CATEGORY-ORIENTED UNIVERSAL ADVERSARIAL PERTURBATIONS FOR DEEPFAKE VIDEO DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fake video detectors rely on different forensic evidence, including frame-level artifacts, pretrained visual representations, and temporal dynamics. This diversity makes it difficult for a universal adversarial perturbation to transfer across different detectors. We categorize existing detectors into three categories: Spatial-Artifact-Oriented (SAO), Semantic-Representation-Oriented (SRO), and Temporal-Dynamics-Oriented (TDO), and develop three category-specific attacks within a category-oriented framework. Specifically, we introduce frequency-aware optimization for SAO, a dual-band static perturbation for SRO, and frame-specific temporal perturbations for TDO. We further develop a unified attack that combines these complementary properties into a single structured perturbation and jointly optimizes it across the three detector categories. Experiments on four surrogate and six unseen detectors, multiple video generators, and external datasets show that the proposed category-specific attacks consistently improve transferability over conventional baselines. The unified attack further provides broader cross-category transfer, although with a trade-off in attack strength compared with separately optimized category-specific attacks. These results highlight the importance of detector forensic evidence in adversarial transfer and motivate robustness evaluation across different detection paradigms.

1 INTRODUCTION

Recent advances in generative models have made fake videos increasingly realistic, creating serious risks to the trustworthiness of digital media. These risks are already visible in real-world incidents. For instance, Deepfake videos have been used to impersonate political leaders and spread false information[1], mimic executives in video conferences to facilitate multi-million-dollar fraud[2], and fabricate endorsements by public officials for investment scams[3]. These incidents have made reliable fake video detection an increasingly important component of digital media forensics.

In response, the research community has developed detectors that examine different traces of video generation. Following prior surveys that categorize deepfake detection methods by the artifacts and forensic cues they exploit (Mirsky & Lee, 2021), deepfake video detectors can be classified into three categories. **Spatial-Artifact-Oriented (SAO)** detectors identify frame-level artifacts, including abnormal textures and frequency patterns (Corvi et al., 2025; Li et al., 2026). **Semantic-Representation-Oriented (SRO)** detectors classify features extracted by pretrained visual encoders such as CLIP and DINOv2 (Radford et al., 2021; Oquab et al., 2024). **Temporal-Dynamics-Oriented (TDO)** detectors analyze inconsistencies across consecutive frames (Internò et al., 2025; Zheng et al., 2025; Zhang et al., 2025).

Although these approaches have advanced fake video detection, their robustness to adversarial manipulation remains a critical question. In image classification, small, often imperceptible perturbations can cause models to make incorrect predictions (Szegedy et al., 2014; Goodfellow et al., 2015). This raises a related question for fake video detection: can a bounded perturbation cause a detector to classify a fake video as real, and can such an attack transfer across videos and detectors? To investigate this question, we develop universal Fake-to-Real attacks tailored to each

[1]Fake Zelenskyy surrender video: reuters.com.
[2]Deepfake video-call fraud exceeding $25 million: reuters.com.
[3]Fake investment endorsements by Singapore officials: csa.gov.sg.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

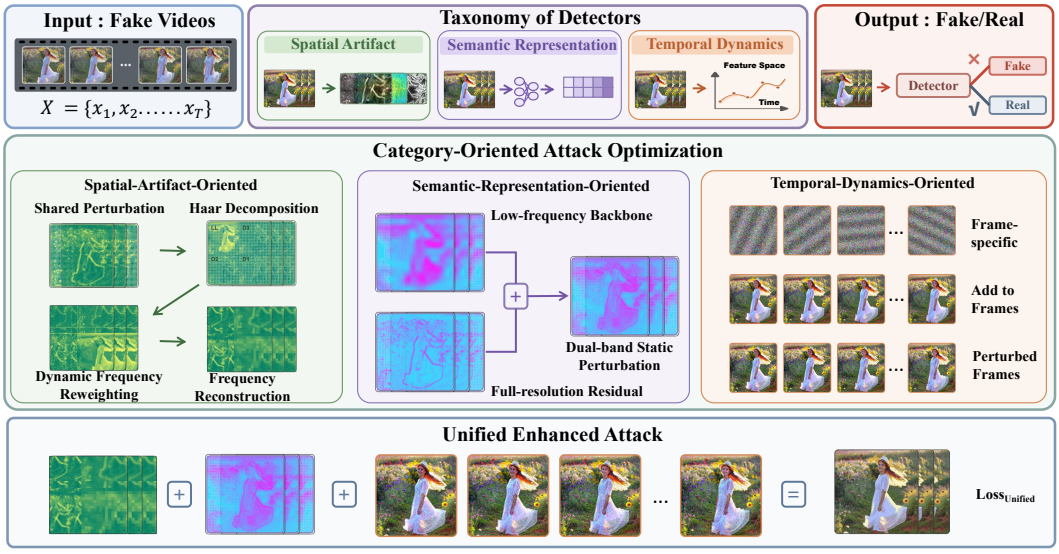

Figure 1: Overview of the proposed attacks. We categorize existing detectors into three categories and develop three category-specific attacks for each group within a category-oriented framework. We further develop a unified attack that combines these complementary properties into a single structured perturbation.

detector category. For SAO detectors, stochastic wavelet reweighting encourages a perturbation to remain effective across spatial-frequency views. For SRO detectors, a dual-band static perturbation combines coarse structure with fine-grained flexibility and is jointly optimized against two visual encoders. For TDO detectors, frame-specific perturbations manipulate the feature changes observed across a video. We further combine these complementary designs into a **unified attack** that optimizes a single structured perturbation across all three categories. Centered temporal components and conflict-aware gradient aggregation help the shared perturbation accommodate their different attack objectives when the target detector's category is unknown. We show the general attack pipeline in Figure 1.

Experiments with four white-box surrogates and six unseen black-box detectors show that these vulnerabilities extend beyond the models used to construct the attacks. On the GenBuster ID split dataset (Wen et al., 2025), the category-specific attacks increase the strict conditional Fake-to-Real flip rate from 14.21% to 64.61% on LGrad (Tan et al., 2023), from 2.73% to 10.41% on ViT-B/16 (Dosovitskiy et al., 2021), and from 3.90% to 85.10% on NSG-VD (Zhang et al., 2025). The unified perturbation also transfers across detector categories, achieving flip rates of 94.37% on LGrad, 20.55% on Swin-Tiny (Liu et al., 2021), and 64.32% on ReStraV (Internò et al., 2025) on the same split, although its effectiveness varies across individual targets.

Overall, our contributions are threefold: (1) We establish a taxonomy of fake video detectors based on their dominant forensic evidence and develop a category-specific attack for each detector family. (2) We integrate these attacks into a unified framework that uses a single perturbation to degrade detectors across different categories, even when the target's detection principle is unknown. (3) Through extensive evaluations across detectors, video generators, and datasets, we demonstrate the effectiveness and transferability of the proposed category-specific and unified attacks.

