# PHYSICS-GUIDED CASCADED REFOCUSING FOR GENERALIZABLE DEEPFAKE VIDEO DETECTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Advances in generative models have dramatically improved the realism of manipulated face videos, posing escalating threats to public security, information credibility, and digital forensics. Dominant detectors follow a fixed-sampling, one-shot paradigm that overlooks the non-uniform temporal distribution of forgery evidence, where traces concentrate within only a few frames, rendering results unstable and subject to considerable randomness. A second active sampling stage promises gains in accuracy and stability yet confronts an intrinsic trade-off: if locating complementary evidence is costly, enlarging the initial sample would be cheaper. We propose a physics-guided cascaded forensic perception framework resolving this trade-off. In the first stage, a Local–Global Resonance Mechanism aggregates cross-frame CLS tokens into global forgery evidence and re-injects it into local patch representations, while predictive uncertainty lets high-confidence samples terminate early and others trigger re-inspection. Candidate clips are then ranked by feature discrepancy and filtered by abnormal facial motion, a discriminative yet network-free signal, enabling targeted acquisition of complementary evidence at negligible cost. Extensive experiments on multiple benchmarks, unseen manipulations, and cross-dataset settings demonstrate consistent improvements in accuracy, stability, and generalization with reasonable efficiency, while qualitative analyses provide interpretable evidence by highlighting facial regions and abnormal motion patterns.

## 1 INTRODUCTION

Recent advances in generative models Ho et al. (2020); Karras et al. (2019); Rombach et al. (2022) have substantially improved the visual realism of manipulated face videos, posing escalating threats to public security, information credibility, and digital forensics Desai et al. (2024). Although existing deepfake detectors achieve impressive accuracy, the dominant paradigm follows a fixed-sampling, one-shot prediction scheme: a predefined number of frames or clips, sampled randomly or at fixed intervals, is processed once to yield a video-level decision. This paradigm implicitly assumes that the sampled regions faithfully represent the entire video. In practice, however, forgery traces are highly non-uniform along the temporal axis, often concentrating within only a few frames or short-lived dynamic patterns. Consequently, the detection outcome of the same video is largely a matter of sampling luck: a random draw may either hit a strongly manipulated segment or miss it entirely. **The fixed-sampling paradigm overlooks the highly non-uniform temporal distribution of forgery evidence, rendering detection results unstable and irreproducible across sampling positions.**

A natural remedy is a second active sampling stage: after an initial random sample, if the collected evidence is deemed insufficient, the detector selectively acquires an additional, more informative clip and consolidates both observations into a final decision. This coarse-to-fine cascaded strategy promises simultaneous gains in accuracy and stability, yet its viability is governed by an intrinsic trade-off: if locating a complementary clip is itself costly, the overall overhead would exceed that of simply enlarging the initial sample coverage, and the second stage would be rendered pointless. **The essential challenge of second-stage active sampling lies in locating video segments that are complementary to the initial sample and rich in forgery cues at a negligible additional cost—that is, securing the performance of the second inspection without sacrificing efficiency.**



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.

To this end, we propose a physics-guided cascaded forensic perception framework for adaptive evidence acquisition. In the first stage, a Local–Global Resonance Mechanism (LGRM) processes the initially sampled clip to produce enhanced representations and a preliminary prediction, whose reliability is gauged by predictive uncertainty: high-confidence samples terminate early, whereas low-confidence ones trigger re-inspection. In the second stage, candidate clips are first ranked by their feature discrepancy from the initial sample and further filtered by physically grounded cues such as abnormal facial motion, so that physically suspicious and complementary evidence can be selected at marginal cost. The prediction over the selected clip is then fused with the initial one to refine the video-level decision. Through this cascaded design, the detector adaptively determines *whether* additional evidence is required and *where* to acquire it.

The two core components of our framework address the two ends of the aforementioned trade-off, respectively. On the one hand, LGRM aggregates frame-level CLS tokens into global forgery evidence and injects it back into local patch representations. This shared representation plays a dual role: within a single inspection, cross-frame interaction amplifies subtle and spatially localized forgery cues; across the two stages, it directly serves as the retrieval key for selecting clips with complementary cues. As a readily available by-product of the first stage, its reuse introduces nearly zero overhead, endowing the second sampling with efficiency by construction. On the other hand, the physics-guided information is equally dual-purpose: as a feature, it is informative and discriminative for deepfake detection and can be directly fed into the classifier to boost performance; as a signal, it is computed without any deep network and thus incurs negligible cost. Together, the two components jointly resolve the seemingly irreconcilable conflict between efficiency and performance in second-stage sampling.

Our main contributions are summarized as follows:

- We reformulate deepfake video detection from one-shot classification into a cascaded forensic perception process that adaptively accumulates evidence according to sample difficulty: easy samples terminate early while challenging ones undergo targeted re-inspection, explicitly handling the efficiency–performance trade-off in second-stage sampling.
- We propose a Local–Global Resonance Mechanism that aggregates cross-frame CLS tokens into global forgery evidence and injects it back into frame-level patch representations. The shared representation serves twice—amplifying localized forgery cues within each inspection while acting as a zero-overhead retrieval key for the second sampling.
- We develop an uncertainty-driven and physics-guided re-inspection strategy, where predictive confidence decides *whether* to re-inspect, and feature discrepancy together with abnormal facial motion jointly guide *where* to acquire complementary evidence, replacing random frame supplementation with targeted forensic evidence acquisition.
- Extensive experiments demonstrate the superiority of our method in detection performance, robustness, and efficiency. Moreover, by comparing the physical cues and corresponding forgery probabilities across multiple sampled clips, our framework provides region-level evidence for tracing the sources of manipulation in different forged segments, thereby enhancing the interpretability of the detection process.

