MixForensics: Blend Before You Encode for Generalizable AI-Generated Video Detection

Ruiqi Liu[1,2,*] ZiAn Wang[3,*] Ruoxin Chen[4] Xuehai Bai[5] Manni Cui[6]
Ziheng Qin[1] Zhiyuan Yan[7] Shu Wu[1,†] Wenhao Wang[8,†]

[1]Institute of Automation, CAS [2]School of Advanced Interdisciplinary Sciences, UCAS
[3]Jilin University [4]Tencent YouTu Lab [5]Independent Researcher
[6]Huazhong University of Science and Technology [7]Peking University [8]Vast Intelligence Lab

*Equal contribution. †Corresponding authors.

Abstract

Detecting AI-generated videos (AIGV) in a generator-agnostic manner is increasingly important as generative models close the gap with real footage, and a video, with many frames, intuitively offers richer forensic evidence than a single image. Yet existing detectors fall short of this expectation: under the encode-then-fuse paradigm, increasing the frame budget from 1 to 8 improves a voting-based image detector by only 5.28 percentage points, and even temporal-modeling backbones such as TimeSformer barely improve on this. Per-frame predictions are also noisy: per-frame logits fluctuate substantially within a video and correlate weakly with any frame-level proxy. We trace the cause not to the aggregation step but to per-frame encoding severing multi-frame cues before they can interact, leaving any post-hoc selection, weighting, or aggregation strategy without a reliable basis. We therefore propose **MixForensics**, a blend-then-encode framework that combines multiple frames in the pixel domain into a single composite image (Stochastic Frame Blending, SFB) and regularizes the encoder for consistency across blended views (Blend Invariance Regularization, BIR). To stress-test this approach, we further introduce **MixForensics-Bench**, a test-only benchmark covering ten recent generators (including five closed-source commercial platforms) and two partial-forgery scenarios (temporal splicing and conditional continuation), with matched controls that isolate generative artifacts from editing artifacts. On AIGVDBench and MixForensics-Bench, MixForensics outperforms the strongest baseline by 1.27 and 8.11 percentage points on the fully-fake splits, leads by 11.49 points on temporal splicing and 9.15 points on conditional continuation, and uses fewer encoder forward passes than per-frame methods.

1 Introduction

Generative video models [1, 2, 3, 4] now produce outputs nearly indistinguishable from real footage, raising concerns about misinformation, identity fraud, and content abuse. Detecting AI-generated videos (AIGV) in a generator-agnostic manner has consequently become a central challenge in digital forensics [5, 6, 7]. Image-level forgery detection has progressed steadily [8, 9, 10, 11], and a video, composed of many frames, should expose even richer forensic cues [12, 13, 14]. Intuitively, detection accuracy should improve as more frames are examined.

However, existing AIGV detectors fall short of this expectation. Nearly all of them follow an encode-then-fuse paradigm: each sampled frame is encoded independently, and the resulting features or predictions are aggregated post-hoc. Within this paradigm two limitations stand out. First, additional frames yield diminishing returns: increasing the frame budget from 1 to 8 improves a voting-based image detector by only 5.28 percentage points while inference cost grows eightfold, and even temporal-modeling methods such as TimeSformer [15] barely do better (Figure 1). Second, per-frame

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

predictions are inherently noisy: within a single AIGV video, per-frame logits fluctuate substantially and correlate weakly with any frame-level proxy such as texture complexity, optical flow magnitude, or frame type (Figure 2). Together, these limitations mean that detectors cannot efficiently scale within this paradigm, whether by adding more frames or by selecting better frames.

These limitations prompt us to rethink how to genuinely exploit multi-frame forensic evidence. We argue that current methods fail because the forensic difficulty of a single frame cannot be quantified in advance: it can be predicted neither from frame-level properties such as texture complexity or optical flow magnitude, nor reliably read out from per-frame outputs, leaving any post-hoc selection or weighting strategy without a reliable basis. The problem is not in the aggregation step but in the fact that multi-frame cues are severed before they reach the encoder. A natural alternative is to blend multiple frames into a single composite image in the pixel domain before feeding them to the encoder, so that multi-frame evidence is fused at the very first stage of feature extraction.

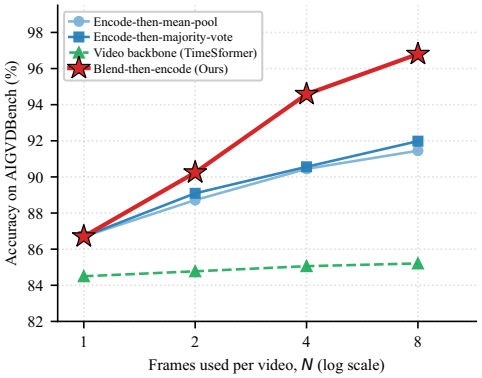


Figure 1: Accuracy vs. frames used per video on AIGVDBench. The two encode-then-X baselines, mean pool (averaging per-frame logits) and majority vote (per-frame label voting), cluster within 1 percentage point at $N = 8$; an in-encoder video backbone (TimeSformer [15]) sits below. Our blend-then-encode method dominates.

We propose MixForensics, a blend-then-encode framework with two components. Stochastic Frame Blending (SFB) selects a contiguous window of subsampled frames and combines them into a single composite image using weights drawn from a symmetric Dirichlet distribution, delivering multi-frame evidence to the encoder in a single forward pass. The composite image varies with the random window position and weights, but the underlying artifact signature does not. We therefore introduce Blend Invariance Regularization (BIR), which constrains feature consistency between two independently blended views of the same video, guiding the encoder to capture video-level forensic signatures rather than specific blending instances.

Existing AIGV benchmarks evaluate detectors almost exclusively on fully forged videos, leaving open the question of how aggregation-based methods behave when only part of a video is generated. We therefore construct MixForensics-Bench, a test-only benchmark covering ten recent generators (five closed-source commercial platforms) and two partial-forgery scenarios: temporal splicing and conditional continuation. Each scenario also includes matched controls that apply the same editing pipeline to purely real content, isolating generative artifacts from editing artifacts.

We evaluate MixForensics on AIGVDBench [7] and MixForensics-Bench. It outperforms the strongest baseline by 1.27 and 8.11 percentage points on the respective fully-fake splits, leads by 11.49 points on temporal splicing and 9.15 points on conditional continuation, and uses fewer encoder forward passes than per-frame methods.

Our contributions are threefold:

- Through controlled pipeline and matched-control experiments, we show that pre-encoding pixel blending exposes a cross-generator forensic signal that per-frame encoding loses and no post-hoc aggregation, selection, or weighting can recover.
- We propose MixForensics, a blend-then-encode framework that achieves competitive cross-generator generalization, up to 8.11 percentage points above the strongest baseline, while using fewer encoder forward passes than per-frame methods.
- We introduce MixForensics-Bench, a test-only benchmark covering ten recent generators (including five closed-source commercial platforms) and two partial-forgery scenarios (temporal splicing and conditional continuation), with matched controls for the partial-forgery scenarios.

2 Related Work

AI-generated image and video detection. Attentional cross-domain CNNs [16] fuse spatial and frequency cues for image manipulation localization. Image-level deepfake detectors exploit frequency-

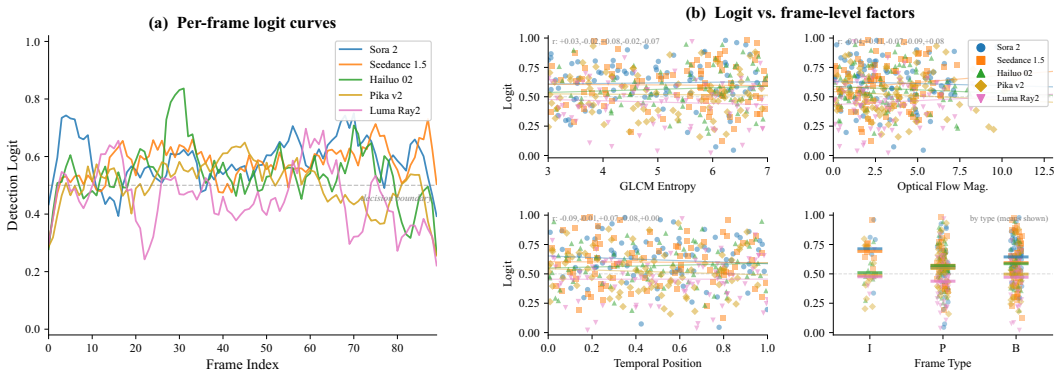


Figure 2: Frame-wise salience heterogeneity. (a) Per-frame detection logit curves within a single AIGV video fluctuate substantially across generators (dashed: decision boundary). (b) Detection logits versus texture complexity, optical flow magnitude, temporal position, and frame type. Pearson correlations for the three continuous variables satisfy $|r| < 0.12$.