2 Threat Model

2.1 Adversary's Goal

The adversary aims to learn a single, bounded, and perceptually imperceptible perturbation that makes fake videos to be classified as real to detectors, and generalizes to videos and detector models unseen during optimization.

Attack effectiveness. The perturbation is designed to induce a targeted Fake-to-Real prediction, causing fake videos that are originally identified as fake rightly to be misclassified as real after the perturbation is applied.

Perceptual imperceptibility. The perturbation should introduce minimal visible degradation and preserve the perceptual quality of the original video.

Universality and transferability. The same perturbation should work on fake videos unseen during optimization, without being adapted to each video. It should also transfer from the surrogate detectors used to construct the attacks to unseen target detectors.

2.2 Adversary's Background Knowledge

The adversary has a collection of fake videos and access to white-box surrogate detectors for learning the perturbation. The target detector is black-box: its architecture, parameters, training data, and gradients are unknown. Based on whether the adversary knows the target detector's broad detection paradigm, we consider two settings: **category-aware** and **category-unaware**.

Category-aware setting. The adversary knows the broad detection paradigm used by the target platform, such as whether it mainly relies on spatial forensic artifacts[4], pretrained visual representations[5], or temporal dynamics[6], but does not know the exact deployed detector. This assumption is practically plausible because the broad technical methodology may be available from public technical descriptions or related publications. The adversary therefore selects surrogate detector(s) from the corresponding category to optimize the perturbation and transfers it to the unknown target model.

Category-unaware setting. The adversary has no prior knowledge of the detection paradigm used by the target platform. In this case, a category-specific attack cannot be selected, and the unified attack is used to jointly consider different detector categories. This setting evaluates whether the learned perturbations can transfer to a target detection pipeline whose technical approach is completely unknown.

3 Method

3.1 Overview and Detector Taxonomy

Existing fake video detectors primarily rely on three types of forensic evidence: frame-level generation artifacts, high-level visual representations, and inter-frame dynamics (Corvi et al., 2025; Veeramachaneni et al., 2025; Internò et al., 2025; Zheng et al., 2025). Our experimental analysis in Section 4.3 provides further empirical support for this three-way categorization. Accordingly, we categorize them as **Spatial-Artifact-Oriented (SAO)**, **Semantic-Representation-Oriented (SRO)**, and **Temporal-Dynamics-Oriented (TDO)** detectors. This taxonomy is determined by the dominant evidence used for detection rather than the backbone architecture. Specifically, SAO detectors identify low-level spatial traces, such as abnormal textures and frequency patterns (Corvi et al., 2025; Li et al., 2026). SRO detectors classify videos using representations extracted by pretrained visual encoders (Veeramachaneni et al., 2025; Oquab et al., 2024). TDO detectors capture abnormal changes or feature trajectories across consecutive frames (Internò et al., 2025; Zheng et al., 2025). We detailedly elaborate on the detection logic and differences of various methods in Appendix A.

3.2 Baseline Attacks

We first propose a baseline attack against video detectors based on the principles of traditional attack methods targeting image detectors. However, because detectors in different categories usually expose different adversarial sensitivities, directly optimizing the same conventional perturbation

[4]Sightengine publicly states that its AI-generated video detector analyzes the visual pixel content of videos without relying on metadata: `https://sightengine.com/detect-ai-generated-videos`.

[5]Sensity publicly describes a deep-learning-based detection system that learns visual features from large collections of known deepfake data: `https://sensity.ai/`.

[6]Reality Defender states that its video detection models identify manipulated movements and other signs of video manipulation: `https://www.realitydefender.com/platform/technology`.

formulation for all detectors may lead to limited transferability. We therefore develop a category-specific attack for each detector category and subsequently integrate their complementary properties into a unified attack.

For baseline attacks, let $X = \{x_1, \ldots, x_T\}$ denote a video containing T sampled frames. The common baseline learns a static pixel-space perturbation δ^{base} that is shared across all frames and videos:

$$x_t^{\text{adv}} = \text{clip}_{[0,1]}(x_t + \delta^{\text{base}}), \qquad \|\delta^{\text{base}}\|_\infty \leq \epsilon. \tag{1}$$

Here, x_t denotes the t-th sampled frame of the input video, and x_t^{adv} is the corresponding perturbed frame. ϵ represents the constraint of the ℓ_∞ norm of the adversarial perturbation. $\text{clip}_{[0,1]}(\cdot)$ keeps the resulting pixel values within the valid range [0,1].

The perturbation δ^{base} is optimized using a targeted classification loss that encourages the surrogate detector to predict the real label. All category-specific attacks use the same training videos, perturbation budget, and optimization schedule as this baseline, while changing only the perturbation parameterization or optimization view.

3.3 Category-Specific Universal Attacks

3.3.1 Frequency-Aware Attacks for SAO

A conventional pixel-space attack may overfit to the particular spatial artifacts captured by the surrogate SAO detector. Because different SAO detectors process spatial-frequency evidence differently, such surrogate-specific patterns may transfer poorly across detectors.

To improve transferability, we optimize the perturbation δ^{SAO} through multiple frequency realizations. Specifically, we apply a multi-level wavelet transform to decompose the SAO perturbation δ^{SAO} into a low-frequency component L and M detail components $\{D_1, \ldots, D_M\}$ at progressively coarser spatial scales:

$$(L, D_1, \ldots, D_M) = \mathcal{W}(\delta^{\text{SAO}}), \tag{2}$$

where $\mathcal{W}$ denotes the wavelet transform. L captures the coarse low-frequency structure of the perturbation, while D_1 captures the finest high-frequency details. The subsequent components $D_2, \ldots, D_M$ represent progressively coarser spatial-frequency structures.

At each optimization step, these wavelet components are dynamically reweighted and reconstructed to obtain a frequency-reweighted realization of δ^{SAO}:

$$\delta^{\text{SAO}'} = \mathcal{W}^{-1}\left(\alpha_L L, \alpha_1 D_1, \ldots, \alpha_M D_M\right). \tag{3}$$

where $\mathcal{W}^{-1}$ denotes the inverse wavelet transform, and $\{\alpha_L, \alpha_1, \ldots, \alpha_M\}$ control the contributions of the corresponding frequency scales. Different reweighting coefficients produce different frequency realizations $\delta^{\text{SAO}'}$ from the same perturbation δ^{SAO}. These realizations are evaluated using the SAO attack objective, and the resulting gradients are backpropagated to update δ^{SAO}. Consequently, patterns that are effective only within a narrow frequency range are less likely to dominate the optimization. The learned perturbation is instead encouraged to remain effective across multiple spatial-frequency views, reducing its dependence on the particular forensic evidence captured by the surrogate detector.