## 2 Related Work

### 2.1 Generalizable Deepfake Detection

Deepfake detection methods typically identify manipulation traces from spatial appearance, frequency statistics, or pretrained representations. Face X-Ray models blending boundaries shared by different manipulation methods Li et al. (2020), while SBI constructs self-blended images to learn generic forgery artifacts without relying on specific generators Shiohara & Yamasaki (2022). Frequency-based methods further exploit transferable spectral clues, including high-frequency artifacts Luo et al. (2021) and spatial-phase inconsistencies Liu et al. (2021). Other studies improve generalization by learning common forgery representations Yan et al. (2023a) or local self-consistency Zhao et al. (2021).

Recent works increasingly leverage foundation models and adaptive learning strategies. ResProto-FD constructs visual-language residual prototypes for unknown forgery detection Jing et al. (2026),

while DFF-Adapter introduces fine-grained manipulation supervision into parameter-efficient DINO tuning Zhang et al. (2026). Identity-aware VLMs combine low-level artifacts with identity-related semantic cues Xu et al. (2026), and SAFE introduces semantic- and frequency-enhanced curriculum learning Yao et al. (2026). CRDA further uses reinforcement learning to dynamically adjust forgery augmentation according to the detector state Chou et al. (2026), whereas DFD-HR performs hierarchical routing to capture diverse forgery traces Sun et al. (2026). These approaches improve representation generalization, but generally adopt a predetermined inference process for each input.

### 2.2 Temporal and Physical Video Forensics

Video-based detectors exploit inconsistencies in facial dynamics and cross-frame representations. ID-Reveal detects identity-specific temporal irregularities Cozzolino et al. (2021), while LipForensics models abnormal mouth motion through visual speech pretraining Haliassos et al. (2020). FTCN emphasizes temporal coherence by restricting spatial convolution Zheng et al. (2021), and AltFreezing balances spatial and temporal learning through alternating optimization Wang et al. (2023). TALL rearranges consecutive frames into a compact thumbnail representation Xu et al. (2023), whereas TI2Net captures temporal identity inconsistency Liu et al. (2023).

More recent approaches investigate style trajectories Choi et al. (2024), temporal surface anomalies Ciamarra et al. (2024), video-level self-blending Yan et al. (2025b), and vulnerability-aware spatiotemporal learning Nguyen et al. (2025). DeepShield jointly analyzes local and global forgery evidence Cai et al. (2025), FakeRadar detects unknown forgeries as representation outliers Li et al. (2025), and temporal-frequency modeling exposes pixel-wise dynamic artifacts Kim et al. (2025). Although these methods demonstrate the importance of temporal evidence, they process predetermined clips or densely aggregate an entire sequence. In contrast, we use facial motion inconsistency to determine where complementary evidence should be acquired after an uncertain initial prediction.

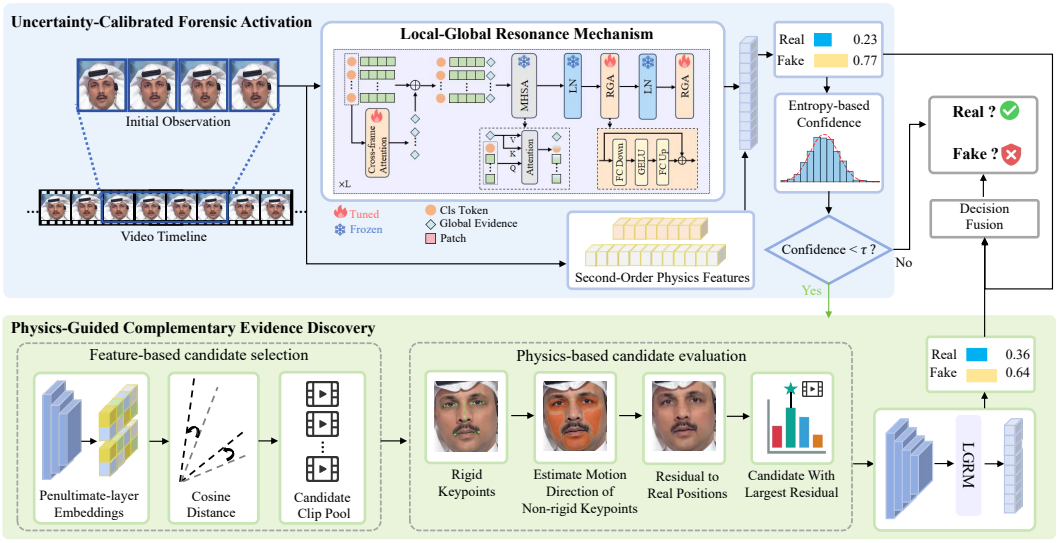



Figure 1: Overview of the proposed cascaded forensic perception framework. An initial clip is analyzed using local–global resonance features and second-order physical cues. Low-confidence samples trigger complementary evidence discovery, where candidates are selected by feature distance and evaluated through facial motion residuals. The selected clip is re-inspected and fused with the initial prediction for the final decision.

## 3 Method

### 3.1 Overview

Given an input video $\mathcal{V} = \{I_t\}_{t=1}^{N}$, our framework performs cascaded forensic perception according to sample difficulty. An initial clip is encoded by the Local–Global Resonance Mechanism

(LGRM), which aggregates cross-frame global evidence and enhances local regions for prediction. The Uncertainty-Calibrated Forensic Activation (UCFA) module then determines whether reinspection is required. For uncertain samples, Physics-Guided Evidence Discovery (PGED) selects a complementary clip with distinctive features and abnormal facial dynamics, whose prediction is fused with the initial result to refine the final decision.