domain cues [17, 18], reconstruction errors [19, 20], common forgery features [21], upsampling artifacts [10], or CLIP [22] features [9]. SynerDetect [23] combines perceptual and generative foundation-model cues through cross-model distillation and contrastive learning. Recent image detectors explore real-centric envelope modeling (REM [24]), real-only representation learning (RealNet [25]), multi-VAE manifold envelope alignment (M^2EA [26]), and manifold-consistent reference reconstruction (MIRROR [27]). DINO-Detect [28] distills a DINOv3 teacher trained on sharp images into a student trained on blurred counterparts to improve blur robustness. TrueMoE [29] uses dual routing over specialized discriminative experts to improve cross-generator generalization. These methods can be applied to videos via per-frame classification with majority voting, but their performance depends on which frames are sampled (Section 1). Video-level methods explicitly model temporal information through architectures such as SlowFast [30] and TimeSformer [15], yet they all follow the encode-then-fuse paradigm and still require per-frame encoding. Recent video-forensics methods such as DeMamba [5], DeCoF [6], ReSTrav [12], and NSG-VD [13] add temporal consistency cues or frequency-domain features for AIGV detection, but do not depart from this paradigm. V-PVP [31] improves video-backbone readouts by retaining patch-level temporal dynamics and inter-patch relations. Hermes [32] uses an evidence reasoning graph and multi-agent verification for explainable AI-generated video detection. Frame sampling [33, 34] and MIL [35] reduce cost but remain sensitive to sampling quality; MixForensics instead blends frames before encoding.

Domain generalization and benchmarks. UnionFormer [36] combines RGB, noise, and object-consistency cues for joint manipulation detection and localization. Cross-generator generalization has been pursued via frequency analysis [18], content-forgery disentanglement [6], and vision-language features [9, 37]; MixForensics is complementary to these approaches. Domain-specific resources such as MedForensics [38] extend synthetic-image detection to medical modalities. Existing benchmarks, including FaceForensics++ [39], Celeb-DF [40], GenVideo [5], GenVidBench [14], and AIGVDBench [7], primarily evaluate fully real or fully fake videos. MixForensics-Bench (Section 4) complements these benchmarks with partial-forgery scenarios and matched controls, while covering ten recent generators (2024–2026), including five closed-source commercial platforms.

3 Method

3.1 Framework

MixForensics consists of two components: Stochastic Frame Blending (SFB), which constructs the composite input, and Blend Invariance Regularization (BIR), which encourages the encoder to learn representations invariant to the blending process. Figure 3 illustrates the architecture.

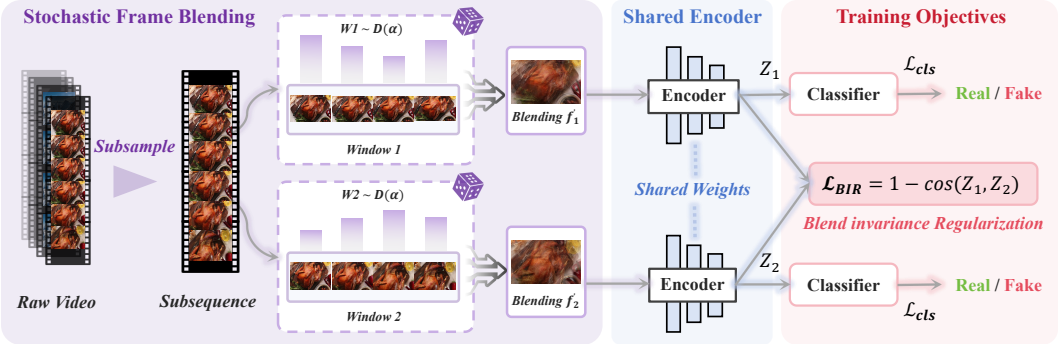


Figure 3: Overview of the MixForensics framework. **Left (SFB):** The video is uniformly subsampled to M candidate frames; a contiguous N-frame window is then drawn and blended with Dirichlet-distributed weights into a single composite image. **Middle:** The composite image is processed by an image encoder, and the CLS token feature is used for classification. **Right (BIR):** Two independently blended views from the same video are constrained to produce consistent features.

3.2 Stochastic Frame Blending (SFB)

Given an input video V of L frames, SFB constructs a single composite image in two stages: (i) uniform subsampling: V is sampled at fixed intervals to obtain M candidate frames spanning the full video; (ii) contiguous windowing: N contiguous frames $\{s_1, \ldots, s_N\}$ are selected from the candidate pool, with the starting position drawn uniformly at random during training and fixed at inference. Blending weights are sampled from a symmetric Dirichlet distribution $\mathbf{w} \sim \mathrm{Dir}(\alpha, \ldots, \alpha)$, and the composite image is

$$\hat{f} = \sum_{j=1}^{N} w_j \cdot s_j, \quad \sum_{j=1}^{N} w_j = 1, \tag{1}$$

where the concentration parameter α controls weight diversity: as $\alpha \to 0$, the window degenerates to a single random frame; as $\alpha \to \infty$, it approaches uniform blending. The composite $\hat{f}$ is resized, cropped, and normalized before being fed into an image encoder E; the CLS token of the final layer serves as the feature $\mathbf{z} = E(\hat{f})$, which is passed to a classification head for prediction.

Design rationale. The two-stage scheme decouples temporal coverage from artifact preservation: M-frame subsampling spans the full video, while a contiguous N-frame window keeps blended frames temporally local to favor the retention of shared artifact statistics. Dirichlet weights expose the model to diverse blending views during training.

3.3 Blend Invariance Regularization (BIR)

The stochasticity of SFB means that the same video produces different composite images across different forward passes. Despite the substantial frame-level variability in forensic salience, the artifacts imposed by a given generator on a given video still form a stable video-level forensic signature. Based on this observation, we introduce BIR as an auxiliary regularizer: if the features extracted by the encoder truly reflect video-level stable signals, then different blends of the same video should yield consistent feature representations.

Concretely, for the same video V, two starting positions are drawn independently from the candidate pool, and SFB is applied separately to each resulting window to produce two composite images $\hat{f}_1$ and $\hat{f}_2$ (with independently sampled window positions and Dirichlet weights). The corresponding features $\mathbf{z}_1 = E(\hat{f}_1)$ and $\mathbf{z}_2 = E(\hat{f}_2)$ are extracted, and the BIR loss is defined as

$$\mathcal{L}_{\mathrm{BIR}} = 1 - \frac{\mathbf{z}_1 \cdot \mathbf{z}_2}{\|\mathbf{z}_1\| \, \|\mathbf{z}_2\|}. \tag{2}$$

Minimizing $\mathcal{L}_{\mathrm{BIR}}$ encourages the encoder to learn feature representations that are invariant to the blending process; empirically this raises intra-video feature consistency from 0.71 to 0.93 and the inter-class margin from 0.42 to 0.58 (Section G). We compare alternative consistency objectives (symmetric KL divergence on logits, cosine similarity on a projection head) in the ablation (Section 5.4).

3.4 Training Objective

The total training loss combines the classification loss and the BIR loss:

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(\hat{f}_1) + \mathcal{L}_{\text{cls}}(\hat{f}_2) + \lambda \cdot \mathcal{L}_{\text{BIR}}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ is the binary cross-entropy loss and λ is a balancing coefficient. Both composite images participate in classification training, maximizing the utilization of training signal.

3.5 Inference

At inference, the M candidate frames obtained by uniform subsampling are partitioned into M/N non-overlapping contiguous N-frame windows, so every sampled candidate frame contributes to the decision. With our default $M{=}32$ and $N{=}8$, this gives four windows per video. Each window is blended with uniform weights into a composite image and classified in one forward pass; per-window scores are aggregated by (i) mean pooling, the default for fully forged videos, or (ii) any-window-fake, which flags the video whenever any window score exceeds a threshold τ and is preferred for partial-forgery scenarios (Section 5.3). Inference therefore costs M/N forward passes per video, an N-fold reduction over per-frame methods at the same frame budget. Uniform weights at inference equal the expected weights of the symmetric Dirichlet distribution used during training.