3.3.2 Dual-Band Attacks for SRO

SRO detectors classify videos using high-level visual representations. However, a conventional full-resolution perturbation can become dominated by fine-grained pixel patterns that strongly affect the surrogate detector, but have limited influence on representations learned by other visual encoders. We therefore parameterize the static perturbation δ^{dual} using two complementary components: a low-resolution backbone δ_{low} and a full-resolution residual δ_{res}:

$$\delta^{\text{dual}} = \text{clip}_{[-\epsilon,\epsilon]}\left(\text{Up}(\delta_{\text{low}}) + \delta_{\text{res}}\right), \tag{4}$$

where $\text{Up}(\cdot)$ denotes differentiable upsampling, and $\text{clip}_{[-\epsilon,\epsilon]}(\cdot)$ performs element-wise clipping to constrain the perturbation within the ℓ_∞ budget ϵ. The low-resolution branch captures smooth, large-scale structures, while the residual branch retains the flexibility to model finer local variations. Both components are jointly optimized through the reconstructed perturbation δ^{dual}, which remains static and is shared across all sampled frames and videos.

3.3.3 FRAME-SPECIFIC ATTACKS FOR TDO

TDO detectors analyze how visual representations evolve across consecutive frames. Applying the same static perturbation to every frame can modify individual-frame features, but provides limited control over their temporal differences. We therefore learn a frame-specific perturbation sequence Δ^{temp} and construct each adversarial frame as x_t^{adv}.

$$\Delta^{\text{temp}} = \{\delta_1, \delta_2, \ldots, \delta_T\}, \tag{5}$$

$$x_t^{\text{adv}} = \text{clip}_{[0,1]}(x_t + \delta_t), \qquad \|\delta_t\|_\infty \leq \epsilon \tag{6}$$

The complete sequence is universal across videos, *i.e.,* the same perturbation δ_t is applied to the t-th sampled position of every video. Unlike a static perturbation, however, different temporal positions are allowed to use different perturbation patterns.

All frame-specific components are jointly optimized using the video-level TDO attack objective. This parameterization enables the attack to modify both individual-frame representations and their frame-to-frame changes, thereby directly influencing the temporal trajectory perceived by the detector. We further study the effect of the number of frame-specific perturbations in Appendix C.

3.4 UNIFIED STRUCTURED OPTIMIZATION

The category-specific attacks are designed for the category-aware setting, where the broad detection paradigm of the target detector is known. In the category-unaware setting, however, the attacker does not know which type of forensic evidence the target detector primarily relies on. We therefore integrate the complementary properties of the three category-specific attacks into a single structured perturbation δ_t^{uni} for each frame t.

Following Section 3.3.2, the shared static component is constructed from δ_{low} and δ_{res}. We additionally introduce temporal components $\Delta^{\text{temp}} = \{\delta_1, \ldots, \delta_T\}$, where δ_t denotes the perturbation associated with the t-th sampled frame. Because the static component already captures perturbation structures shared across frames, unconstrained temporal components may learn redundant frame-shared patterns. We therefore remove their temporal mean to obtain the centered temporal perturbation for the t-th frame:

$$\hat{\delta}_t = \text{clip}_{[-\epsilon,\epsilon]}\left(\delta_t - \frac{1}{T}\sum_{k=1}^{T}\delta_k\right). \tag{7}$$

This operation encourages the temporal component to focus on position-dependent variations, rather than duplicating the shared static component. Thus, the unified perturbation applied to the t-th frame is

$$\delta_t^{\text{uni}} = \text{clip}_{[-\epsilon,\epsilon]}\left(\text{Up}(\delta_{\text{low}}) + \delta_{\text{res}} + \hat{\delta}_t\right). \tag{8}$$

During optimization, we use the frequency-aware transformation introduced in Section 3.3.1 to improve the transferability of the unified perturbation δ_t^{uni} across different spatial-frequency views. In this way, we can obtain the unified perturbation sequence Δ^{uni}, as follows. The same δ_t^{uni} is evaluated by the SAO, SRO, and TDO surrogate detectors.

$$\Delta^{\text{uni}} = \{\delta_1^{\text{uni}}, \ldots, \delta_T^{\text{uni}}\} \tag{9}$$

Considering directly combining gradients of different categories can cause unstable optimization, we use PCGrad (Yu et al., 2020) to jointly optimize $\{\delta_{\text{low}}, \delta_{\text{res}}, \delta_{\text{temp},1}, \ldots, \delta_{\text{temp},T}\}$. Please see Appendix B for detailed gradient aggregation procedure.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on four fake-video datasets: GenBuster (Wen et al., 2025), which includes ID, OOD, and WILD splits; GenVideo (Chen et al., 2024); DeepTraceReward (DTR) (Fu et al., 2025); and DeCoF (Ma et al., 2025). Together, they cover diverse data distributions and video generators, such as EasyAnimate (Xu et al., 2024), Sora (Brooks et al., 2024), and CogVideoX (Yang et al., 2025).

Table 1: Strict conditional Fake-to-Real flip rates (%) of the matched baseline and category-specific attacks. FAA, DBA, and FSA denote frequency-aware attack, dual-band attack, and frame-specific attack, respectively. Bold values indicate the better result within each detector and dataset. GB denotes GenBuster, with GB-ID, GB-OOD, and GB-WILD referring to its three evaluation splits; DTR denotes DeepTraceReward.

Category	Detector	Attack	GB-ID	GB-OOD	GB-WILD	GenVideo	DTR	DeCoF
SAO	WaveRep	Baseline	86.90	94.60	99.26	72.19	66.36	85.09
		FAA	**99.70**	**99.79**	**100.00**	**98.81**	**98.48**	**99.86**
	Native Scale	Baseline	22.43	13.51	42.31	3.23	7.89	11.01
		FAA	**41.29**	**42.38**	**84.62**	**13.48**	**21.05**	**30.59**
	LGrad	Baseline	14.21	11.48	14.29	8.51	3.16	12.50
		FAA	**64.61**	**70.49**	**80.00**	**59.57**	**46.32**	**76.88**
SRO	CLIP	Baseline	71.55	64.08	77.24	87.06	90.30	95.48
		DBA	**83.64**	**75.48**	**91.06**	**90.00**	**92.12**	**97.04**
	DINOv2	Baseline	19.26	12.30	20.00	11.74	23.78	28.73
		DBA	**24.01**	**23.59**	**24.56**	**16.52**	**25.73**	**29.45**
	ViT-B/16	Baseline	2.73	0.65	4.04	1.48	0.94	5.57
		DBA	**10.41**	**7.27**	**26.26**	**9.48**	**10.97**	**18.92**
	Swin-Tiny	Baseline	4.47	2.09	11.54	10.23	12.90	19.14
		DBA	**20.69**	**13.55**	**31.73**	**20.26**	**25.48**	**32.71**
TDO	ReStraV	Baseline	8.30	13.40	20.90	2.11	15.00	1.71
		FSA	**99.90**	**99.80**	**100.00**	**99.37**	**100.00**	**97.95**
	D3	Baseline	12.30	10.00	17.10	**0.18**	3.14	8.48
		FSA	**88.10**	**89.80**	**92.90**	0.00	**45.30**	**37.19**
	NSG-VD	Baseline	3.90	6.70	0.00	2.78	5.06	3.63
		FSA	**85.10**	**89.00**	**100.00**	**33.60**	**43.67**	**33.87**