### 3.2 Local–Global Resonance Mechanism

Given a clip $\mathbf{V} \in \mathbb{R}^{T\times C\times H\times W}$, each frame is divided into $N = HW/P^2$ non-overlapping patches of size $P \times P$. The flattened patch $\mathbf{x}_{t,i} \in \mathbb{R}^{P^2C}$ is projected into a $D$-dimensional embedding, together with a learnable CLS token $\mathbf{c}_t$ and spatial positional embedding $\mathbf{E}_s$. The frame- and clip-level representations are initialized as:

$$\begin{aligned} \mathbf{X}_t^0 &= [\mathbf{c}_t; \mathbf{W}\mathbf{x}_{t,1}; \cdots; \mathbf{W}\mathbf{x}_{t,N}] + \mathbf{E}_s, \\ \mathbf{Z}^0 &= [\mathbf{X}_1^0, \mathbf{X}_2^0, \ldots, \mathbf{X}_T^0], \end{aligned} \tag{1}$$

where $\mathbf{X}_t^0 \in \mathbb{R}^{(N+1)\times D}$ and $\mathbf{Z}^0 \in \mathbb{R}^{T\times(N+1)\times D}$.

To enhance weak forgery traces, LGRM is inserted into each CLIP Transformer block. At the $l$-th layer, the CLS tokens from all frames are aggregated and processed by cross-frame self-attention:

$$\begin{aligned} \mathbf{C}^l &= [\mathbf{c}_1^l, \mathbf{c}_2^l, \ldots, \mathbf{c}_T^l], \\ \mathbf{E}^l &= \mathbf{C}^l + \mathrm{MHSA}\big(\mathbf{C}^l\big), \end{aligned} \tag{2}$$

where $\mathbf{E}^l = [\mathbf{e}_1^l, \ldots, \mathbf{e}_T^l]$ represents global forgery evidence. This operation preserves frame-specific information while integrating shared and complementary cues across frames.

The evidence token $\mathbf{e}_t^l$ is then appended to the corresponding spatial sequence and interacts with intra-frame patches:

$$\begin{aligned} \widetilde{\mathbf{X}}_t^l &= [\mathbf{c}_t^l; \mathbf{x}_{t,1}^l; \cdots; \mathbf{x}_{t,N}^l; \mathbf{e}_t^l], \\ \hat{\mathbf{X}}_t^l &= \widetilde{\mathbf{X}}_t^l + \mathrm{MHSA}\big(\widetilde{\mathbf{X}}_t^l\big). \end{aligned} \tag{3}$$

Here, the global token guides attention toward local regions consistent with cross-frame forgery evidence.

**Resonance-Guided Adapter.** Since the pretrained CLIP backbone is frozen, we introduce a lightweight adapter $\mathcal{A}(\cdot)$ to adapt the resonance-enhanced features. It consists of a down-projection, nonlinear activation, and up-projection, and is inserted around the frozen MLP:

$$\mathbf{X}_t^{l+1} = \hat{\mathbf{X}}_t^l + \mathcal{A}\Big(\mathrm{MLP}_{\mathrm{frozen}}\Big(\mathcal{A}\big(\mathrm{LN}(\hat{\mathbf{X}}_t^l)\big)\Big)\Big). \tag{4}$$

The resulting clip representation $\mathbf{Z}^{l+1} = [\mathbf{X}_1^{l+1}, \ldots, \mathbf{X}_T^{l+1}]$ is forwarded to the next layer. Through stacked blocks, global evidence aggregation and local feature enhancement progressively reinforce subtle and localized forgery artifacts.

### 3.3 Uncertainty-Calibrated Forensic Activation

We first sample a continuous clip of $T$ frames from the input video:

$$\mathcal{C}^{(0)} = \{I_{s_0}, I_{s_0+1}, \ldots, I_{s_0+T-1}\}, \tag{5}$$

where $s_0$ is the starting position and $T = 8$ in our experiments. Since motion in real-world videos is constrained by physical dynamics, acceleration provides an effective measure of temporal motion variation. Inspired by this property, we introduce second-order temporal differences into the initial perception stage to capture manipulation-induced dynamic inconsistencies.

Given the frame-level feature sequence $\{\mathbf{F}_t\}_{t=1}^T$ extracted from $\mathcal{C}^{(0)}$, the first- and second-order differences are defined as:

$$\Delta\mathbf{F}_t = \mathbf{F}_{t+1} - \mathbf{F}_t, \qquad \Delta^2\mathbf{F}_t = \mathbf{F}_{t+1} - 2\mathbf{F}_t + \mathbf{F}_{t-1}. \tag{6}$$

Here, $\Delta\mathbf{F}_t$ describes feature-level velocity, while $\Delta^2\mathbf{F}_t$ approximates acceleration under a unit temporal interval. Authentic facial motion generally exhibits coherent temporal evolution, whereas

abrupt second-order variations may reveal unnatural dynamics introduced by manipulation. We therefore enhance the original representation as:

$$\widetilde{\mathbf{F}}_t = \mathbf{F}_t + \phi\big(\Delta^2 \mathbf{F}_t\big), \tag{7}$$

where $\phi(\cdot)$ is a learnable projection for adapting the second-order dynamic features.

Based on the enhanced feature sequence, the initial perception network $\mathcal{F}_\theta$ produces a fake probability and a forensic representation:

$$\left(p^{(0)}, \mathbf{z}^{(0)}\right) = \mathcal{F}_\theta\left(\{\widetilde{\mathbf{F}}_t\}_{t=1}^{T}\right), \tag{8}$$

where $p^{(0)} \in [0,1]$ and $\mathbf{z}^{(0)} \in \mathbb{R}^d$ denote the initial prediction and penultimate-layer representation, respectively.

To determine whether the current observation contains sufficient evidence, we calculate the normalized predictive entropy:

$$H\left(p^{(0)}\right) = -\frac{p^{(0)} \log p^{(0)} + \big(1-p^{(0)}\big) \log\big(1-p^{(0)}\big)}{\log 2}. \tag{9}$$

The corresponding confidence is defined as $c^{(0)} = 1 - H(p^{(0)})$, and forensic re-inspection is activated according to:

$$g(\mathcal{V}) = \mathbb{I}\left[c^{(0)} < \tau\right], \tag{10}$$

where $\tau$ is the activation threshold. High-confidence samples retain the initial prediction, whereas low-confidence samples are forwarded to the subsequent evidence discovery stage. This design enhances the initial representation with physically inspired dynamics while allocating additional computation only to ambiguous videos.