4 MixForensics-Bench

We construct **MixForensics-Bench**, a test-only benchmark totaling 8,000 videos from ten recent generators and two partial-forgery scenarios (Figure 4).

Generators and real data. The benchmark covers five closed-source generators (Sora 2 [41], Seedance 1.5 [42], Hailuo 02 [43], Pika v2 [44], Luma Ray2 [45]) and five open-source generators (Wan 2.2 [3], HunyuanVideo 1.5 [4], LTX-2 [46], CogVideoX1.5 [47], SkyReels-V2 [48]), all dating from 2024–2026 and spanning diverse architectures (DiT [49], MoE-DiT [3], Expert Transformer [47], Diffusion Forcing [50], among others). Real videos (2,000) are sampled from OpenVid-HD [51] with SHA-256 deduplication against the training set. All videos are standardized to 720p and 24 fps. Real and generated base clips are trimmed to 3 s (72 frames); conditional-continuation samples retain the length resulting from concatenation.

Evaluation splits. Each generator contributes 100 fully fake videos (1,000 total). The temporal splicing split (1,500 videos) replaces a contiguous segment with generated content at forgery ratios $\{10\%, 30\%, 50\%, 70\%, 90\%\}$; the conditional continuation split (1,000 videos) extends real clips via I2V generation from all ten generators. Both partial-forgery splits include matched real-real controls (1,500 + 1,000), constructed by applying the same editing pipeline to purely real source clips, so that the effect of generated content is isolated from editing artifacts. Each split uses its own negative class: the fully fake split uses the 2,000 unedited real videos, while the partial-forgery splits use their respective real-real controls.

Design rationale. MixForensics-Bench is test-only, with 9 of the 10 generators unseen during training; CogVideoX1.5 is used in T2V mode here and in V2V mode in the training set. Matched real-real controls isolate generative signals from editing artifacts by ensuring that compression, resolution changes, and boundary discontinuities are present in both classes. The inclusion of five closed-source commercial platforms ensures realistic evaluation conditions beyond the open-source ecosystem. The training protocol is described in Section 5; Section A provides full construction details.

5 Experiments

5.1 Experimental Setup

Training and evaluation. Our training set uses three generators as fake sources: OpenSora [52], CogVideoX1.5 [47] (V2V), and EasyAnimate [53]. We evaluate on two benchmarks: MixForensics-Bench (Section 4, 10 generators) and the open-source split of AIGVDBench [7] (20 generators

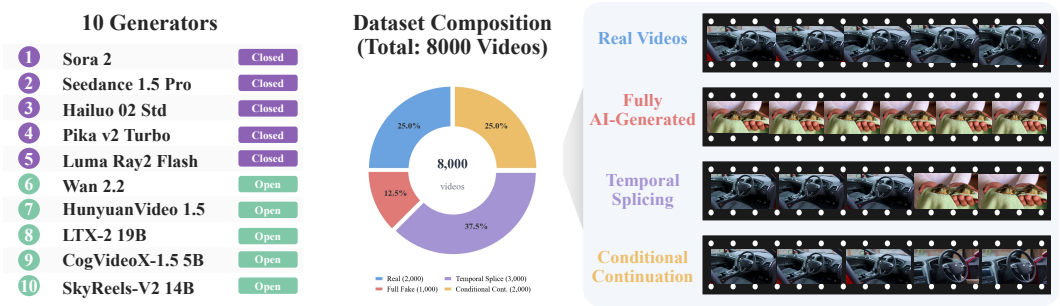


Figure 4: Dataset composition of MixForensics-Bench (8,000 videos): 2,000 real (OpenVid-HD [51]), 1,000 fully fake from 10 generators (9 unseen plus CogVideoX1.5 T2V; 5 closed-source), 1,500 temporal splicing at five forgery ratios, and 1,000 conditional continuation. Both partial-forgery splits include matched real-real controls (1,500 + 1,000). All videos use 720p / 24 fps; base clips are 3 s, while continuation samples retain their concatenated length.

spanning T2V, I2V, and V2V). AIGVDBench uses a held-out test split that includes both generators represented in training and unseen generators; three generator names overlap with the training set. On MixForensics-Bench, only CogVideoX1.5 T2V shares a generator name with the training set; the remaining nine generators are entirely unseen. Further details on benchmark selection and split rationale are provided in Section B. Balanced accuracy (%), denoted ACC in the figures and tables, serves as the primary metric.

Implementation. We adopt CLIP ViT-L/14 [22] as the image encoder, fine-tuned with LoRA [54] (rank$=8$, $\alpha=16$, applied to all 24 Transformer layers). The classification head is a single linear layer. Default hyperparameters are: blending window size $N=8$, Dirichlet concentration $\alpha=0.5$, and BIR weight $\lambda=0.1$. At inference, each video is uniformly subsampled to $M=32$ frames and partitioned into four non-overlapping $N=8$-frame windows, each producing one composite image for classification (Section 3.5). We train with AdamW (lr $=2\times10^{-5}$, weight decay 10^{-4}) for 20 epochs using a cosine learning rate schedule on $4\times$A800 GPUs. All reported results use seed 42.

Baselines. We compare against eleven methods from three categories: video-classification models with general temporal modeling (TimeSformer [15] and SlowFast [30]), image-level detectors that perform per-frame classification with majority vote (NPR [10], UnivFD [9], Effort [11], DDA [55], and MIRROR [27]), and video-forensics methods designed for AI-generated video detection (DeMamba [5], DeCoF [6], ReSTrav [12], and NSG-VD [13]). For the two main benchmarks, all methods are trained on the same three-generator training set and follow the unified evaluation protocol of AIGVDBench [7], retaining their original architectures and author-recommended hyperparameters. The additional GenVidBench-143k evaluation uses its own training split (Section C). At inference, image-level methods apply per-frame classification with majority vote, while video-forensics and video-classification methods use their respective temporal heads.

5.2 Cross-Generator Generalization

AIGVDBench. Table 1 reports results on the open-source split of AIGVDBench [7] (20 generators; 3,000 test videos each). MixForensics with mean pooling achieves 96.80% versus 95.53% for DeMamba, a 1.27-point gain. Any-window-fake raises accuracy to 98.63%. The per-generator results show gains over per-frame methods on realistic outputs such as HunyuanVideo and Wan 2.1.

MixForensics-Bench. Table 2 reports results on the MixForensics-Bench fully fake split (10 generators; 9 entirely unseen and CogVideoX1.5 T2V which shares its name with the V2V model in our training set). MixForensics achieves the highest average ACC across all ten generators. Video-forensics baselines (DeMamba [5], DeCoF [6], ReSTrav [12], NSG-VD [13]) and video-classification baselines (TimeSformer [15], SlowFast [30]) partially mitigate this issue through temporal modeling, but still lag behind MixForensics, suggesting that pre-encoding blending captures complementary information that post-encoding fusion misses. Mean pooling yields 70.60% and any-window-fake yields 69.14%. A single noisy window on a real video can trigger a false alarm under the any-

Table 1: ACC (%) on the open-source split of AIGVDBench, covering 20 generators across 12 T2V, 6 I2V, and 2 V2V tasks (3,000 test videos per generator). Three names overlap with the training set (OpenSora, CogVideoX, EasyAnimate); results include both seen-generator and cross-generator evaluation on held-out videos. Best average **bolded**; second underlined.