Fake video detectors. We evaluate representative detectors from all three categories. For SAO, WaveRep (Corvi et al., 2025) is used as the white-box surrogate detector, while Native Scale (Li et al., 2026) and LGrad (Tan et al., 2023) are used as black-box target detectors. For SRO, CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2024) are used as white-box surrogate detectors, while ViT-B/16 (Dosovitskiy et al., 2021) and Swin-Tiny (Liu et al., 2021) are used as black-box target detectors. For TDO, ReStraV (Internò et al., 2025) is used as the white-box surrogate detector, while D3 (Zheng et al., 2025) and NSG-VD (Zhang et al., 2025) are used as black-box target detectors.

All surrogate detectors are accessible during perturbation optimization, whereas the black-box target detectors are excluded from optimization and checkpoint selection, and are used only to evaluate transferability.

Evaluation metric. We report the strict conditional Fake-to-Real flip rate. Only fake videos that a detector correctly classifies before perturbation are included in the denominator:

$$\mathrm{FR}_{F\to R} = \frac{|\{x : f(x) = \text{Fake} \ \wedge \ f(x+\delta) = \text{Real}\}|}{|\{x : f(x) = \text{Fake}\}|} \times 100\%. \tag{10}$$

This definition prevents naturally misclassified fake videos from being counted as successful attacks. We keep the evaluated videos, clean predictions, detector thresholds, and frame-sampling strategy fixed when comparing attacks.

Implementation Details. All attacks use an ℓ_∞ budget of $\epsilon = 0.02$ in the normalized $[0, 1]$ pixel space, with detector parameters frozen. We optimize for 2,000 iterations using Adam with cosine learning-rate decay. Inputs follow the original detector settings: SAO uses detector-specific preprocessing, SRO uses 224×224 inputs, and TDO uses 24 frames at 224×224.

4.2 Category-Specific Attack Results

In this section, we evaluate the effectiveness of category-specific attacks by comparing each attack with its matched baseline. As reported in Table 1, the proposed SAO attack improves the flip rate for every detector and dataset. The gains are especially pronounced for the black-box LGrad detector, where the improvement ranges from 43.16% to 65.71% across the six datasets. The consistent

Table 2: Strict conditional Fake-to-Real flip rates (%) of the category-specific baselines and the unified attack optimized with conflict-aware gradient aggregation.

Category	Detector	Attack	GB-ID	GB-OOD	GB-WILD	GenVideo	DTR	DeCoF
SAO	WaveRep	Baseline	**86.90**	**94.60**	**99.26**	72.19	**66.36**	**85.09**
		Unified	81.90	92.83	**99.26**	**76.21**	**66.36**	84.81
	Native Scale	Baseline	22.43	13.51	42.31	3.23	7.89	11.01
		Unified	**46.86**	**54.67**	**93.59**	**19.41**	**29.28**	**38.05**
	LGrad	Baseline	14.21	11.48	14.29	8.51	3.16	12.50
		Unified	**94.37**	**96.72**	**100.00**	**85.58**	**90.53**	**95.00**
SRO	CLIP	Baseline	71.55	64.08	77.24	87.06	90.30	95.48
		Unified	**91.36**	**92.28**	**96.48**	**88.50**	**91.82**	**95.79**
	DINOv2	Baseline	19.26	**12.30**	20.00	11.74	**23.78**	**28.73**
		Unified	**24.89**	9.23	**31.25**	**13.10**	19.87	24.00
	ViT-B/16	Baseline	2.73	0.65	4.04	1.48	0.94	5.57
		Unified	**4.84**	**2.51**	**12.60**	**3.35**	**4.39**	**10.14**
	Swin-Tiny	Baseline	4.47	2.09	11.54	10.23	12.90	19.14
		Unified	**20.55**	**13.09**	**31.19**	**12.15**	**16.45**	**22.68**
TDO	ReStraV	Baseline	8.30	13.40	20.90	2.11	15.00	1.71
		Unified	**64.32**	**73.28**	**83.72**	**65.30**	**81.91**	**61.20**
	D3	Baseline	12.30	10.00	**17.10**	0.18	3.14	**8.48**
		Unified	**21.73**	**20.26**	11.43	**0.90**	**12.53**	2.12
	NSG-VD	Baseline	3.90	6.70	0.00	2.78	5.06	3.63
		Unified	**9.04**	**12.20**	**11.11**	**4.96**	**7.50**	**8.40**

gains on Native Scale and LGrad indicate that optimizing the perturbation through diverse frequency realizations reduces its dependence on the particular spatial cues used by WaveRep.

The SRO attack also improves all four representation-based detectors across all datasets. Improvements on the white-box CLIP and DINOv2 classifiers show that the dual-band parameterization retains attack strength on the optimization models. The absolute flip rates on these black-box models remain lower than those on CLIP, suggesting that semantic-representation transfer is more sensitive to the feature space and training procedure of the target detector.

For TDO detectors, replacing a frame-repeated perturbation with a frame-specific temporal perturbation produces the largest overall improvement. The attack reaches near-saturated flip rates on ReStraV across all six domains and transfers strongly to NSG-VD. It also substantially improves transfer to D3 in five of the six datasets. Overall, these results support the central premise of our approach: explicitly aligning the perturbation parameterization with the dominant forensic evidence improves both in-category effectiveness and black-box transfer. Additional ablations on the temporal perturbation length and perturbation budget are provided in Appendix C, with the corresponding results shown in Figure 3 and Figure 4.

4.3 Transfer Mechanism Analysis

In this section, we provide complementary evidence to explain the transfer behavior of the three category-specific attacks. As shown in Figure 2(a), for SAO, we visualize the baseline perturbation δ^{base} and the perturbation δ^{SAO} learned by the frequency-aware attack. We observe that δ^{base} appears as spatially diffuse pixel noise, whereas δ^{SAO} exhibits more location-selective and coherent structures. This structural change is consistent with reduced reliance on surrogate-specific local gradients and may contribute to improved transferability across different SAO detectors.

