

# SoK: The Practical Robustness Frontier: Cost-Aware Vulnerability Analysis for Deepfake Detectors

Shaikh Akib Shahriyar<sup>1</sup>, Ben Brown<sup>1</sup>, Armon Barton<sup>2</sup>, Ning Zhang<sup>3</sup>, and Matthew Wright<sup>1</sup>

<sup>1</sup>Rochester Institute of Technology, Rochester, NY, USA

<sup>2</sup>Naval Postgraduate School, Monterey, CA, USA

<sup>3</sup>Washington University in St. Louis, St. Louis, MO, USA

## Abstract

Deepfake detection is the ideal attacker use case for adversarial examples: digital data that is meant to fool both humans and detection algorithms alike. Adversarial robustness studies, however, are limited and fragmented across isolated attacks, detector families and defenses. In this SoK, we systematize adversarial robustness in deepfake detection through a cost-aware evaluation framework. We evaluate 12 detectors across four architectural families – Naive CNNs, Spatial, Frequency and Video – under a tiered threat model spanning zero-query transferability, low-query black-box attacks, high-query attacks and adaptive white-box attacks. To compare these regimes, we introduce the *Practical Robustness Frontier (PRF)*, which measures attack success as a function of query budget and use thresholds to quantify when detectors become operationally vulnerable.

Our analysis reveals several architecture-dependent failure modes. Zero-query transfer is strong enough to threaten closed-API deployment, showing that opacity and rate limits alone are insufficient. Under low-query attacks, frequency-aware detectors break earliest, suggesting that explicit spectral cues can become efficient attack surfaces. In contrast, temporally coupled video architectures tend to exhibit greater robustness. We further evaluate adversarial training, boundary shaping and generative purification defenses, showing that none of these provide robustness given moderate query budgets. Overall, our findings argue that future deepfake robustness evaluations should move beyond single-scalar metrics and report cost-aware attack frontiers under realistic deployment constraints.

## 1 Introduction

The democratization of generative AI has catalyzed an arms race in media forensics. Due to increasing realism and access, deepfakes are becoming widely used and highly effective for social engineering, biometric fraud and creation of non-consensual synthetic media [2, 28, 62]. In response, the

security community has developed many deepfake detection methods utilizing diverse spatial, spectral and spatiotemporal information [11, 33, 76].

Despite significant architectural advancements, the evaluation of deepfake detectors remains misaligned with real-world security threats. Recent, comprehensive Systematization of Knowledge (SoK) efforts [11, 35] have extensively mapped the *generalization gap*, evaluating how detectors degrade when exposed to novel, out-of-distribution generation tools (e.g., training on FaceSwap videos and testing on Diffusion-based fakes). Even if a detector is robust to new types of deepfakes, however, it can still be vulnerable to adversarial examples. In fact, deepfake detection is one of the most realistic applications for adversarial examples: digital-only, high-dimensional data, where misclassification can lead to massive benefits to the attacker, such as facilitating social engineering and fraud.

Prior work has shown that deepfake detectors are vulnerable to adversarial examples [25, 33, 64]. However, they leave a critical question open: how much access is required to make evasion practical? A detector that fails under a strong white-box attack, such as AutoAttack [16] at  $\epsilon = 8/255$ , and a detector that fails after a few black-box queries present different operational risks [26, 60, 66]. Against deployed detection systems, adversaries will likely face defenses seeking to prevent attackers from using the detector as an oracle to help create videos that bypass it. These may include hiding classifier outputs other than real/fake binary labels, rate limits on using the detector and monitoring of unusual query patterns [3, 26]. Even unprotected systems would use rate limits to keep costs constrained or charge users per query. These constraints make the *accessibility* of a vulnerability central: two detectors may have similar white-box failure rates, yet one may approach that failure ceiling after tens of black-box queries while another may require thousands. Robustness evaluation thus should not only test *whether* a vulnerability exists but also *how much feedback and interaction are required to exploit it*.