	I2V						T2V											V2V		
Method	Easy Animate	LTX	Pyramid Flow	SEINE	SVD	Video Crafter	Acc Video	Animate Diff	CogVideo X1.5	Easy Animate	Hunyuan Video	IPOC	LTX	Open Sora	Pyramid Flow	Rep Video	Video Crafter	Wan 2.1	CogVideo X	LTX	**Avg.**
Video-classification																				
TimeSformer	95.23	84.53	99.38	98.93	99.95	75.38	62.60	61.32	86.72	96.07	69.55	86.75	96.50	100.00	99.47	72.90	94.72	56.37	94.25	73.65	85.21
SlowFast	89.73	79.82	96.15	93.47	97.23	82.67	73.85	72.43	83.18	90.52	79.37	82.63	91.28	99.83	96.32	80.45	88.17	68.93	89.42	68.53	85.20
Image-level (per-frame + majority vote)																				
NPR	79.37	78.27	79.70	79.65	79.68	76.48	57.07	76.05	78.20	77.65	59.87	78.78	78.55	79.72	79.68	74.68	79.03	63.52	79.52	78.55	75.70
UnivFD	72.53	74.37	77.62	82.03	83.22	72.15	74.70	81.78	82.65	80.38	74.63	83.77	81.82	84.22	83.80	79.62	83.90	75.88	81.95	73.87	79.24
Effort	73.33	73.52	76.65	79.78	81.22	71.75	73.77	80.22	80.85	79.83	74.40	81.88	80.32	82.20	81.88	79.68	82.00	75.95	80.17	72.52	78.10
DDA	81.23	86.65	76.42	89.65	89.67	89.58	84.48	88.45	82.28	79.87	83.45	88.90	86.77	89.33	78.68	84.05	89.63	87.78	80.93	84.93	85.14
MIRROR	49.97	49.95	49.97	85.97	70.05	94.92	50.35	61.52	50.10	50.23	50.20	50.53	50.05	50.12	49.98	50.17	95.32	50.55	50.03	50.03	58.00
Video-forensics																				
DeMamba	100.00	98.77	100.00	100.00	100.00	98.87	69.22	88.76	98.21	99.62	76.62	99.06	99.52	100.00	100.00	93.36	98.52	92.61	100.00	97.42	95.53
DeCoF	82.50	82.25	88.42	90.75	92.72	57.18	79.07	80.83	91.98	84.60	76.00	93.18	91.32	93.33	93.43	90.42	79.90	73.98	90.90	81.93	84.73
ReSTrav	60.17	68.40	64.52	55.02	54.97	73.73	80.25	88.00	61.57	66.13	79.32	68.67	60.82	79.60	67.65	63.85	80.38	60.98	78.88	72.55	69.27
NSG-VD	50.73	53.66	51.85	52.03	50.34	53.41	56.33	52.48	52.18	49.68	48.88	52.21	52.96	59.33	50.78	51.91	48.22	48.43	53.57	53.42	52.12
Ours (mean)	100.00	98.97	100.00	99.98	100.00	99.80	83.05	92.13	99.87	99.98	79.28	99.98	98.50	100.00	100.00	98.53	99.90	87.00	99.98	99.07	96.80
Ours (any-win.)	100.00	99.85	100.00	100.00	100.00	99.87	91.62	96.67	99.95	100.00	90.17	99.97	99.88	100.00	100.00	99.80	99.95	94.93	100.00	99.87	**98.63**

Table 2: Cross-generator ACC (%) on the MixForensics-Bench fully fake split: 10 generators (5 closed-source, 5 open-source) with 100 fake videos each, evaluated against 2,000 real videos from OpenVid-HD [51]. 9 generators are entirely unseen; CogVideoX1.5 T2V shares its name with the V2V model in our training set. Best result **bolded**; second underlined.

	Closed-Source					Open-Source					
Method	Sora 2	Seedance 1.5	Hailuo 02	Pika v2	Luma Ray2	Wan 2.2	HunyuanVideo 1.5	LTX-2	CogVideoX1.5	SkyReels-V2	**Avg.**
Video-classification											
TimeSformer	63.75	41.27	47.60	51.88	74.20	46.52	40.95	65.42	65.78	69.22	56.66
SlowFast	53.73	50.88	52.35	51.88	52.27	55.27	54.75	51.83	53.73	54.25	53.09
Image-level (per-frame + majority vote)											
NPR	54.23	47.65	53.90	51.38	56.08	46.83	52.15	54.75	53.60	52.40	52.30
UnivFD	59.83	59.15	57.35	56.67	56.10	57.80	59.25	58.42	59.92	58.33	58.28
Effort	78.47	43.17	62.92	53.70	75.25	63.50	43.83	62.35	72.58	69.10	62.49
DDA	41.67	55.15	53.80	41.92	57.33	49.60	48.25	59.85	54.48	56.08	51.81
MIRROR	50.27	50.85	50.15	50.67	50.40	50.02	52.75	50.38	53.90	50.58	51.00
Video-forensics											
DeMamba	78.58	52.10	56.88	48.35	69.72	57.50	51.65	63.27	59.83	56.15	59.40
DeCoF	52.52	50.80	55.90	50.77	52.02	52.73	50.77	51.77	51.62	52.05	52.10
ReSTrav	59.62	44.10	70.53	71.88	47.70	58.12	30.48	59.73	68.03	51.33	56.15
NSG-VD	51.25	48.58	52.90	43.17	53.65	51.48	58.40	53.75	57.10	53.95	52.42
Ours (mean)	88.17	62.75	71.30	61.67	67.85	59.12	64.58	71.38	81.95	77.20	**70.60**
Ours (any-win.)	80.62	61.27	68.17	62.83	68.05	58.70	67.33	71.92	78.40	74.10	69.14

window rule, making mean pooling preferable on this split. We use any-window-fake by default for partial-forgery scenarios and report both rules for comparison.

Why most baselines collapse on MixForensics-Bench. The results show a substantial generalization gap: DeMamba [5] drops from 95.53% on AIGVDBench to 59.40% on MixForensics-Bench. Several methods score near chance (51–53%), with particularly poor results on Seedance 1.5 and Pika v2. Effort [11], the strongest baseline, reaches 62.49%, versus our 70.60%, an 8.11-point gain. These results suggest that pre-encoding blending captures cues that transfer better to unseen generator families.

5.3 Partial-Forgery Detection

Tables 3 and 4 report accuracy on the temporal splicing and conditional continuation splits. Partial forgery is inherently more difficult: forged frames constitute only a fraction of the video, so both logit-level and pixel-level averaging dilute the forensic signal, driving all methods toward chance performance at low forgery ratios (Section D).

Under mean pooling, MixForensics outperforms all baselines across forgery ratios, with the margin widening at higher ratios where blending windows sample predominantly fake frames. The any-window-fake rule flags a video whenever a single window exceeds the decision threshold; baselines retain their original aggregation strategies (Section 5.1). Conditional continuation proves harder than temporal splicing owing to smoother I2V transitions. Both splits employ matched real-real controls as the negative class, where real videos are processed through the identical splicing or continuation

Table 3: ACC (%) on the temporal splicing split of MixForensics-Bench: 1,500 spliced videos (10 generators across five forgery ratios $\{10\%, 30\%, 50\%, 70\%, 90\%\}$), evaluated against 1,500 matched real-real controls. All methods use the same training set.

Method	CogVideoX1.5	Hailuo 02	HunyuanVideo 1.5	LTX-2	Luma Ray2	Pika v2	Seedance 1.5	SkyReels-V2	Sora 2	Wan 2.2	**Avg.**
Video-classification											
TimeSformer	56.23	52.30	67.90	53.90	54.83	52.47	52.53	54.97	61.50	52.70	55.93
SlowFast	58.53	52.43	52.53	55.10	55.30	52.80	52.63	57.20	55.20	55.80	54.75
Image-level (per-frame + majority vote)											
NPR	54.80	52.40	51.70	54.13	52.13	49.80	51.90	51.13	53.13	45.40	51.65
UnivFD	57.63	56.80	57.90	57.10	56.30	57.30	56.57	55.30	57.50	57.00	56.94
Effort	54.83	54.60	54.17	52.57	53.60	53.83	54.33	53.40	54.00	53.17	53.85
DDA	53.20	52.60	56.20	58.87	57.93	48.60	52.20	55.60	46.93	46.87	52.90
MIRROR	52.03	51.53	51.20	52.73	52.00	52.47	52.53	52.20	50.80	51.37	51.89
Video-forensics											
DeMamba	54.50	52.40	52.73	53.50	59.23	52.23	53.07	53.40	55.50	53.23	53.98
DeCoF	53.17	53.07	52.80	52.17	53.40	53.47	54.20	53.53	53.40	53.33	53.25
ReSTrav	52.67	54.50	52.00	53.10	52.20	52.37	51.50	52.87	52.63	53.77	52.76
NSG-VD	57.50	51.53	55.20	49.50	52.23	48.17	51.20	58.17	55.90	57.10	53.65
Ours (mean)	64.53	63.67	73.50	62.73	63.47	62.30	72.27	67.03	72.43	66.07	<u>66.80</u>
Ours (any-win.)	66.40	66.50	72.00	64.43	65.60	64.50	74.50	68.30	74.80	67.27	**68.43**

Table 4: ACC (%) on the conditional continuation split of MixForensics-Bench: 1,000 I2V-extended videos (10 generators, 100 each) evaluated against 1,000 matched real-real controls. All methods use the same training set.