### 3.4 Physics-Guided Complementary Evidence Discovery

For an activated video, we generate $T$-frame candidate clips with temporal stride $\Delta$. Clips overlapping with or located too close to the initial observation are excluded:

$$\mathcal{P} = \{\mathcal{C}_k \mid \mathcal{C}_k = \{I_{s_k}, \ldots, I_{s_k+T-1}\},\ |s_k - s_0| \geq T + \delta\}, \tag{11}$$

where $\delta$ controls the temporal exclusion margin. This constraint ensures that the additional observation is acquired from a previously unobserved region.

Temporal separation alone does not guarantee information complementarity. We therefore extract the representation $\mathbf{z}_k = \mathcal{E}_\theta(\mathcal{C}_k)$ for each candidate using the encoder of the initial perception network and calculate its cosine distance from the initial representation:

$$d_k = 1 - \frac{\big(\mathbf{z}^{(0)}\big)^\top \mathbf{z}_k}{\big\|\mathbf{z}^{(0)}\big\|_2 \, \|\mathbf{z}_k\|_2}. \tag{12}$$

The $K$ candidates with the largest representation discrepancies are retained:

$$\mathcal{P}_K = \mathrm{TopK}\left(\mathcal{P}, \{d_k\}, K\right). \tag{13}$$

Although representation discrepancy suppresses redundant observations, large differences may also result from natural pose, illumination, or expression variations. We therefore introduce a facial motion prior to identify candidates containing physically unexplained local dynamics.

For two adjacent frames, facial landmarks are divided into a relatively rigid subset $\mathcal{R}$, including the eye corners and nose region, and a local dynamic subset $\mathcal{L}$ around the eyes and mouth. A partial affine transformation is estimated from the rigid landmarks:

$$\mathbf{A}_t = \arg\min_{\mathbf{A}} \sum_{i \in \mathcal{R}} \rho\left(\|\mathbf{p}_{t+1,i} - \mathbf{A}\widetilde{\mathbf{p}}_{t,i}\|_2^2\right), \tag{14}$$

where $\widetilde{\mathbf{p}}_{t,i} = [x_{t,i}, y_{t,i}, 1]^\top$ is the homogeneous landmark coordinate and $\rho(\cdot)$ is implemented using RANSAC. The estimated transformation captures global translation, rotation, and scale changes between adjacent frames.

We then apply $\mathbf{A}_t$ to the local landmarks and measure the discrepancy between their predicted and observed positions. The physical inconsistency of candidate $\mathcal{C}_k$ is defined as:

$$R_k = \frac{1}{|\Omega_k|} \sum_{t \in \Omega_k} \frac{1}{|\mathcal{L}|} \sum_{i \in \mathcal{L}} \frac{\|\mathbf{p}_{t+1,i} - \mathbf{A}_t \widetilde{\mathbf{p}}_{t,i}\|_2}{d_t^{\text{eye}} + \epsilon}, \tag{15}$$

where $\Omega_k$ denotes the valid adjacent-frame pairs in the candidate, $d_t^{\text{eye}}$ is the inter-ocular distance, and $\epsilon$ prevents numerical instability. The scale normalization reduces the influence of face size and camera distance.

The complementary evidence selected for re-inspection is:

$$k^* = \underset{k:\mathcal{C}_k \in \mathcal{P}_K}{\arg\max} R_k, \qquad \mathcal{C}^{(r)} = \mathcal{C}_{k^*}. \tag{16}$$

This coarse-to-fine search first identifies observations that complement the initial representation and then prioritizes the candidate containing the strongest unexplained facial dynamics. It therefore prevents the evidence search from being dominated by either redundant content or irrelevant appearance variation.

Table 1: Cross-dataset evaluation. Different methods are compared on four datasets using AUC scores (%). The best results are highlighted in bold, and the second-best results are underlined.