For SRO, as shown in Figure 2(b), we compare the normalized radial spectral energy distributions of the baseline perturbation δ^{base} and the dual-band perturbation δ^{dual}, so as to analyze how their energy is distributed across spatial-frequency bands. We observe that δ^{dual} concentrates more energy in the low-frequency region than δ^{base}. Such low-frequency perturbations have been shown to improve black-box transferability across models (Sharma et al., 2019), providing a plausible explanation for the consistent gains in Table 1.

For TDO, as shown in Figure 2(c), we follow the trajectory analysis of ReStraV (Internò et al., 2025) to visualize the distributions of eight temporal-geometric features, including the mean, minimum,

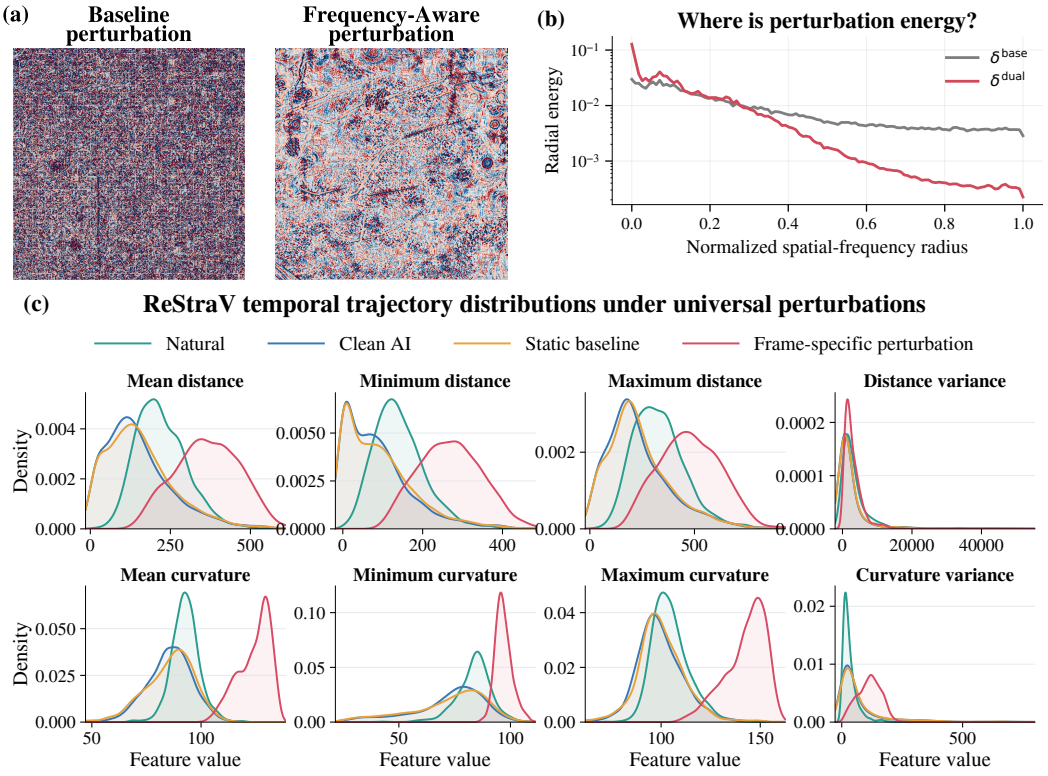

Figure 2: Mechanism-level analysis of category-oriented adversarial transfer: (a) spatial structures of SAO perturbations; (b) radial spectral energy of SRO perturbations; (c) ReStraV trajectory features under different perturbations.

maximum, and variance of inter-frame distance and curvature. We compare four conditions: natural videos (green), clean AI-generated videos (blue), AI-generated videos with the static baseline perturbation (orange), and AI-generated videos with our frame-specific temporal perturbation (red). We observe that the static baseline distributions largely overlap with those of clean AI-generated videos, indicating that a temporally shared perturbation has limited influence on the representation trajectories used by ReStraV. In contrast, the temporal perturbation produces substantial shifts across multiple distance- and curvature-based features, moving the distributions away from clean AI-generated videos and generally toward, or in some cases beyond, the natural-video distributions.

4.4 Unified Attack under the Category-Unaware Setting

In this section, we evaluate the effectiveness of the single perturbation learned by the unified procedure in Section 3.4. Table 2 compares it with the corresponding category-specific baselines in Table 1. This experiment models the category-unaware setting, where the attacker cannot select an attack according to the target detector's forensic principle. We discover that the joint perturbation shows strong cross-domain transferability for WaveRep, LGrad, and ReStraV, while its transferability to representation detectors and heterogeneous temporal detectors is more limited. Native Scale also exhibits substantial domain dependence. These results demonstrate that the unified attack can simultaneously cover all three forensic-evidence categories, while the attack strength for each individual category is not fully preserved under multi-objective optimization. This indicates a trade-off between broad cross-category transferability and category-specific attack strength. The unified attack remains effective on SAO detectors and ReStraV, while the improvements on SRO detectors and some TDO detectors are smaller. Notably, the relatively low performance of NSG-VD on GenBuster WILD is primarily attributable to the small number of fake videos that are correctly identified by the detector.

4.5 Adaptive Defense

We further evaluate whether a simple test-time defense can mitigate the proposed attacks. Specifically, we use an adaptive JPEG consistency defense that combines the detector responses to the original input and a strongly compressed JPEG view. The defense reduces the Fake-to-Real flip rate to 0% for 8 of the 11 evaluated detectors, while the remaining three are reduced to at most 1.22%. However, this robustness comes at a substantial cost to clean detection performance, with Real accuracy decreasing by 44.28–99.70 percentage points. Detailed results and analysis are provided in Appendix D.

5 Related Work

5.1 Universal adversarial perturbations

Universal adversarial perturbations show that a single input-agnostic perturbation can induce prediction errors on many previously unseen inputs without per-sample optimization (Moosavi-Dezfooli et al., 2017). Subsequent studies have further demonstrated that adversarial examples optimized on accessible surrogate models can transfer to unknown target models in black-box settings (Papernot et al., 2017; Liu et al., 2017; Dong et al., 2018; Xie et al., 2019; Dong et al., 2019; Naseer et al., 2019). These findings motivate transferable universal perturbations against fake-video detectors. However, conventional formulations typically optimize a single pixel-space perturbation for one classifier or a set of closely related models. They do not explicitly account for the fact that fake-video detectors can rely on fundamentally different spatial, representation-level, and temporal evidence. In contrast, we study how the dominant forensic evidence of a detector affects transferability and design category-specific perturbation parameterizations accordingly.