Method	CogVideoX1.5	Hailuo 02	HunyuanVideo 1.5	LTX-2	Luma Ray2	Pika v2	Seedance 1.5	SkyReels-V2	Sora 2	Wan 2.2	**Avg.**
Video-classification											
TimeSformer	55.35	55.85	51.85	53.85	53.35	51.35	53.85	55.85	51.85	55.85	53.90
SlowFast	56.15	55.15	58.15	56.65	56.15	58.15	55.15	55.15	57.15	60.65	56.85
Image-level (per-frame + majority vote)											
NPR	51.05	49.55	50.05	53.55	49.55	50.55	54.55	51.55	53.05	53.05	51.65
UnivFD	49.10	52.60	54.60	52.60	54.60	51.60	53.10	51.10	52.60	48.60	52.05
Effort	53.15	54.15	56.15	51.15	54.65	56.65	55.15	53.15	53.15	54.15	54.15
DDA	52.85	49.85	51.85	50.35	53.35	50.85	52.85	50.85	50.85	53.35	51.70
MIRROR	51.65	51.65	53.15	48.65	53.65	51.15	54.15	52.65	49.65	52.15	51.85
Video-forensics											
DeMamba	54.90	57.40	55.90	58.40	58.40	55.40	57.40	58.40	57.90	57.90	57.20
DeCoF	51.20	54.20	53.70	54.20	53.20	56.20	53.70	50.20	51.70	52.70	53.10
ReSTrav	50.30	48.30	52.30	49.80	50.80	50.30	47.80	49.80	51.30	49.30	50.00
NSG-VD	54.40	55.90	56.40	52.90	56.40	52.90	54.40	56.40	56.40	55.90	55.20
Ours (mean)	68.85	66.35	64.35	66.35	62.35	67.85	68.35	67.35	62.85	64.35	<u>65.90</u>
Ours (any-win.)	63.55	61.55	67.05	67.05	66.05	66.55	67.05	66.05	69.05	69.55	**66.35**

pipeline with purely authentic content, so the reported accuracy isolates generative artifacts from editing artifacts. Were the model exploiting boundary or blending cues rather than generator-specific signals, it would misclassify these controls as fake, depressing accuracy.

Baseline behavior under partial forgery. On the splicing split, video-classification models (TimeSformer [15], SlowFast [30]) and UnivFD [9] achieve the highest baseline accuracy (54–58% on average). On the continuation split, DeMamba [5] (57.20%) and SlowFast [30] (56.85%) achieve the highest baseline accuracy. No single baseline consistently leads across both scenarios, reflecting the difficulty of partial-forgery detection. MixForensics surpasses every baseline on average on both splits, with the gap widening further under the any-window-fake rule, where a single window score exceeding the threshold suffices to flag the video.

5.4 Ablation and Analysis

Component ablation. Table 5 ablates the key design choices on the MixForensics-Bench fully fake split. The left column varies the window size N; the right column reports accuracy when individual components are removed: "w/o Dirichlet" uses uniform weights during training, "w/o BIR" disables the BIR loss, and "w/o contiguous" replaces contiguous window sampling with random sampling within the candidate pool. Accuracy peaks at $N{=}8$: increasing N from 2 to 8 raises accuracy by 5.50 percentage points, while $N{=}16$ is 0.27 points below the default. Removing Dirichlet weights, BIR, or contiguous sampling reduces accuracy by 3.13, 5.85, and 2.20 points, respectively. Detailed sweeps of α and λ are in Section F; pipeline variant isolation appears in Section H.

Table 5: Component ablation (seed 42). Default ($N{=}8$): 70.60%; arrows show drop.

N	ACC	w/o	ACC
2	65.10 ↓5.50	Dirichlet	67.47 ↓3.13
4	67.35 ↓3.25	BIR	64.75 ↓5.85
16	70.33 ↓0.27	contiguous	68.40 ↓2.20

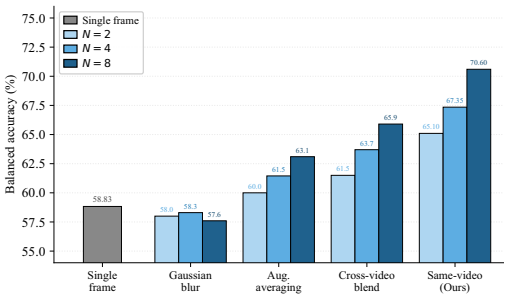


Figure 5: Pixel blending vs. matched controls on MixForensics-Bench (fully fake; seed 42; balanced ACC %). Same-video blending outperforms Gaussian blur, augmentation averaging, and cross-video blending.

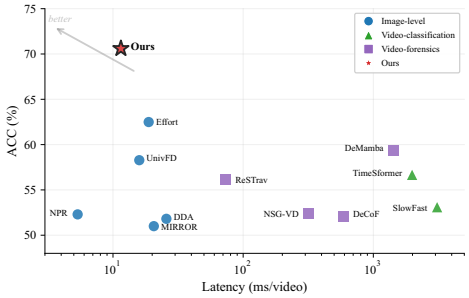


Figure 6: Inference efficiency ($N = 8$). MixForensics achieves the best ACC at low latency on a single A800 GPU.

Blending controls. We compare same-video pixel blending with a single-frame baseline and three matched controls (Figure 5): Gaussian blur with kernel size matched to the blending strength, augmentation averaging over multiple augmented views of a single frame, and cross-video blending with frames drawn from different fake videos. Gaussian blur degrades performance relative to the single-frame baseline, indicating that generic smoothing alone does not explain the blending gains. Same-video blending also outperforms cross-video blending, consistent with a benefit from retaining recurring within-video forensic cues.

Inference efficiency. Figure 6 compares inference cost. Each window of $N{=}8$ frames is processed in a single encoder forward pass; with $M{=}32$ frames partitioned into four non-overlapping windows, MixForensics requires only four forward passes per video, an $8\times$ reduction relative to per-frame methods that require 32 passes on the same frame budget. Pixel-level blending adds negligible overhead (<1 ms per window). Compared to video-level baselines (TimeSformer [15], SlowFast [30]), which require per-frame encoding plus additional temporal modeling, MixForensics achieves comparable or higher accuracy at a fraction of the computational cost.

6 Conclusion

MixForensics combines frames before encoding through SFB and aligns blended-view representations through BIR. Controlled comparisons support the benefit of pre-encoding fusion for generalizable AI-generated video detection. On the fully-fake benchmarks, the method exceeds the strongest baseline by up to 8.11 percentage points while using fewer encoder forward passes than per-frame methods. MixForensics-Bench adds temporal-splicing and conditional-continuation tests with matched controls; the any-window-fake rule improves sensitivity to partial forgeries.

Limitations. SFB discards temporal order information, which may limit performance on tasks that require motion-level analysis. When the forged-frame ratio is low, blending windows may contain predominantly real frames, diluting the forgery signal and reducing detection sensitivity. Heavy video compression can further degrade the fine-grained artifacts that blending aims to retain. Future work could explore order-aware blending strategies and adaptive window selection to address these limitations. Extending the framework to handle longer videos and investigating its interaction with diverse compression pipelines are also promising directions. We hope that MixForensics-Bench, with its partial-forgery scenarios and matched controls, will serve as a useful testbed for future research in video forensics.

Broader impacts. AI-generated video detection can support efforts against misinformation and content abuse. Missed detections can allow synthetic content to pass as authentic, while false positives can wrongly flag authentic footage, causing reputational harm or erroneous moderation.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62372454). The authors declare no competing interests.

References

[1] OpenAI. Sora. `https://openai.com/index/sora`, 2024.

[2] Kwai. Kling. `https://kling.kuaishou.com/`, 2024.

[3] 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.

[4] Bing Wu, Chang Zou, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Jack Peng, Jianbing Wu, Jiangfeng Xiong, Jie Jiang, et al. HunyuanVideo 1.5 technical report. *arXiv preprint arXiv:2511.18870*, 2025.

[5] 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. *arXiv preprint arXiv:2405.19707*, 2024.

[6] Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting AI-generated video via frame consistency. *arXiv preprint arXiv:2402.02085*, 2024.