| Method | Venue | Input Type | Datasets | | | |
|---|---|---|---|---|---|---|
| | | | CDF-v2 | DFD | DFDC | FFIW |
| DAM Zhou et al. (2021) | CVPR'21 | Video | 75.3±1.6 | - | - | - |
| LipForensics Haliassos et al. (2020) | CVPR'21 | Video | 82.4±1.4 | - | 73.5±1.8 | - |
| FTCN Zheng et al. (2021) | ICCV'21 | Video | 86.9±1.2 | - | 74.0±1.6 | 70.9±1.7 |
| SST Choi et al. (2024) | CVPR'24 | Video | 89.0±0.9 | 96.1±0.6 | - | - |
| AltFreezing Wang et al. (2023) | CVPR'24 | Video | 89.5±0.8 | 98.5±0.4 | - | 73.0±1.5 |
| NACO Zhang et al. (2024) | ECCV'24 | Video | 89.5±0.9 | - | 76.7±1.3 | - |
| FCG Han et al. (2025) | CVPR'25 | Video | 92.3±0.7 | - | 81.2±1.1 | - |
| STA Yan et al. (2025b) | CVPR'25 | Video | 94.7±0.5 | 96.5±0.5 | 84.3±0.9 | 92.1±0.7 |
| MPSTL Miao et al. (2026) | TOMM'26 | Video | 91.6±0.8 | - | 76.0±1.4 | - |
| TALL-Swin Xu et al. (2023) | ICCV'23 | Frame | 90.8±0.8 | - | 76.8±1.2 | - |
| UCF Yan et al. (2023a) | ICCV'23 | Frame | 83.7±1.3 | 86.7±1.1 | 74.2±1.5 | 69.7±1.8 |
| CADDM Dong et al. (2023) | CVPR'23 | Frame | 93.9±0.6 | 99.0±0.3 | 71.8±1.4 | 80.6±1.0 |
| LSDA Yan et al. (2024) | CVPR'24 | Frame | 91.1±0.8 | - | 77.0±1.2 | - |
| UDD Fu et al. (2025) | AAAI'25 | Frame | 93.1±0.6 | 95.5±0.6 | 81.2±1.0 | - |
| Effort Yan et al. (2025a) | ICML'25 | Frame | 95.6±0.4 | 96.5±0.5 | 84.3±0.8 | 92.1±0.6 |
| AUNet Bai et al. (2023) | CVPR'23 | Frame | 92.8±0.7 | **99.2±0.2** | 73.8±1.3 | 81.5±0.9 |
| CDFA Lin et al. (2024) | ECCV'24 | Frame | 94.6±0.5 | - | 81.2±0.9 | - |
| FreqBlender Zhou et al. (2024) | NeurIPS'24 | Frame | 94.5±0.5 | - | 74.6±1.3 | 86.1±0.8 |
| SJEDD Zou et al. (2025) | TCSVT'25 | Frame | 92.2±0.7 | - | 84.1±0.8 | - |
| FA Cui et al. (2025) | CVPR'25 | Frame | 95.7±0.4 | 97.2±0.4 | <u>87.2±0.6</u> | - |
| GenD Yermakov et al. (2026) | WACV'26 | Frame | <u>96.0±0.3</u> | 97.0±0.4 | 87.1±0.5 | <u>92.8±0.5</u> |
| Ours | - | Video | **96.6±0.2** | 98.2±0.1 | **91.4±0.2** | **94.0±0.2** |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETTINGS

**Dataset.** We evaluate our method on a diverse set of commonly used deepfake video benchmarks, which cover different manipulation techniques, data scales, and real-world content distributions. (1) **FaceForensics++ (FF++)** includes 1,000 pristine videos and 4,000 manipulated videos produced by four face manipulation algorithms. (2) **FFIW-10k (FFIW)** contains 8,250 authentic videos and the same number of forged samples, resulting in a balanced evaluation set. (3) **DeepFake Detection Challenge (DFDC)** consists of 1,133 real videos and 4,080 fake videos collected under relatively diverse and realistic conditions. (4) **DeepfakeDetection (DFD)** contains 3,434 videos, including

363 real samples and 3,071 manipulated samples. (5) **CelebDF-v2 (CDF-v2)** provides 590 authentic videos and 5,639 high-quality deepfake videos. (6) **DF40** covers 40 deepfake generation techniques, from which five representative manipulation methods are selected in our experiments. Following prior studies, we adopt video-level Area Under the Curve (AUC) as the evaluation metric on all deepfake datasets.

**Implementation Details.** We train our model on the c23 compression setting of FF++ and follow the data preprocessing protocol provided by DeepfakeBench Yan et al. (2023b). During training, each input consists of a clip containing 8 consecutive frames. At inference time, one 8-frame clip with a randomly selected starting position is sampled from each video. Model optimization is performed using Adam Kinga et al. (2015) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and zero weight decay. The learning rate is initialized as $8 \times 10^{-5}$ and gradually reduced to $5 \times 10^{-5}$ over 50 epochs using a cosine decay schedule. The entropy-confidence threshold $\tau$ is set to $0.55$ based exclusively on the FF++ validation set and is fixed for all cross-dataset evaluations. All experiments are conducted for 50 epochs on four NVIDIA GeForce RTX 4090 GPUs with a total batch size of 32. To ensure a fair comparison, each model is evaluated five times using different randomly sampled initial clips, and we report the mean AUC and standard deviation.

### 4.2 Generalization Experiments

**Cross-dataset evaluation.** Table 1 compares our method with representative frame- and video-based detectors under cross-dataset settings. Our approach achieves the best performance on CDF-v2, DFDC, and FFIW, with AUC scores of 96.6%, 91.4%, and 94.0%, respectively. It outperforms the second-best method by 4.2 percentage points on DFDC and 1.2 percentage points on FFIW. Although several frame-based methods perform slightly better on DFD, our method still achieves a competitive AUC of 98.2%. Among methods reporting results on all four datasets, our approach obtains the highest average AUC of 95.1%, exceeding GenD by 1.8 percentage points.

More importantly, our method shows the lowest performance variation across all four datasets, with standard deviations of only 0.2, 0.1, 0.2, and 0.2, respectively. This indicates that the proposed cascaded re-inspection strategy reduces sensitivity to the initially sampled clip. When the first-stage evidence is insufficient, the detector actively searches for a complementary and physically suspicious segment rather than relying on a single sampling position, leading to more stable and reproducible predictions.

The strong cross-dataset performance mainly benefits from the joint use of complementary representation cues and physically grounded facial dynamics. LGRM enhances subtle cross-frame forgery evidence and provides a global representation for retrieving dissimilar candidate clips, while the physics-guided branch identifies abnormal local motion relative to global facial movement. Their combination enables the detector to acquire clips that are both complementary and forensically informative, improving cross-domain generalization while reducing dependence on random sampling.

