5 and diffusion inversion to obtain surrogate initial-noise descriptors ϵ_t, then constructs adjacent differences $\epsilon_{t+1} - \epsilon_t$ and analyzes them with multi-domain and multi-scale statistics. DELTA does not require diffusion inversion; it directly measures transitions in the generative encoder space,

$$\Delta_h z_t = \frac{z_{t+h} - z_t}{h}. \tag{15}$$

Thus, DELTA differs primarily in the representation being differenced and in the information retained: channel-specific signed changes, temporal position, and multiple temporal horizons.

H.2 Encoding

Wan2.1 and CogVideoX1.5 receive one frame per encoder call; temporal evidence comes from differencing their maps and processing the resulting sequence. We therefore also evaluate the image VAE from Stable Diffusion v1.5 (Rombach et al., 2022) with the same frame sampling, temporal horizons, four spatial statistics, ActionFormer, and evaluation protocol. Its four latent channels occupy the corresponding slots of the 256-dimensional DELTA layout, while unused channel slots are zero-padded.

Table 11: Frame-wise encoder comparison under the official ActivityForensics intra-domain and open-world protocols. Values are temporal AP and three-threshold mAP (%).

	Intra-domain				Open-World			
Encoder	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	mAP ↑	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	mAP ↑
Stable Diffusion v1.5 (Rombach et al., 2022)	86.03	79.61	56.99	74.21	86.49	79.60	57.97	74.69
CogVideoX1.5 (Yang et al., 2025)	89.80	84.91	61.46	78.72	89.65	84.96	64.70	79.77
Wan2.1 (Wan et al., 2025)	**91.37**	**87.69**	**73.06**	**84.04**	**92.13**	**88.72**	**78.08**	**86.31**

In Table 11, the image VAE outperforms ActionFormer intra-domain but is weaker open-world. Both video-generation encoders perform better than the image VAE, with Wan2.1 giving the strongest result. Thus, temporal differencing is useful beyond one encoder family, while the underlying generative representation materially affects localization quality.

H.3 Spatial-Statistic Ablation

Signed mean captures net directional change, absolute mean captures magnitude without sign cancellation, RMS emphasizes large responses, and standard deviation measures spatial dispersion.

The ablation retains each statistic alone and removes each one in turn. All nine variants use identical temporal gaps, data splits, and ActionFormer settings. Omitted statistic slots are set to zero in the 256-dimensional descriptor; retained values are unchanged. Table 12 reports the complete ablation.

Table 12: Spatial-statistic ablation under the official ActivityForensics intra-domain and open-world protocols. Values are temporal AP (%); mAP averages AP@0.75, AP@0.85, and AP@0.95. Bold/underline indicate best/second-best.

	Intra-domain				Open-World			
Statistics	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	mAP ↑	AP@0.75 ↑	AP@0.85 ↑	AP@0.95 ↑	mAP ↑
Signed mean only	82.28	76.64	54.88	71.27	81.83	76.46	53.76	70.68
Absolute mean only	85.53	78.97	47.80	70.77	85.14	80.54	52.44	72.71
RMS only	83.30	74.62	42.09	66.67	83.31	77.05	44.13	68.16
Spatial std. only	81.74	72.12	36.72	63.53	83.23	76.24	32.52	64.00
Without signed mean	87.83	82.41	55.07	75.10	88.79	85.01	59.84	77.88
Without absolute mean	90.10	86.60	71.51	82.74	91.58	88.25	74.22	84.68
Without RMS	<u>91.37</u>	87.34	<u>73.06</u>	83.92	91.08	87.98	76.91	85.32
Without spatial std.	91.08	**87.94**	**73.61**	**84.21**	<u>91.96</u>	**88.80**	<u>77.18</u>	<u>85.98</u>
Full	**91.37**	<u>87.69</u>	73.06	<u>84.04</u>	**92.13**	<u>88.72</u>	**78.08**	**86.31**

Combining statistics outperforms every individual statistic on both splits. Removing signed mean causes the largest decrease, 8.94/8.43 mAP points intra-domain/open-world, while removing absolute mean costs 1.30/1.63 points. The effects of removing RMS (0.12/0.99 points lower) or spatial standard deviation (0.17 points higher/0.33 points lower) are much smaller. These results support the joint use of directional and magnitude cues.

H.4 Scalar Transition Diagnostic

The scalar adjacent-latent diagnostic is useful for visualization but is not a universal detector. Under fixed-duration sampling, the equal-generator median shift between fake and real videos is -0.0153 on GenVideo and -0.0367 on FakeParts. Several individual generators and tasks remain close to zero, and Sora shifts in the opposite direction. After resize-before-JPEG preprocessing, the GenVideo shift becomes -0.1265, further showing that its magnitude depends on the input pipeline. These observations motivate retaining vector-valued latent differences instead of thresholding a single smoothness statistic. Besides, these observations also suggest that generated-video evidence is not captured by a universal shift in transition magnitude. This also motivates the quantized video-level representation: rather than relying on exact continuous amplitudes, FSQ groups nearby transition responses into discrete configurations from which recurring spatial transition patterns can be learned. Table 13 shows higher average AUROC for the FSQ-based detector on FakeParts, supporting the use of quantization for video-level detection.

Table 13: FakeParts video-level detection with and without FSQ. Values are AUROC (%); Avg. is the equal-weight mean over the eight tasks.

Configuration	TI2V	T2V	Interp.	Extrap.	Inpaint.	Outpaint.	Face	Style	Avg. ↑
FSQ	**77.24**	**84.43**	77.63	**53.31**	**57.44**	**94.93**	**69.15**	**81.92**	**74.51**
No-FSQ	72.36	82.74	**78.45**	46.23	56.43	93.20	57.06	77.95	70.55

I Efficiency and Resource Use

Table 14 reports measured resource use for the two ActionFormer pipelines on the same 100 ActivityForensics test videos (193 frames each) and one NVIDIA A100-SXM4-80GB GPU, using 10 warm-up videos and three timed repetitions. Video-to-feature time includes decoding, preprocessing, encoder inference, and descriptor construction; end-to-end latency adds ActionFormer inference. Model loading is excluded. Parameter counts include the encoder and ActionFormer, peak memory is maximum allocated GPU memory, and storage is the measured feature-array size.

Table 14: Efficiency and resource use under the ActivityForensics input protocol. Latency is measured in seconds per video and storage in MiB per video.

Method	Params (M) ↓	Video-to-feature ↓	Proposal head ↓	End-to-end ↓	Video/s ↑	Peak GPU (GiB) ↓	Storage (MiB) ↓
ActionFormer (Zhang et al., 2022)	435.08	8.71	0.82	9.53	0.105	6.49	0.566
DELTA	133.64	8.93	0.78	9.71	0.103	0.51	0.188

Wall-clock latency is similar because decoding and preprocessing dominate: these stages take 8.06 seconds per video for CLIP and 7.44 for DELTA, while encoder inference takes 0.65 and 1.42 seconds. DELTA reduces the total parameter count by 69.3%, peak allocated GPU memory by 92.1%, and stored feature size by 66.8%, with a 1.9% increase in end-to-end latency. Its practical efficiency advantage is therefore memory and storage rather than wall-clock speed.