[7] Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for AI-generated video detection. *arXiv preprint arXiv:2601.11035*, 2026.

[8] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8695–8704, 2020.

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

[10] 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 (CVPR)*, pages 28130–28139, 2024.

[11] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024.

[12] Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. AI-generated video detection via perceptual straightening. In *Advances in Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=LsmUgStXby`.

[13] Shuhai Zhang, Zihao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for AI-generated video detection. In *Advances in Neural Information Processing Systems*, 2025.

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

[15] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021.

[16] Shuaibo Li, Shibiao Xu, Wei Ma, and Qiu Zong. Image manipulation localization using attentional cross-domain CNN features. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5614–5628, 2023.

[17] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision (ECCV)*, pages 86–103. Springer, 2020.

[18] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[19] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4113–4122, 2022.

[20] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22445–22455, 2023.

[21] Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. UCF: Uncovering common features for generalizable deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22412–22423, 2023.

[22] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021.

[23] Shuaibo Li, Yijun Yang, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, Qing Zhang, and Lei Zhu. SynerDetect: Hierarchical synergistic learning for generalizable AI-generated image detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 6405–6414, 2026.

[24] Ruiqi Liu, Yi Han, Zhengbo Zhang, Liwei Yao, Zhiyuan Yan, Jialiang Shen, ZhiJin Chen, Manni Cui, Boyi Sun, Lubin Weng, et al. Beyond artifacts: Real-centric envelope modeling for reliable AI-generated image detection. In *European Conference on Computer Vision (ECCV)*, pages 225–243. Springer Nature Switzerland, 2026.

[25] Shuaibo Li, Laixin Zhang, Wei Ma, Jianwei Guo, Shibiao Xu, Zhijie Qiu, and Hongbin Zha. RealNet: Efficient and unsupervised detection of AI-generated images via real-only representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 38889–38898, 2026.

[26] Ruiqi Liu, Yi Han, Boyi Sun, Houxin He, Yuxuan Zhou, and Yan Wang. M^2EA: Multi-VAE manifold envelope alignment for challenging AI-generated image detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9147–9151. IEEE, 2026.

[27] Ruiqi Liu, Manni Cui, Ziheng Qin, Zhiyuan Yan, Ruoxin Chen, Yi Han, Zhiheng Li, Junkai Chen, ZhiJin Chen, Kaiqing Lin, et al. MIRROR: Manifold ideal reference ReconstructOR for generalizable AI-generated image detection. *arXiv preprint arXiv:2602.02222*, 2026.

[28] Jialiang Shen, Jiyang Zheng, Yunqi Xue, Huajie Chen, Yu Yao, Hui Kang, Ruiqi Liu, Helin Gong, Yang Yang, Dadong Wang, et al. DINO-Detect: A simple yet effective framework for blur-robust AI-generated image detection. *arXiv preprint arXiv:2511.12511*, 2025.

[29] Laixin Zhang, Shuaibo Li, Wei Ma, and Hongbin Zha. TrueMoE: Dual-routing mixture of discriminative experts for synthetic image detection. *arXiv preprint arXiv:2509.15741*, 2025.

[30] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019.

[31] Manni Cui, Ziheng Qin, ZiAn Wang, Ruiqi Liu, Dianyuan Zou, Jianglan Wei, Han Zhou, Yu Liu, Jingrui Xu, Wenhao Wang, et al. Rethinking the readout: Unlocking video backbones for AI-generated video detection. *arXiv preprint arXiv:2607.15321*, 2026.

[32] Shuaibo Li, Pengfei Hao, Hongtao Wu, Jianfeng Dong, Ping Li, Xiaohong Liu, and Lei Zhu. Hermes: An evidence-driven agentic framework for trustworthy and explainable AI-generated video detection. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL `https://openreview.net/forum?id=mgtSv2NUik`.

[33] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2016.

[34] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7083–7093, 2019.

[35] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6479–6488, 2018.

[36] Shuaibo Li, Wei Ma, Jianwei Guo, Shibiao Xu, Benchong Li, and Xiaopeng Zhang. UnionFormer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12523–12533, 2024.

[37] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 1006–1015, 2024.

[38] Shuaibo Li, Zhaohu Xing, Hongqiu Wang, Pengfei Hao, Xingyu Li, Zekai Liu, and Lei Zhu. Toward medical deepfake detection: A comprehensive dataset and novel method. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 626–637. Springer Nature Switzerland, 2026.

[39] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1–11, 2019.

[40] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3207–3216, 2020.

[41] OpenAI. Sora2. `https://openai.com/index/sora-2`, 2025.

[42] ByteDance. Seedance 1.5 Pro. `https://seed.bytedance.com/en/seedance1_5_pro`, 2025.

[43] MiniMax. Hailuo AI. `https://hailuoai.video`, 2025.

[44] Pika Labs. Pika. `https://pika.art`, 2025.

[45] Luma AI. Luma Ray2. `https://lumalabs.ai/ray2`, 2025.

[46] Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

[47] 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. *arXiv preprint arXiv:2408.06072*, 2024.

[48] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. SkyReels-V2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

[49] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.

[50] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pages 24081–24125, 2024.

[51] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. OpenVid-1M: A large-scale high-quality dataset for text-to-video generation. *arXiv preprint arXiv:2407.02371*, 2024.

[52] Zangwei Zheng, Xiangyu Peng, Yuxuan Lou, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, et al. Open-Sora 2.0: Training a commercial-level video generation model in $200k. *arXiv preprint arXiv:2503.09642*, 2025.

[53] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. EasyAnimate: A high-performance long video generation method based on transformer architecture. *arXiv preprint arXiv:2405.18991*, 2024.

[54] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[55] Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes AI-generated image detector easier generalizable. In *Advances in Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=C39ShJwtD5`.

[56] OpenAI. GPT-5. `https://openai.com/index/introducing-gpt-5`, 2025.

[57] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7310–7320, 2024.

[58] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. ModelScope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023.

[59] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2Video-Zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, 2023.

[60] Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pages 65618–65642, 2024.

[61] Zhiqiang Xia, Zhaokang Chen, Bin Wu, Chao Li, Kwok-Wai Hung, Chao Zhan, Yingjie He, and Wenjiang Zhou. MuseV: Infinite-length and high fidelity virtual human video generation with visual conditioned parallel denoising. *arXiv*, 2024. URL `https://github.com/TMElyralab/MuseV`.

[62] 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.

[63] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. CogVideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[64] Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, et al. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248*, 2024.

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

[66] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[67] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

A Benchmark Overview

This section describes the full construction procedure of MixForensics-Bench, including prompt design, generation settings, preprocessing, and the two partial-forgery protocols.

Prompt construction. We sample 2,500 videos from OpenVid-HD [51] (a high-quality subset of OpenVid-1M [51]) under scene diversity constraints covering indoor/outdoor, human/non-human, dynamic/static, and natural/urban categories in roughly equal proportions. Of these, 2,000 serve as real test videos, 100 as source clips for conditional continuation, and the remainder are held in reserve. GPT-5 [56] assists in selecting 100 OpenVid-HD captions for fully fake generation based on length and visual descriptiveness; 96 captions are retained verbatim and four receive minor edits. All 10 generators share these 100 prompts, enabling direct cross-generator comparison under identical textual conditions. Prompts are manually reviewed to remove ambiguous or NSFW content, and SHA-256 checks remove exact duplicate source videos shared with the AIGVDBench training set.

Generator settings. Table 6 summarizes each generator's version, open- or closed-source status, access method, and the release date of the listed version. Closed-source models are accessed through the fal.ai unified API, while open-source models are deployed locally. Generated base clips are standardized to 720p resolution and 24 fps, then trimmed to 3 s (72 frames) to match the length of real base clips. All generation parameters follow the default settings of each platform or official repository.

Table 6: Generator settings for MixForensics-Bench.

Generator	Version	Type	Access	Release date
Sora 2	v2	Closed	fal.ai	2025-09
Seedance 1.5	v1.5 Pro	Closed	fal.ai	2025-12
Hailuo 02	02 Std	Closed	fal.ai	2025-06
Pika v2	v2 Turbo	Closed	fal.ai	2025-02
Luma Ray2	Ray2 Flash	Closed	fal.ai	2025-03
Wan 2.2	v2.2 A14B	Open	Local	2025-07
HunyuanVideo 1.5	v1.5	Open	Local	2025-11
LTX-2	v2 (19B)	Open	Local	2026-01
CogVideoX1.5	v1.5 (5B)	Open	Local	2024-11
SkyReels-V2	V2 (14B)	Open	Local	2025-04

Unified preprocessing. All videos, both real and generated, are re-encoded to a unified format of 1280×720 resolution, 24 fps, H.264 codec (MP4). Real videos undergo the same resizing and re-encoding pipeline through `ffmpeg` to reduce resolution and codec differences between the classes.