To bridge this gap between benchmark robustness and op-

erational security, we introduce the **Practical Robustness Frontier (PRF)**. The PRF maps Attack Success Rate (ASR) as a function of the adversary’s query budget, allowing robustness to be evaluated in terms of attacker access to feedback rather than a single success rate. We summarize the PRF using three classes of diagnostics:

1. **Zero-Query Transfer:** The ASR achieved by adversarial examples crafted on a surrogate model and evaluated on the target without any target queries. This measures whether closed-API restrictions can be bypassed.
2. **Early-Break Thresholds:** The query budgets at which ASR first reaches *Leakage* (5%), *Practical Break* (20%) and *Catastrophic Break* (50%), capturing when evasion becomes measurable, operationally significant, or severe.
3. **High-Query Convergence ( $Q_{95}$ ):** The query budget required for a black-box attacker to reach 95% of the model’s white-box failure ceiling, measuring how quickly black-box attacks approach the strongest measured white-box vulnerability.

**Research Questions (RQs).** This SoK pivots the scientific focus from the binary question of “Can this detector be broken?” to the operational question of “How efficiently can it be broken and how do we maximize that cost?” We highlight that cost here means the ability to bypass rate-limits and other defenses rather than just effort or time. We systematically evaluate the practical robustness of deepfake detection guided by three core research questions:

- **RQ1:** How do different detector architectures shape adversarial vulnerability across zero-query, low-query and high-query threat models?
- **RQ2:** What does a cost-aware evaluation reveal that is hidden by single-point attack success rates?
- **RQ3:** How do representative defenses shift the Practical Robustness Frontier under query-bounded attacks?

**Contributions.** To address these questions, we make the following contributions:

1. **A systematized evaluation landscape for adversarial deepfake detection.** We organize the evaluation space around 12 detectors, four types of attack and eight representative defenses. The detector suite spans four architectural families – Naive CNN, Spatial, Frequency and Video – while the attack regimes cover zero-query transfer, low-query black-box attacks, high-query convergence and white-box reference attacks. The defense suite covers adversarial training, boundary shaping and generative purification. This systematization enables controlled comparison across detector designs, attacker access levels and defense mechanisms.

2. **Practical Robustness Frontier (PRF).** We introduce a cost-aware robustness framework that reports attack success as a function of query budget. We define early-break thresholds for low-query attacks and use  $Q_{95}$  to measure how quickly black-box attacks approach a white-box reference.
3. **Cost-aware vulnerability findings across detector types.** Our evaluation shows that zero-query transfer attacks already threaten closed-API detectors, frequency-aware models expose efficient low-query attack directions and video architectures slow black-box convergence more effectively than frame-level detectors.
4. **Defenses tested under operational constraints.** We show that all tested defenses – adversarial training, boundary shaping and purification – fail to hold up against black-box attacks with modest query budgets, in most cases taking robust AUC to worse than random guessing within 1,500 queries. This puts the onus on systems to prevent adversarial queries as much as possible instead of relying on such defenses.

## 2 Background and Related Work

### 2.1 Deepfake Generation and Artifacts

Deepfake generation methods differ in the artifacts they tend to leave and these artifacts have shaped the design of forensic detectors [70, 74]. Early autoencoders such as DeepFaceLab [53] often require alignment, warping, blending and post-processing steps that can introduce spatial boundary artifacts and temporal flicker [37]. GAN-based synthesis improved photorealism, but common generator and upsampling architectures have been shown to leave spectral fingerprints and frequency-domain irregularities [19, 20, 30]. Diffusion and latent-diffusion models avoid some GAN-specific upsampling artifacts, but may introduce other detectable traces, including reconstruction-error cues exploited by DIRE [57, 78]. More broadly, synthetic media may exhibit geometric, semantic, or temporal inconsistencies when the generation pipeline does not explicitly enforce those constraints [23, 50, 52]. These shifting artifacts motivate our analysis of how detector inductive biases affect adversarial robustness across attacks.

### 2.2 Deepfake Detection

Deepfake detectors differ in the forensic evidence they are designed, or trained, to exploit. Some detectors learn frame-level RGB cues from manipulated images [59, 68], while others encode stronger priors about spatial artifacts such as blending, warping, or texture inconsistencies [38], or about spectral traces left by synthesis pipelines [43, 55]. Prior work has shown that detectors can achieve strong in-distribution

accuracy while relying on non-causal shortcuts tied to generator, compression, identity, or acquisition artifacts, limiting cross-dataset generalization [69, 79].