5.2 Adversarial attacks against synthetic-media detectors

Previous work has revealed adversarial vulnerabilities in detectors for GAN-generated images and manipulated facial media, including white-box evasion, transfer-based attacks, and universal perturbations (Carlini & Farid, 2020; Gandhi & Jain, 2020; Neekhara et al., 2021; Diao et al., 2024; Saberi et al., 2024; Mavali et al., 2024). These studies establish that synthetic-media detectors can be highly sensitive to visually bounded perturbations. Nevertheless, their primary focus is image-level detection or frame-based DeepFake detection, where the decision is largely determined by the appearance of an individual image or face frame. General AI-generated video detectors introduce a broader range of detection mechanisms: some preserve subtle spatial generation traces, some classify pretrained visual representations, and others model feature trajectories across frames (Corvi et al., 2025; Li et al., 2026; Internò et al., 2025; Zheng et al., 2025; Zhang et al., 2025).

Our work differs by evaluating targeted Fake-to-Real perturbations across these heterogeneous detection paradigms, including transfer to held-out detectors, unseen generation distributions, and external video datasets.

6 Conclusion

We studied universal Fake-to-Real attacks against three complementary categories of fake video detectors. Controlled category-wise comparisons show that frequency-aware spatial optimization, dual-band static perturbation, and frame-specific temporal parameterization generally improve transfer over their respective baselines. The unified experiment further demonstrates that one bounded unified perturbation can simultaneously affect all three forensic evidence categories across unseen generators, datasets, and detector architectures.

The experiments also reveal a breadth–strength trade-off. Joint optimization retains strong spatial transfer and attacks ReStraV reliably, but is less effective against heterogeneous representation and temporal black boxes than independently optimized category-specific attacks. Future work should improve multi-objective balancing, evaluate multiple source generators, and report perceptual quality together with full clean and perturbed detection metrics.

AI USE STATEMENT

In this work, we use generative AI tools exclusively for language polishing. We have not used generative AI tools for information retrieval, discovery, or research ideation.

ETHICS AND REPRODUCIBILITY STATEMENT

This paper has studied universal Fake-to-Real attacks against three complementary categories of fake video detectors. There are no ethical issues with this paper.

Besides, we have provided technical details for SAO, SRO, and TDO attacks in Section 3.3.1, Section 3.3.2, and Section 3.3.3, and provided experimental details in Section 4.1, Appendix B. The code will be released when the paper is accepted.

REFERENCES

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI Technical Report, 2024. February 2024.

Nicholas Carlini and Hany Farid. Evading deepfake-image detectors with white- and black-box attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: AI-generated video detection on million-scale GenVideo benchmark. *arXiv preprint arXiv:2405.19707*, 2024.

Riccardo Corvi, Davide Cozzolino, Ekta Prashnani, Shalini De Mello, Koki Nagano, and Luisa Verdoliva. Seeing what matters: Generalizable AI-generated video detection with forensic-oriented augmentation. In *Advances in Neural Information Processing Systems*, 2025.

Yunfeng Diao, Naixin Zhai, Changtao Miao, Xun Yang, and Meng Wang. Vulnerabilities in AI-generated image detection: The challenge of adversarial attacks. *arXiv preprint arXiv:2407.20836*, 2024.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Xingyu Fu, Siyi Liu, Yinuo Xu, Pan Lu, Guangqiuse Hu, Tianbo Yang, Taran Anantasagar, Christopher Shen, Yikai Mao, Yuanzhe Liu, Keyush Shah, Chung Un Lee, Yejin Choi, James Zou, Dan Roth, and Chris Callison-Burch. Learning human-perceived fakeness in ai-generated videos via multimodal llms. *arXiv preprint arXiv:2509.22646*, 2025.

Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors. *arXiv preprint arXiv:2003.10596*, 2020.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `https://arxiv.org/abs/1412.6572`.

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.

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

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Long Ma, Zhiyuan Yan, Qinglang Guo, Yong Liao, Haiyang Yu, and Pengyuan Zhou. Detecting AI-generated video via frame consistency. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025. doi: 10.1109/ICME59968.2025.11210049.

Sina Mavali, Jonas Ricker, David Pape, Yash Sharma, Asja Fischer, and Lea Schönherr. Fake it until you break it: On the adversarial robustness of ai-generated image detectors. *arXiv preprint arXiv:2410.01574*, 2024.

Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys*, 54(1):1–41, 2021. doi: 10.1145/3425780.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Muhammad Muzammal Naseer, Salman H. Khan, Muhammad Haris Khan, Fahad Shahbaz Khan, and Fatih Porikli. Cross-domain transferability of adversarial perturbations. In *Advances in Neural Information Processing Systems*, 2019.

Paarth Neekhara, Brian Dolhansky, Joanna Bitton, and Cristian Canton Ferrer. Adversarial threats to deepfake detection: A practical perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2021.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Mehrdad Saberi, Vinu Sankar Sadasivan, Keivan Rezaei, Aounon Kumar, Atoosa Chegini, Wenxiao Wang, and Soheil Feizi. Robustness of ai-image detectors: Fundamental limits and practical attacks. In *International Conference on Learning Representations*, 2024.

Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 3389–3396, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL https://arxiv.org/abs/1312.6199.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12105–12114, 2023.

Keerthi Veeramachaneni, Praveen Tirupattur, Amrit Singh Bedi, and Mubarak Shah. Leveraging pre-trained visual models for ai-generated video detection. *arXiv preprint arXiv:2507.13224*, 2025.

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

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

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.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, 2025.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

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.

Chende Zheng, Ruiqi Suo, Chenhao Lin, Zhengyu Zhao, Le Yang, Shuai Liu, Minghui Yang, Cong Wang, and Chao Shen. D3: Training-free AI-generated video detection using second-order features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12852–12862, 2025.

A Detailed Explanation of Different Detectors

Existing fake video detectors differ substantially in the forensic cues they rely on. Some methods identify local artifacts within individual frames, some determine whether the visual content is consistent with representations learned from large-scale real images, while others detect abnormal changes across consecutive frames. Such differences make it difficult for a single attack strategy to effectively attack all detectors. Based on their dominant detection principles, we broadly categorize existing fake video detection methods into three categories: **Spatial-Artifact-Oriented (SAO)**, **Semantic-Representation-Oriented (SRO)**, and **Temporal-Dynamics-Oriented (TDO)**.

Spatial-Artifact-Oriented (SAO). SAO methods mainly detect fake videos by identifying spatial artifacts within individual frames. These artifacts are often introduced during video generation and may appear as abnormal textures, frequency patterns, local inconsistencies, or other low-level traces that distinguish generated frames from real ones. Therefore, the detection decision is primarily determined by the spatial characteristics of each frame, while temporal relationships between frames play a limited role. Representative methods in this category include WaveRep (Corvi et al., 2025), Native Scale (Li et al., 2026), and LGrad (Tan et al., 2023).