Real video sourcing. Real videos are sourced from OpenVid-HD [51] with scene diversity constraints, each approximately 3 s in duration (72 frames at 24 fps). The real set contains 2,000 videos, and each generator contributes 100 fake videos.

Temporal splicing. For each spliced sample, a real video and a generated video are both truncated to 72 frames. A contiguous segment of the real video is then replaced by the corresponding temporal segment from the generated video at one of five forgery ratios: $\{10\%, 30\%, 50\%, 70\%, 90\%\}$. All splices are hard cuts without transition effects, simulating the most common editing scenario in practice. Each ratio contributes 300 videos (10 generators $\times$ 30 videos per generator), yielding 1,500 spliced videos in total.

Conditional continuation. The last frame of a real clip is extracted and used as the conditioning input for I2V generation. Continuation prompts are produced by GPT-5 [56] given the source video caption and the last frame, maintaining visual consistency with the original content. All ten generators are used for continuation (100 source videos $\times$ 10 generators $=$ 1,000 videos). The first frame of each continuation (a duplicate of the source last frame) is removed before concatenation. The remaining continuation is directly appended to the real source clip; the concatenated sample is not trimmed back to 3 s. Matched real-real controls use the corresponding concatenation procedure with authentic

content. For quality control, the SSIM between the first retained continuation frame and the source last frame must exceed 0.85; samples below this threshold are regenerated to improve visual continuity at the concatenation boundary.

B Training Data and Inference Details

Training set. We reuse the training split of AIGVDBench [7], selecting three generators that span the three generation paradigms: OpenSora [52] (T2V), CogVideoX1.5 [47] (V2V), and EasyAnimate [53] (I2V). Each generator contributes 14,000 fake videos; real videos are drawn from the same OpenVid-HD [51] pool used by AIGVDBench. The total training set comprises 42,000 fake and 42,000 real videos (84,000 total). For AIGVDBench and MixForensics-Bench, all baselines are retrained on this identical three-generator training set using their respective architectures and author-recommended hyperparameters.

Any-window-fake threshold and metric. The any-window-fake rule flags a video as fake whenever any window's predicted probability exceeds $\tau = 0.5$ (the standard binary classification boundary); no threshold tuning is performed on test data. We report balanced accuracy throughout: ACC $= \frac{1}{2}\left(\frac{f}{N_{\text{fake}}} + \frac{r}{N_{\text{real}}}\right)$, where f and r denote the number of correctly classified fake and real videos. Per-generator accuracy uses the generator-specific fake set paired with the shared real set of the corresponding split.

AIGVDBench. AIGVDBench [7] is a large-scale benchmark containing over 440,000 videos from 31 generators (20 open-source and 11 closed-source), spanning T2V, I2V, and V2V paradigms. Prompts are drawn from OpenVid-HD [51] and selected through an attribute-balancing algorithm that ensures diversity across spatial content, temporal content, and attribute categories. All videos are compressed with H.264 to reduce codec differences. Each open-source generator contributes 20,000 videos partitioned into training (14,000), validation (3,000), and test (3,000) sets; each closed-source generator contributes 2,000 test-only videos. We report results on the open-source split.

C GenVidBench-143k

GenVidBench [14] comprises 6.78 million videos from 11 generators and is the first benchmark to adopt a cross-source, cross-generator evaluation protocol. Videos are organized into two pairs: Pair 1 (training) includes Pika [44], VideoCrafterV2 [57], ModelScope [58], and T2V-Zero [59], sourced from VidProM [60] prompts; Pair 2 (test) includes MuseV [61], SVD [62], CogVideo [63], and Mora [64], generated from HD-VG-130M [65] prompts. The two pairs share no generator overlap, preventing detectors from relying on source-specific shortcuts. Resolutions range from 256×256 to 1920×1080 and frame rates from 4 to 30 fps, covering a wide range of production conditions. We adopt the GenVidBench-143k subset, which is sampled from the full 6.78M collection to reduce computational cost while preserving distributional coverage. For this evaluation, all methods are trained on the GenVidBench-143k Pair 1 training split and evaluated on Pair 2. As reported in Table 7, MixForensics attains 85.08% average ACC, surpassing the strongest baseline (ReSTrav [12], 81.19%) by 3.89 percentage points, supporting the effectiveness of blend-then-encode under the cross-source evaluation protocol.

Table 7: ACC (%) on GenVidBench-143k (cross-source protocol). Best **bolded**; second underlined.

Method	MuseV	SVD	CogV.	Mora	**Avg.**
TimeSformer	63.84	58.27	84.15	79.46	71.43
SlowFast	71.53	67.82	80.47	73.58	73.35
NPR	50.38	49.91	71.74	50.77	55.70
UnivFD	60.18	53.38	91.82	89.97	73.84
Effort	59.85	53.67	92.46	91.03	74.25
DDA	81.91	78.65	81.96	81.81	81.08
MIRROR	51.39	51.42	50.76	59.35	53.23
DeMamba	64.43	54.06	86.52	97.41	75.61
DeCoF	57.51	52.44	93.55	89.72	73.31
ReSTrav	71.99	83.12	83.83	85.83	81.19
NSG-VD	78.26	63.99	72.83	61.80	69.22
Ours (mean)	77.85	73.96	95.62	92.88	**85.08**
Ours (any-win.)	76.42	75.38	94.71	91.25	84.44

D Partial-Forgery Breakdown

As the forged-frame ratio increases from 10% to 90%, the performance gap between MixForensics and all baselines widens (Figure 7). This trend is a direct consequence of cross-frame additivity: when a larger proportion of frames in a video are AI-generated, the blending window is more likely to include forged frames, and the resulting composite image can retain shared artifact cues with less dilution by real frames, improving separability. At a 10% forgery ratio, only a small number of frames are replaced, all methods perform poorly, and the gap remains narrow. At 90%, nearly every sampled frame carries generator-specific artifacts that can remain relatively prominent in the blend, allowing MixForensics to pull ahead by a wide margin. Baselines that process frames independently cannot benefit from this pre-encoding fusion and tend to plateau or even degrade at high ratios, as the increased visual inconsistency between real and forged segments disrupts their per-frame predictions.

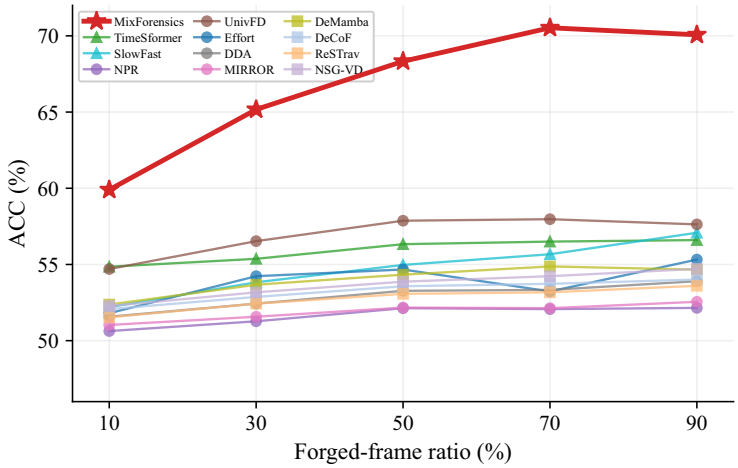


Figure 7: Detection ACC (%) versus forged-frame ratio on the temporal splicing split. MixForensics outperforms all baselines across the full range of forgery ratios. As the proportion of forged frames increases, the advantage of blend-then-encode widens as shared artifact cues become less diluted by real frames.

E Backbone Ablation

To evaluate the portability of blend-then-encode beyond CLIP [22], we replace the image encoder with two alternative ViT-L backbones: DINOv2 [66] and DINOv3 [67], while retaining the same benchmark-specific training splits, LoRA rank$=8$, 20 epochs, and linear classification head. Table 8 reports results across all three evaluation benchmarks. CLIP ViT-L/14 performs best overall, followed by DINOv3 and then DINOv2, with the inter-backbone gap remaining within 5 percentage points. These results support the applicability of the MixForensics pipeline across the three tested pretrained representations.