For adversarial robustness, these forensic cues matter because they also shape the attacker’s optimization problem. Explicit priors can improve detection when the expected artifact is present, but they can also create predictable attack surfaces when the attack basis aligns with the detector’s cue space. Conversely, detectors that aggregate evidence across time may impose additional constraints on adversarial perturbations by requiring a sequence-level decision to be fooled. Section 3.1 formalizes this intuition into the detector taxonomy used throughout our PRF evaluation.

### 2.3 Adversarial Attacks on Vision Models

Adversarial examples are inputs modified by a small perturbation  $\delta$ , often constrained by  $\|\delta\|_p \leq \epsilon$ , to change a model’s prediction while preserving perceptual similarity [22, 44, 67]. Attacks are commonly distinguished by the adversary’s access to the target model. In white-box attacks, the adversary has access to model parameters and gradients, enabling direct gradient-based optimization [9, 44]. In black-box attacks, the adversary relies on queries, which may return confidence scores, logits, or only hard labels [12, 24, 26]. Transfer attacks are an even lower-interaction case, where adversarial examples are crafted on a surrogate model and evaluated on the target without target queries [42, 51, 72].

### 2.4 Adversarial Defenses

Adversarial defenses generally try to either harden the classifier, reshape the decision space, or transform inputs before classification. Adversarial training hardens the classifier by training on adversarially perturbed examples, but often introduces a trade-off between clean and robust performance [44, 73]. Other defenses modify the decision geometry by adding auxiliary regions or labels that make boundary crossing more difficult [6, 36, 65], while purification defenses preprocess inputs to remove adversarial perturbations before classification [49, 61, 63, 65].

These approaches raise different evaluation challenges. Training-based defenses should be assessed not only by aggregate robust accuracy, but also by false accepts, clean utility and query-bounded robustness. Transformation-based and stochastic defenses require adaptive evaluation, since apparent robustness can result from gradient obfuscation rather than genuine resistance [4, 8, 71]. Section 6 evaluates representative defenses from each family under the PRF framework.

### 2.5 Related SoK and Survey Works

Prior surveys and benchmarks have made substantial progress in organizing the deepfake detection literature, but their

primary focus has been generalization rather than adversarial robustness. Foundational surveys catalog manipulation techniques, datasets and forensic countermeasures [29, 45, 70]. More recent benchmarking efforts evaluate detector performance across datasets, manipulation methods and deployment-relevant base-rate assumptions [34, 35]. These works are complementary to ours: they clarify how detectors behave under distribution shift and realistic prevalence, but they do not address adversarial evasion.

A smaller body of work directly studies adversarial attacks on deepfake detectors, including white-box, black-box and transfer attacks [21, 25, 75]. These studies establish that detector vulnerabilities are not merely hypothetical: adversarial examples can transfer across detectors and black-box attacks can reduce detector performance. However, they generally report attack success under a fixed setting, rather than measuring how success changes with attacker access and query budget. As a result, they do not show how quickly different detector families fail, how much querying is needed to approach worst-case failure, or whether defenses remain effective against adaptive attacks.

Our SoK addresses this gap by evaluating adversarial robustness as a costed frontier rather than a single attack success rate. We systematize 12 detector architectures across four families and evaluate attacks and defenses under a tiered threat model. This framing allows us to compare not only whether detectors and defenses fail, but how much attacker interaction is required to make them fail.

## 3 Systematization of Deepfake Detectors

To evaluate adversarial robustness across detector designs, we first define a controlled suite of target models. We aim to select models that represent distinct forensic assumptions while remaining reproducible under a common training and evaluation pipeline. This section describes the taxonomy and selection criteria used to construct that suite.

Our scope is visual face-forgery detection rather than synthetic-media detection in full generality. We use this setting because it offers established datasets, reproducible detector implementations and a diverse set of architectural families that can be evaluated under a common adversarial protocol. This scope is also important in practice: face-swap and face-forgery attacks are prominent in remote identity verification and online abuse, including biometric fraud and non-consensual synthetic media [2, 28]. Future work should extend PRF-style evaluation to newer diffusion- and transformer-based generators, long-form video synthesis and multimodal audio-visual forgeries.