Semantic-Representation-Oriented (SRO). SRO methods rely on high-level visual representations rather than explicitly designed forensic artifacts. They typically employ powerful pretrained visual encoders, such as CLIP, DINOv2, or other vision transformers, to extract semantic representations from video frames and train classifiers to distinguish real and fake content in the resulting feature space. Compared with SAO methods, their decisions are less dependent on a particular low-level artifact and more closely related to how generated content is represented by general-purpose visual models. Representative detectors include classifiers built on CLIP (Radford et al., 2021), DINOv2 (Oquab et al., 2024), ViT (Dosovitskiy et al., 2021), and Swin (Liu et al., 2021) representations.

Temporal-Dynamics-Oriented (TDO). TDO methods exploit the temporal dimension of videos. Instead of treating frames independently, they examine how visual representations, objects, or generation artifacts evolve across consecutive frames. Although modern video generators can produce visually realistic individual frames, maintaining consistent temporal dynamics remains challenging. Consequently, abnormal frame-to-frame changes, representation trajectories, or temporal geometric patterns can provide useful evidence for detecting fake videos. Representative methods include ReStraV, D3, and NSG-VD (Internò et al., 2025; Zheng et al., 2025; Zhang et al., 2025).

B Details of Conflict-Aware Joint Optimization

The unified attack jointly considers the attack objectives of the three detector categories. Let $c \in \{\mathrm{SAO}, \mathrm{SRO}, \mathrm{TDO}\}$ denote a detector category, and let $\mathcal{L}_c$ denote the corresponding category-specific attack objective. Specifically, $\mathcal{L}_{\mathrm{SAO}}$, $\mathcal{L}_{\mathrm{SRO}}$, and $\mathcal{L}_{\mathrm{TDO}}$ represent the attack losses evaluated on the SAO, SRO, and TDO surrogate detectors, respectively.

Let $\theta = \{\delta_{\mathrm{low}}, \delta_{\mathrm{res}}, \delta_{\mathrm{temp},1}, \ldots, \delta_{\mathrm{temp},T}\}$ denote the learnable parameters of the unified perturbation. For each detector category c, we compute the gradient of its attack objective with respect to the unified perturbation parameters:

$$g_c = \nabla_\theta \mathcal{L}_c, \tag{11}$$

where g_c denotes the update direction induced by category c.

Because the gradient magnitudes of different detector categories may vary substantially, we normalize each category gradient as

$$\bar{g}_c = \frac{g_c}{\|g_c\|_2 + \eta}, \tag{12}$$

where $\bar{g}_c$ is the normalized gradient and η is a small constant for numerical stability.

We then detect gradient conflicts using the inner product between normalized category gradients. For two detector categories i and j, if

$$\bar{g}_i^\top \bar{g}_j < 0, \tag{13}$$

their gradients point in conflicting directions. Following PCGrad (Yu et al., 2020), we remove from $\bar{g}_i$ the component that opposes $\bar{g}_j$:

$$\bar{g}_i \leftarrow \bar{g}_i - \frac{\bar{g}_i^\top \bar{g}_j}{\|\bar{g}_j\|_2^2 + \eta} \bar{g}_j. \tag{14}$$

After resolving pairwise conflicts, the adjusted category gradients are aggregated as

$$g_{\text{uni}} = \frac{1}{3} \sum_{c \in \{\text{SAO,SRO,TDO}\}} \bar{g}_c, \tag{15}$$

where g_{uni} denotes the final gradient used to update the unified perturbation parameters θ.

This procedure prevents a single detector category from dominating the optimization solely because of its gradient magnitude and reduces destructive interference among the three attack objectives.

C Additional Experimental Results

We conduct ablation studies on two key design choices: the number of independently learned temporal perturbation frames T and the perturbation budget ϵ. Unless otherwise specified, we report the strict conditional Fake-to-Real flip rate on the white-box detectors. For SRO, we report the mean flip rate of CLIP and DINOv2.

Effect of temporal perturbation length. Figure 3 varies the number of independently learned perturbation frames from $T = 1$ to $T = 32$. When $T = 1$, the perturbation reduces to a temporally shared static level. The TDO surrogate ReStraV benefits substantially from introducing temporal degrees of freedom: its flip rate increases from approximately 15% at $T = 1$ to over 92% at $T = 8$, reaches about 98% at $T = 16$, and saturates near 100% for $T \geq 24$. This result confirms that a frame-specific parameterization is necessary for effectively manipulating the temporal trajectories used by TDO detectors.

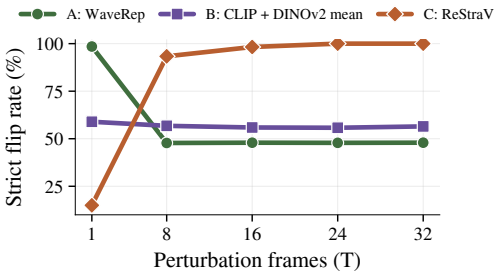

Figure 3: Ablation on temporal length T. ReStraV benefits substantially from frame-specific perturbations, whereas SAO and SRO detectors do not.

In contrast, increasing T does not improve the two frame-oriented detector categories. WaveRep achieves its highest flip rate with a static perturbation ($T = 1$), while its performance decreases to approximately 48% once frame-specific perturbations are introduced. The mean SRO performance of CLIP and DINOv2 remains stable, varying only from roughly 56% to 59% across all temporal lengths. These results are consistent with our category-oriented design: SAO and SRO detectors primarily rely on frame-level evidence and therefore do not require additional temporal variation.

Effect of perturbation budget. Figure 4 evaluates the sensitivity of the three detector categories to the ℓ_∞ perturbation budget on DeepTraceReward and DeCoF. As expected, larger budgets generally improve attack effectiveness. However, the sensitivity differs substantially across detector categories.

For WaveRep and ReStraV, a moderate budget of $\epsilon = 0.02$ already produces near-saturated flip rates on both datasets. On DeepTraceReward, WaveRep increases from about 56% at $\epsilon = 0.005$ to approximately 99% at $\epsilon = 0.02$, while ReStraV increases from 87% to nearly 100%. A similar trend is observed on DeCoF, where both detectors reach approximately 98%–100% at $\epsilon = 0.02$. Increasing the budget beyond 0.02 produces only marginal additional gains for these two categories.

This result suggests that changing high-level visual representations is more difficult than manipulating spatial artifacts or temporal trajectories under a small perturbation budget. Nevertheless, $\epsilon = 0.02$ provides a favorable trade-off between attack strength and perceptual imperceptibility: it already yields near-saturated SAO and TDO attacks while retaining meaningful SRO effectiveness. We therefore use $\epsilon = 0.02$ in all main experiments.