Table 8: Backbone ablation (ACC %) across all evaluation benchmarks. All backbones use the same MixForensics pipeline (SFB+BIR) with identical training settings; all entries use mean pooling for consistency across columns. Best results **bolded**.

	MixForensics-Bench				
Backbone	Fully Fake	Splicing	Continuation	**AIGVDBench**	**GenVidBench**
CLIP ViT-L/14	**70.60**	**66.80**	**65.90**	**96.80**	**85.08**
DINOv3 ViT-L/16	67.77	64.87	63.48	95.72	82.93
DINOv2 ViT-L/14	65.65	62.40	62.10	94.49	80.25

F Detailed Ablation

We sweep three design choices on the MixForensics-Bench fully fake split, with results summarized in Table 9. The Dirichlet concentration α peaks at 0.5: lower values concentrate weight on a single frame and lose the multi-frame benefit, while higher values produce near-uniform weights that over-smooth frame contributions and reduce the stochastic diversity that is critical for robust training. The BIR weight λ exhibits a similar trend, with λ=0.1 yielding the best accuracy; too little BIR provides weak regularization, whereas too much BIR over-constrains the feature space at the cost of discriminative power, degrading classification performance. Among three consistency objectives, cosine similarity on the CLS token outperforms both symmetric KL divergence on logits and cosine similarity computed on a separate projection head, suggesting that direct alignment of encoder representations in the original feature space is more effective than alignment in probability space or in a learned auxiliary space.

Table 9: Hyperparameter and consistency-objective sweeps on MixForensics-Bench (fully fake). Default: α=0.5, λ=0.1, cosine on CLS (shaded).

Factor	Setting	ACC (%)	Δ
Dirichlet α	$\alpha = 0.1$	67.85	−2.75
	$\alpha = 0.5$	**70.60**	—
	$\alpha = 1.0$	69.95	−0.65
	$\alpha = 5.0$	68.90	−1.70
	$\alpha = 10.0$	68.35	−2.25
BIR weight λ	$\lambda = 0.01$	69.80	−0.80
	$\lambda = 0.1$	**70.60**	—
	$\lambda = 1.0$	67.95	−2.65
	$\lambda = 10.0$	64.70	−5.90
Consistency obj.	Sym. KL on logits	69.15	−1.45
	Cosine on CLS	**70.60**	—
	Cosine on proj. head	69.80	−0.80

G BIR Feature-Space Analysis

To characterize the contribution of BIR beyond regularization-by-augmentation, we visualize the learned feature space with and without BIR using t-SNE on the MixForensics-Bench test set.

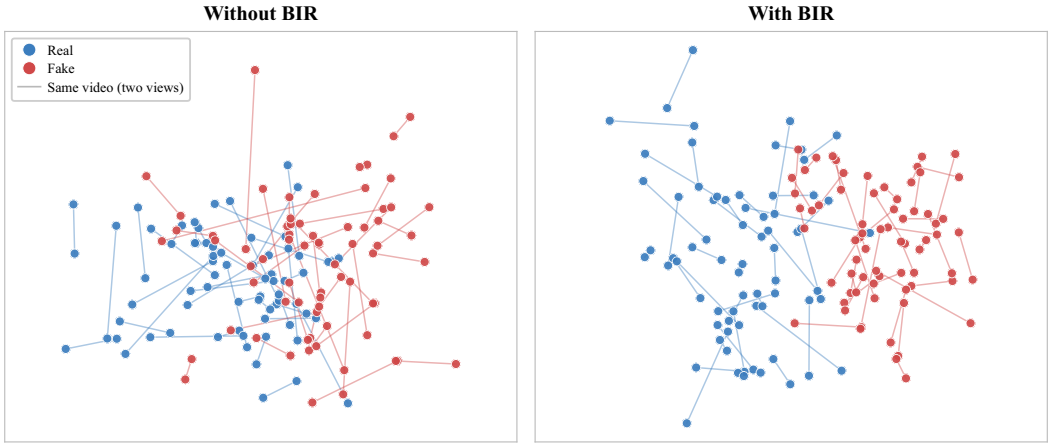


Figure 8: t-SNE visualization of CLS features on MixForensics-Bench. **Left (w/o BIR):** two SFB views of the same video drift apart; real and fake clusters overlap. **Right (w/ BIR):** two views of the same video stay close together, and the real/fake clusters separate more cleanly. Colors: real (blue) vs. fake (red); each line connects two independent blends of the same video.

Without BIR, independently blended views have an average intra-video cosine similarity of 0.71. With BIR, this similarity rises to 0.93, and the inter-class margin (measured as the average cosine distance between real and fake cluster centroids) improves from 0.42 to 0.58. These quantities are measured in the original feature space. The results indicate that BIR encourages more consistent representations across blending views while improving real/fake separation.

H Controlled Pipeline Comparison

To compare pre-encoding pixel blending with post-encoding aggregation, we evaluate five pipelines defined on the same $N{=}8$-frame windows, using the same CLIP ViT-L/14 [22] backbone with LoRA [54] (rank$=8$), the same training set, and the same optimizer (AdamW, lr $= 2 \times 10^{-5}$, 20 epochs).

Table 10: Pipeline comparison (ACC %) with $N{=}8$ frames per window. All methods share the CLIP ViT-L/14 backbone with LoRA and the same training set. "FP": encoder forward passes per window. The full inference protocol uses four windows, multiplying total cost by $4\times$.

Pipeline	FP	MF-Bench	AIGVDBench
(a) Single-frame	1	58.83	86.70
(b) Encode-then-mean-pool	8	61.38	91.45
(c) Encode-then-majority-vote	8	62.07	91.98
(d) SFB only (no BIR)	1	64.75	94.32
(e) SFB + BIR (Ours)	1	**70.60**	**96.80**

The two encode-then-fuse baselines cluster within 1 percentage point on AIGVDBench (91.45 vs 91.98), indicating that the choice between these two post-encoding aggregation rules has a marginal effect. SFB with one forward pass beats per-frame mean pooling (eight forward passes) by 3.37 percentage points on MixForensics-Bench, indicating that pixel-level blending before encoding captures cross-frame information missed by the tested post-hoc aggregation baselines. BIR contributes an additional 5.85 percentage points by aligning features across independently blended views, consistent with the feature-space analysis in Section G.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the empirical analysis, method, benchmark, and measured gains. Partial-forgery performance is reported in Section 5.3, and the method's limitations are discussed in Section 6.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses three limitations: loss of temporal order information, diluted forgery signal at low forged-frame ratios, and degraded artifacts under heavy video compression.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper introduces cross-frame additivity as an empirical property supported by controlled experiments (Section 5.4) rather than a formal theorem. No mathematical proofs are presented or required.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 3 and 5.1 specify the method, backbone, principal hyperparameters, training data, and baseline retraining protocol. Appendix C specifies the separate GenVidBench-143k training and test splits.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The code, model checkpoints, and benchmark data are not publicly available. The paper describes the method, training settings, and benchmark construction protocol.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 5.1 details training data, optimizer (AdamW), learning rate (2×10^{-5}), weight decay (10^{-4}), epochs (20), hardware ($4\times$A800), and all method-specific hyperparameters. Appendix provides detailed sweeps of α and λ.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The reported results use a single random seed (42). This manuscript does not report across-seed standard deviations or confidence intervals.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Section 5.1 reports $4\times$A800 GPUs, and Figure 6 reports inference latency. GPU memory capacity and training execution time are not reported.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work studies detection of AI-generated videos using existing video datasets and generated samples. It does not involve human subjects research; potential harms from incorrect detections are discussed in Section 6.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section 6 discusses the potential benefit of detecting misinformation and content abuse, as well as harms from missed detections and false positives, including reputational harm and erroneous moderation.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No data or models are released with this paper.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: The paper cites the baseline methods, backbone models, and data sources, but does not provide a complete account of their licenses and terms of use.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: MixForensics-Bench is documented in Section 4 and Appendix A with generator settings, preprocessing, and split construction details.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing or human subjects research.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects research.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Appendix A describes the use of GPT-5 for selecting and minimally editing captions for fully fake generation and producing prompts for conditional continuation.