### 3.1 A Bias-Diverse Detector Suite

Comprehensive frameworks such as DeepfakeBench [80] help standardize deepfake detection evaluation, but using many

Table 1: **Taxonomy of Target Detection Architectures.** We group 12 selected detectors into four bias-diverse families used throughout our PRF evaluation. FF++ AUC reports clean performance of our implementations on the FaceForensics++ test split under the implementation and training protocol described in Section 4.

| Focus of Methodology       | Detector Name        | Venue | Year | Key Techniques                                                     | FF++ Clean AUC |
|----------------------------|----------------------|-------|------|--------------------------------------------------------------------|----------------|
| <b>Naive CNNs</b>          | Xception [59]        | ICCV  | 2019 | Depthwise separable CNN without explicit forensic preprocessing    | 0.976          |
|                            | EfficientNet-B4 [68] | ICML  | 2019 | Compound-scaled CNN for efficient frame-level classification       | 0.970          |
|                            | Meso4Inception [1]   | WIFS  | 2018 | Compact shallow CNN with multi-scale Inception modules             | 0.786          |
| <b>Spatial Detectors</b>   | FWA [38]             | CVPRW | 2019 | Detection of affine face warping and blending artifacts            | 0.939          |
|                            | IID [40]             | ECCV  | 2022 | Disentanglement of identity-related features and manipulation cues | 0.916          |
|                            | CLIP [56]            | ICML  | 2021 | Vision-language pretrained representation learning                 | 0.839          |
| <b>Frequency Detectors</b> | F3Net [55]           | ECCV  | 2020 | Learnable frequency decomposition and local frequency statistics   | 0.978          |
|                            | SPSL [41]            | CVPR  | 2021 | Spatial-phase analysis of generative upsampling artifacts          | 0.929          |
|                            | SRM [43]             | CVPR  | 2021 | High-frequency residual filters inspired by steganalysis           | 0.965          |
| <b>Video Detectors</b>     | I3D [10]             | CVPR  | 2017 | 3D convolutional aggregation over temporal volumes                 | 0.968          |
|                            | TimeSformer [7]      | ICML  | 2021 | Divided space-time attention over video clips                      | 0.969          |
|                            | XCLIP [47]           | ECCV  | 2022 | Video-language model with cross-frame communication                | 0.959          |

closely related detectors can obscure the relationship between detector design and adversarial behavior. We therefore prioritize *bias diversity*: each selected model should represent a distinct assumption about where manipulation evidence appears and how that evidence is aggregated.

We organize the detector suite into four families: **Naive CNNs**, **Spatial**, **Frequency** and **Video** detectors. This does not imply that all models within a family are highly similar or that the families are perfectly independent. Rather, they provide a systematization axis for asking how forensic inductive biases shape transferability, query efficiency and the Practical Robustness Frontier (PRF).

As summarized in Table 1, we select 12 models using three criteria: (i) *bias diversity*: the suite should cover all families; (ii) *reproducibility*: each model must have an available implementation or a sufficiently specified architecture that can be integrated into our evaluation pipeline; (iii) *meaningful clean performance*: each model must perform well on FaceForensics++ [59], so that adversarial degradation is interpretable. We use a few historically important or architecturally distinctive baselines with lower clean AUC, because they help separate architectural effects from raw clean accuracy. Details on the detector models are given in Appendix A.

## 3.2 The Cost-Constrained Threat Model

White-box attacks are useful for estimating a strong empirical vulnerability reference, but they are insufficient for characterizing deployed deepfake detection systems. In operational settings, detectors may be exposed through APIs that impose monetary costs, rate limits, hard-label outputs, or query monitoring [13, 26]. We therefore organize the adversarial landscape by the attacker’s query budget and access to feedback.