Table 2: Cross-method evaluation on different face manipulation techniques.

| Method | Uniface | FaceDancer | FSGAN | InSwap | SimSwap |
|---|---|---|---|---|---|
| RECCE | 84.2 | 78.3 | 88.4 | 79.5 | 73.0 |
| SBI | 64.4 | 44.7 | 87.9 | 63.3 | 56.8 |
| CORE | 81.7 | 71.7 | 91.1 | 79.4 | 69.3 |
| IID | 79.5 | 79.0 | 86.4 | 74.4 | 64.0 |
| UCF | 78.7 | 80.0 | 88.1 | 76.8 | 64.9 |
| LSDA | 85.4 | 75.9 | 83.2 | 81.0 | 72.7 |
| CDFA | 76.5 | 75.4 | 84.8 | 72.0 | 76.1 |
| FIA-USA | 91.8 | 83.0 | 86.3 | 87.4 | 91.0 |
| STA | <u>96.0</u> | <u>91.2</u> | **96.4** | <u>93.7</u> | <u>93.1</u> |
| Ours | **98.1** | **92.3** | <u>95.9</u> | **98.4** | **95.1** |

**Cross-method evaluation.** Table 2 reports the generalization performance across five unseen face manipulation techniques. Our method achieves the best results on four of the five manipulation methods, with AUC scores of 98.1%, 92.3%, 98.4%, and 95.1% on Uniface, FaceDancer, InSwap, and SimSwap, respectively. It also obtains a competitive AUC of 95.9% on FSGAN, only 0.5 %

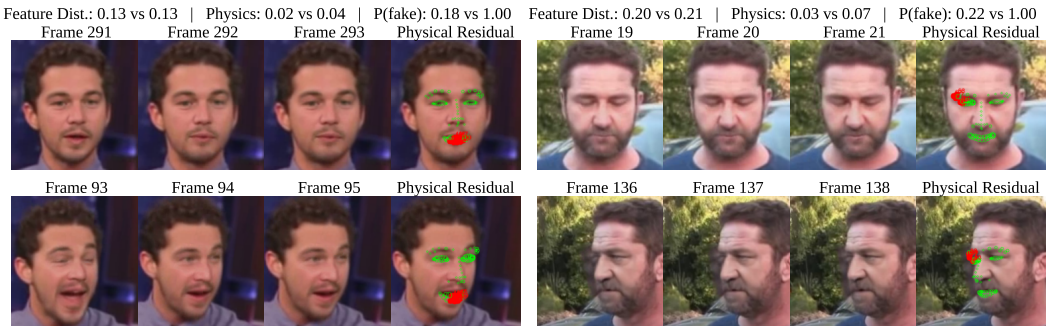
below the best-performing method. Compared with the strongest competing approach, STA, our method yields particularly notable gains of 4.7 % on InSwap and 2.1% on Uniface. Moreover, our method achieves an average AUC of 96.0% across the five manipulation techniques, exceeding STA by 1.9 %. These results demonstrate that the proposed framework is not restricted to artifacts associated with a particular generation pipeline. We attribute this robustness to the incorporation of physically grounded facial dynamics, which provides manipulation-independent evidence beyond appearance-level artifacts. Furthermore, for uncertain samples, the re-inspection mechanism actively retrieves temporal segments that are complementary to the initial observation and exhibit suspicious local motion patterns. The aggregation of initial and supplementary forensic evidence therefore reduces the risk of missing sparsely distributed forgery cues and improves generalization to previously unseen manipulation methods.

Table 3: Ablation of model components.

| LGRM | UCFA | PGED | | CDF-v2 | DFD | DFDC | FFIW |
|---|---|---|---|---|---|---|---|
| | | Feature | Motion | | | | |
| | | | | 83.5 | 84.3 | 78.2 | 81.7 |
| ✓ | | | | 91.4 | 92.5 | 82.8 | 88.3 |
| ✓ | ✓ | | | 94.7 | 96.9 | 87.1 | 91.8 |
| ✓ | ✓ | ✓ | | 95.1 | 97.2 | 87.5 | 92.0 |
| ✓ | ✓ | | ✓ | 95.4 | 97.7 | 88.4 | 92.6 |
| ✓ | ✓ | ✓ | ✓ | **96.6** | **98.2** | **91.4** | **94.0** |

### 4.3 Ablation Study

**Component Ablation.** Table 3 evaluates the contribution of the main components. Adding LGRM improves the baseline AUC from 83.5%, 84.3%, 78.2%, and 81.7% to 91.4%, 92.5%, 82.8%, and 88.3% on CDF-v2, DFD, DFDC, and FFIW, respectively, demonstrating the effectiveness of local–global feature interaction. Incorporating UCFA further improves the results to 94.7%, 96.9%, 87.1%, and 91.8%, showing that uncertainty-guided activation benefits samples with insufficient initial evidence. We further disentangle the two cues in PGED. Feature-based selection improves performance by retrieving complementary observations, while motion-based selection yields larger gains, especially on DFDC, indicating that abnormal facial dynamics provide informative cross-domain cues. Combining both cues achieves the best AUCs of 96.6%, 98.2%, 91.4%, and 94.0%. These results demonstrate that feature discrepancy and motion inconsistency are complementary: the former reduces redundant observations, while the latter prioritizes physically suspicious segments, jointly improving cross-dataset generalization.




(a) Mouth-region case (b) Eye-region case

Figure 2: Qualitative interpretability analysis of physics-guided re-inspection.

### 4.4 Visualization Analysis

To further analyze the interpretability of the physics-guided re-inspection mechanism, we visualize two representative cases. In each case, the paired candidate clips have nearly identical feature discrepancies: $0.13$ vs. $0.13$ for the mouth case and $0.20$ vs. $0.21$ for the eye case. This indicates that feature discrepancy alone is insufficient to determine which clip contains stronger forensic evidence.

In contrast, the physical cue further separates these semantically similar candidates. In the mouth case, the physical residual increases from $0.02$ to $0.04$, while the forgery probability rises from $0.18$ to $1.00$. In the eye case, the residual increases from $0.03$ to $0.07$, accompanied by a probability increase from $0.22$ to $1.00$. The residual responses are concentrated around the mouth and eye regions, respectively, and are consistent with the local dynamic abnormalities observed through manual inspection, demonstrating the complementary and region-level interpretability of the physical cue.