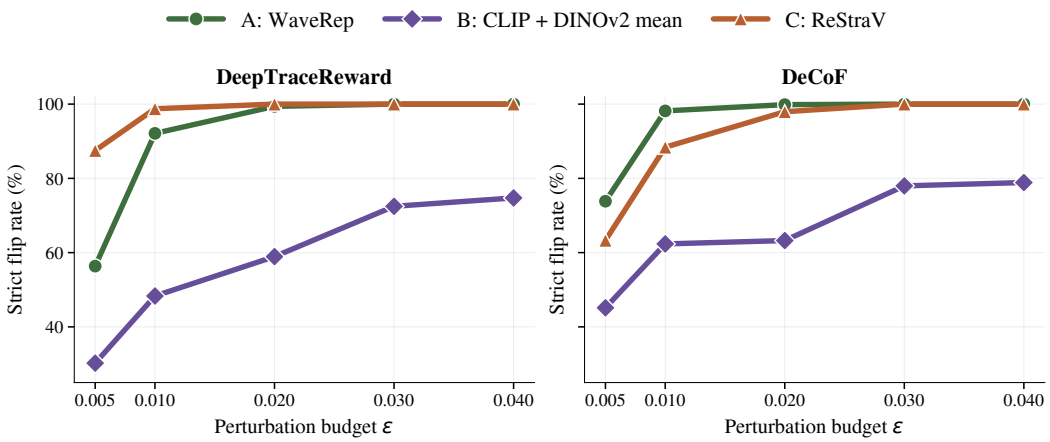

Figure 4: Sensitivity to the perturbation budget ϵ on DeepTraceReward and DeCoF. Larger budgets generally improve targeted Fake-to-Real flip rates, but SAO and TDO detectors saturate around $\epsilon = 0.02$, whereas SRO detectors remain more budget-sensitive.

Table 3: Changes in clean detection accuracy (percentage points) after applying the adaptive defense on dataset. ↑ and ↓ denote accuracy increases and decreases, respectively.

Category	Detector	Fake Acc. Change	Real Acc. Change
SAO	WaveRep	↑0.60	↓82.83
	LGrad	↑71.39	↓98.80
	FreDect	↑88.86	↓88.55
	Native Scale	↑8.43	↓99.70
SRO	CLIP	↑0.60	↓75.00
	DINOv2	↑7.53	↓81.02
	Swin-Tiny	↑6.63	↓86.75
	Transformer	↑3.02	↓56.93
TDO	ReStraV	↑75.90	↓67.47
	D3	↑13.25	↓44.28
	NSG-VD	↑52.41	↓93.07

D Adaptive Defenses

We further evaluate a simple test-time adaptive JPEG consistency defense against the proposed attacks. For each input, the defense additionally constructs a strongly compressed JPEG view and evaluates both the original and compressed inputs using the same detector. The two predictions are converted into Fake evidence relative to the detector's original decision threshold, and the final decision uses the stronger evidence after adjusting the JPEG branch. In this sense, the defense is adaptive at inference time: when adversarial perturbations weaken the Fake evidence on the original input, the compressed view can provide stronger evidence and override the raw prediction without retraining the detector.

The defense substantially reduces the strict Fake-to-Real flip rates on the evaluated detectors. Among the 11 detectors, the flip rate is reduced to 0% for 8 detectors, while the remaining three are reduced to at most 1.22%. However, this robustness improvement comes at a considerable cost to clean detection performance. As shown in Table 3, Fake accuracy increases for all evaluated detectors, whereas Real accuracy decreases by 44.28–99.70 percentage points. The balanced accuracy also decreases for 9 of the 11 detectors, indicating that the defense becomes strongly biased toward predicting Fake.

These results show that a simple JPEG-based input transformation can substantially suppress the proposed attacks, but much of this robustness is obtained by shifting the detector toward more conservative Fake predictions. Therefore, the resulting loss in clean detection utility remains substan-

tial. More effective defenses that improve adversarial robustness without sacrificing clean detection performance are still needed.

E Qualitative Visualization of Adversarial Videos

Figure 5 compares clean and adversarial frames from representative fake videos. Across different video contents and temporal positions, the perturbations introduce only subtle visual changes while successfully modifying the inputs used for detector evaluation.

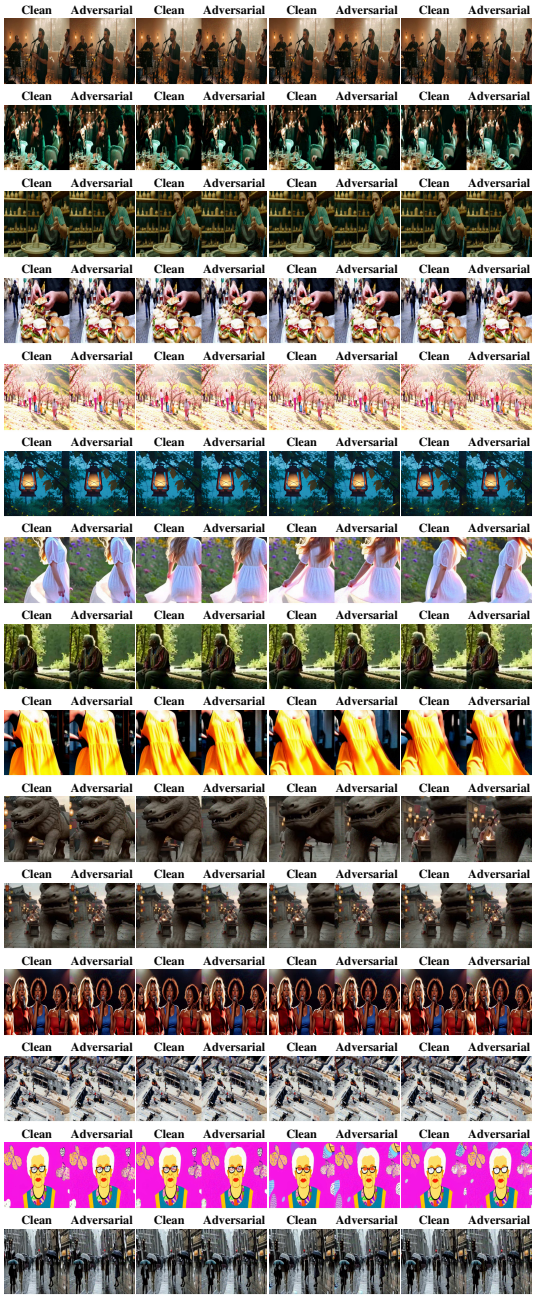

Figure 5: Qualitative comparison between clean and adversarial video frames. Each row corresponds to one video, and four sampled temporal positions are shown from left to right. For each position, the clean frame and its adversarial counterpart are displayed side by side.