### 4.5 Efficiency Analysis

Table 4: Accuracy–efficiency comparison under different inference strategies on CDF-v2. The activation rate denotes the proportion of videos that trigger re-inspection. Latency includes candidate encoding, landmark processing, and all detector forward passes.

| Method | Activation (%) ↓ | FLOPs (G) ↓ | Latency (ms/video) ↓ | AUC (%) ↑ | ACC (%) ↑ |
|---|---|---|---|---|---|
| One-clip | – | **10.0** | **15.5** | 93.7 | 87.4 |
| Fixed multi-clip | 100 | 44.4 | 32.3 | 94.1 | 88.3 |
| Compute-matched multi-clip | – | 18.6 | 19.7 | 94.7 | 88.9 |
| Ours | **21.86** | 17.5 | 23.4 | **96.6** | **90.2** |

As shown in Table 4, the proposed adaptive re-inspection strategy achieves a favorable trade-off between detection performance and computational cost. The one-clip baseline requires only 10.0 GFLOPs and 15.5 ms per video, but achieves relatively limited performance of 93.7% AUC and 87.4% ACC. Applying fixed multi-clip inference to all videos increases the computational cost substantially to 44.4 GFLOPs and 32.3 ms, while providing only marginal performance gains. In contrast, our method activates the second-stage re-inspection for only 21.86% of the videos, resulting in an average cost of 17.5 GFLOPs and 23.4 ms per video. Despite requiring fewer FLOPs than the compute-matched multi-clip baseline (17.5 vs. 18.6 GFLOPs), our method improves AUC from 94.7% to 96.6% and ACC from 88.9% to 90.2%. Compared with fixed multi-clip inference, it reduces the computational cost by 60.6% in terms of FLOPs and 27.6% in latency, while improving AUC and ACC by 2.5 and 1.9 percentage points, respectively. These results indicate that the performance improvement does not simply arise from processing more clips; instead, selectively allocating additional computation to uncertain samples yields a more effective use of the inference budget.

## 5 Conclusion

In this work, we reformulate deepfake video detection from fixed-sampling, one-shot prediction into an adaptive cascaded forensic perception process. The proposed framework integrates local–global representation interaction, uncertainty-driven re-inspection, and physics-guided complementary evidence discovery to determine both whether and where additional evidence should be acquired. Extensive experiments demonstrate strong cross-dataset and cross-manipulation generalization, improved sampling stability, and a favorable accuracy–efficiency trade-off. Moreover, the physical residuals provide region-level explanations by localizing suspicious dynamics around facial regions such as the mouth and eyes, consistent with manual inspection. Overall, the proposed framework offers an effective and interpretable solution for robust and efficient deepfake video forensics, while future work will extend it toward online evidence acquisition for streaming scenarios.

### AI USE STATEMENT

In this work, we did not use generative AI tools for any task requiring disclosure under the ICLR 2027 AI Policy. Generative AI tools were used solely to edit the manuscript for readability, specifically to identify and correct grammatical and spelling errors and to improve the clarity and fluency of the language. No generative AI tool was used to generate or alter the research ideas, methodology, scientific claims, citations, data, code, figures, experimental results, analyses, or conclusions. All AI-assisted edits were reviewed by the authors, who verified that the revisions preserved the intended technical meaning and did not introduce unsupported content. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

### ETHICS STATEMENT

This work develops a defensive method for detecting manipulated facial videos and is intended to support research in digital media forensics. The experiments use previously released research benchmarks, including FaceForensics++, FFIW-10k, the DeepFake Detection Challenge dataset, DeepfakeDetection, Celeb-DF-v2, and DF40. These datasets were used solely for research purposes and in accordance with their applicable terms of use. This study did not recruit or interact with human participants, collect new personal data, or release a new dataset.

Facial-video benchmarks may contain identifiable individuals and may reflect demographic, geographic, and collection biases. Consequently, the reported results should not be interpreted as demonstrating uniform performance across all populations or real-world settings. Deepfake detectors may also produce false positives or false negatives and should not be used as the sole basis for consequential decisions, identity attribution, surveillance, or legal conclusions. Although the proposed method is intended for defensive forensic applications and does not generate or enhance manipulated media, advances in detection may potentially inform adaptive attacks. Any practical deployment should therefore incorporate appropriate privacy safeguards, lawful and consensual data use, human oversight, security controls, and evaluation on data representative of the intended application domain. The authors take responsibility for the ethical conduct and reporting of this research.

### REPRODUCIBILITY STATEMENT

We are committed to ensuring the reproducibility of our work. All datasets used in this paper, including FaceForensics++, FFIW-10K, DFDC, DeepfakeDetection, Celeb-DF-v2, and DF40, are publicly available. The proposed framework is fully specified in Section 3: the Local–Global Resonance Mechanism is given by Eqs. 1–4, the Uncertainty-Calibrated Forensic Activation by Eqs. 5–10, and Physics-Guided Complementary Evidence Discovery by Eqs. 11–16, with architectural and optimization hyperparameters listed in the *Implementation Details* paragraph. To reduce the sensitivity of results to sampling randomness, all cross-dataset evaluations adopt a single activation threshold $\tau$ selected exclusively on the FF++ validation set, and each model is evaluated five times with different randomly sampled initial clips; the mean AUC and standard deviation are reported in Tables 1–4. The complete source code, configuration files, and scripts for reproducing all main results will be released at github upon acceptance.

## REFERENCES

Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, and Weiming Hu. Aunet: Learning relations between action units for face forgery detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 24709–24719, June 2023.

Yinqi Cai, Jichang Li, Zhaolun Li, Weikai Chen, Rushi Lan, Xi Xie, Xiaonan Luo, and Guanbin Li. Deepshield: Fortifying deepfake video detection with local and global forgery analysis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12524–12534, 2025.

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake video detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1133–1143, 2024.

Yuxuan Chou, Tao Yu, Wen Huang, Tao Dai, Shu-Tao Xia, et al. Improving deepfake detection with reinforcement learning-based adaptive data augmentation. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 40, pp. 3381–3389, 2026.

Andrea Ciamarra, Roberto Caldelli, and Alberto Del Bimbo. Temporal surface frame anomalies for deepfake video detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3837–3844, 2024.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 15108–15117, 2021.

Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 19207–19217, 2025.

Akshar Prabhu Desai, Tejasvi Ravi, Mohammad Luqman, Mohit Sharma, Nithya Kota, and Pranjul Yadav. Gen-ai for user safety: A survey. In *Proc. IEEE Int. Conf. Big Data (BigData)*, pp. 5315–5324, December 2024.

Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, and Zheng Ge. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3994–4004, June 2023.

Xinghe Fu, Zhiyuan Yan, Taiping Yao, Shen Chen, and Xi Li. Exploring unbiased deepfake detection via token-level shuffling and mixing. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 39, pp. 3040–3048, February 2025.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. *arXiv preprint arXiv:2012.07657*, 2020.

Yue-Hua Han, Tai-Ming Huang, Kai-Lung Hua, and Jun-Cheng Chen. Towards more general video-based deepfake detection through facial component guided adaptation for foundation model. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 22995–23005, June 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pp. 6840–6851, December 2020.

Jiuyao Jing, Yu Zheng, and Chunlei Peng. Resproto-fd: Visual-language residual prototype sets for generalized face forgery detection. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 40, pp. 5548–5556, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4401–4410, June 2019.

Taehoon Kim, Jongwook Choi, Yonghyun Jeong, Haeun Noh, Jaejun Yoo, Seungryul Baek, and Jongwon Choi. Beyond spatial frequency: Pixel-wise temporal frequency-based deepfake video detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 11198–11207, 2025.

Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In *Proc. Int. Conf. Learn. Represent. (ICLR)*, volume 5, 2015.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5001–5010, June 2020.

Zhaolun Li, Jichang Li, Yinqi Cai, Junye Chen, Xiaonan Luo, Guanbin Li, and Rushi Lan. Faker-adar: Probing forgery outliers to detect unknown deepfake videos. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 13382–13392, 2025.

Yuzhen Lin, Wentang Song, Bin Li, Yuezun Li, Jiangqun Ni, Han Chen, and Qiushi Li. Fake it till you make it: Curricular dynamic forgery augmentations towards general deepfake detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 104–122, September 2024.

Baoping Liu, Bo Liu, Ming Ding, Tianqing Zhu, and Xin Yu. Ti2net: temporal identity inconsistency network for deepfake detection. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 4691–4700, 2023.

Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 772–781, June 2021.

Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16317–16326, 2021.

Hui Miao, Yuanfang Guo, Leo Yu Zhang, Jiantao Zhou, and Yunhong Wang. A multi-grained parallel spatio-temporal learning architecture for deepfake video detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 2026.

Dat Nguyen, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Vulnerability-aware spatio-temporal learning for generalizable deepfake video detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 10786–10796, 2025.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10684–10695, June 2022.

Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18720–18729, June 2022.

Jiamu Sun, Zhiyuan Yan, Ke-Yue Zhang, Taiping Yao, and Shouhong Ding. Dfd-hr: Generalizable deepfake detection via hierarchical routing learning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 13984–13995, 2026.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4129–4138, June 2023.

Junhao Xu, Jingjing Chen, Yang Jiao, Jiacheng Zhang, Zhiyu Tan, Hao Li, and Yu-Gang Jiang. Identity-aware vision-language model for explainable face forgery detection. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 40, pp. 11278–11286, 2026.

Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 22658–22668, October 2023.

Zhiyuan Yan, Yong Zhang, Yanbo Fan, and Baoyuan Wu. Ucf: Uncovering common features for generalizable deepfake detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 22412–22423, June 2023a.

Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 36, pp. 4534–4565, 2023b.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 8984–8994, June 2024.

Zhiyuan Yan, Jiangming Wang, Zhendong Wang, Peng Jin, Ke-Yue Zhang, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Effort: Efficient orthogonal modeling for generalizable ai-generated image detection. In *Proc. Int. Conf. Mach. Learn. (ICML)*, July 2025a.

Zhiyuan Yan, Yandan Zhao, Shen Chen, Mingyi Guo, Xinghe Fu, Taiping Yao, Shouhong Ding, Yunsheng Wu, and Li Yuan. Generalizing deepfake video detection with plug-and-play: Video-level blending and spatiotemporal adapter tuning. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 12615–12625, June 2025b.

Yulin Yao, Kangfeng Zheng, Bin Wu, Chunhua Wu, Jujie Wang, Jiaqi Gao, Minjiao Yang, and Dan Luo. Safe: Semantic-and frequency-enhanced curriculum for cross-domain deepfake detection. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 40, pp. 11919–11927, 2026.

Andrii Yermakov, Jan Cech, Jiri Matas, and Mario Fritz. Deepfake detection that generalizes across benchmarks. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, 2026.

Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency representation for face forgery video detection. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 407–424, September 2024.

Tianxiang Zhang, Peipeng Yu, Zhihua Xia, Longchen Dai, Xiaoyu Zhou, and Hui Gao. Fine-grained dino tuning with dual supervision for face forgery detection. In *Proc. AAAI Conf. Artif. Intell. (AAAI)*, volume 40, pp. 12780–12788, 2026.

Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 15023–15033, 2021.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 15044–15054, 2021.

Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, Junyu Dong, et al. Freqblender: Enhancing deepfake detection by blending frequency knowledge. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pp. 44965–44988, December 2024.

Tianfei Zhou, Wenguan Wang, Zhiyuan Liang, and Jianbing Shen. Face forensics in the wild. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 5778–5788, June 2021.

Mian Zou, Baosheng Yu, Yibing Zhan, Siwei Lyu, and Kede Ma. Semantics-oriented multitask learning for deepfake detection: A joint embedding approach. *IEEE Trans. Circuits Syst. Video Technol.*, 2025.