### 3.2.1 Zero-Query (Gray-Box) Transferability

In the zero-query setting, the adversary has no interaction with the target detector during attack construction. Instead, the adversary crafts perturbations offline against a surrogate model and submits the resulting examples to the target once. This setting captures transfer-based evasion, where adversarial examples generated for one model can remain effective against another [42]. In our SoK, zero-query transfer is diagnostic: successful transfer suggests that the surrogate and target rely on overlapping non-robust features or related inductive biases [27], while weaker transfer suggests less alignment between the surrogate attack and the target decision rule. We instantiate this tier using strong white-box attacks on the surrogate, including AutoAttack [16] and C&W [9] and evaluate the transferred examples on the target.

### 3.2.2 Low-Query Black-Box (1-1,000 Queries)

The low-query setting models an attacker who can interact with the target detector but must remain within a limited query budget. We use this tier as the primary deployment-relevant regime because repeated queries may be subject to rate limits and detection by stateful monitoring systems [13, 26]. We evaluate two feedback models, choosing attacks that match the feedback available to the adversary:

- **Score-Based (SimBA-DCT):** When the API exposes continuous scores or logits, the adversary can use score changes to guide the search. We use SimBA [24] because it is a simple, query-efficient score-based attack and we use its low-frequency Discrete Cosine Transform (DCT) basis to test whether structured frequency directions provide especially efficient evasion against face-forgery detectors.

- **Decision-Based (HSJ):** When the API exposes only a hard real/fake label, the adversary must estimate the decision boundary from labels alone. We use HopSkipJump (HSJ) [12] because it is a standard hard-label boundary-estimation attack and evaluate both  $\ell_2$  and  $\ell_\infty$  variants to measure how norm constraints affect query efficiency under hard-label feedback.

### 3.2.3 High-Query Convergence (1,000–25,000 Queries)

The high-query setting serves as a diagnostic stress test rather than a model of ordinary production API access. It asks whether low-query robustness persists when the attacker is allowed many more boundary-estimation queries. This approximates an offline or weakly monitored oracle, a local replica, or a service without effective query controls. We use HSJ- $\ell_\infty$  because it preserves the hard-label feedback assumption while providing a strong high-budget black-box attack. In the PRF analysis, this setting measures how quickly black-box attacks approach the white-box reference using the  $Q_{95}$  convergence metric.

### 3.2.4 White-Box and Adaptive Threat Models

Finally, we evaluate a white-box adversary with access to the detector architecture and gradients. This tier provides an empirical reference ceiling, denoted  $ASR_{WB}$ , against which black-box convergence can be compared. For undefended models, we use the parameter-free AutoAttack ensemble [16] as the primary white-box evaluator because it reduces dependence on attack-specific hyperparameter tuning. For defenses that introduce stochasticity, input transformations, or non-differentiable components, standard gradients may yield misleading robustness estimates. We therefore use EOT [5] and BPDA [4], following established guidance on adaptive defense evaluation [71].

## 3.3 Evaluation Metrics

Static metrics such as clean AUC or white-box ASR summarize only one point in the security landscape. They do not show how attack success changes as the adversary receives more feedback or a larger query budget. We therefore introduce the **Practical Robustness Frontier (PRF)**, a cost-aware curve that maps attack success to attacker effort.

For a detector  $f_\theta$ , attack algorithm  $A$  and maximum query budget  $Q$ , let  $A(x_i, f_\theta, Q)$  denote the adversarial example returned for input  $x_i$  after at most  $Q$  target queries. The black-box attack success rate at budget  $Q$  is

$$ASR_{BB}(Q) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[f_\theta(A(x_i, f_\theta, Q)) \neq y_i].$$

The PRF plots  $ASR_{BB}(Q)$  against  $\log_{10}(Q)$ . A steep curve indicates that additional queries rapidly increase attack success,

while a flatter curve indicates lower marginal returns from additional probing.

We summarize attacker effort using three diagnostics. First, *zero-query transfer* measures ASR and AUC degradation for adversarial examples crafted on a surrogate and evaluated on the target without target queries. This corresponds to the no-interaction endpoint of the threat model.

Second, we report *early-break thresholds*:

$$Q_\tau = \min \{Q \in \mathbb{N} \mid ASR_{BB}(Q) \geq \tau\},$$

for  $\tau \in \{0.05, 0.20, 0.50\}$ . We refer to these thresholds as *Leakage* (5% ASR), *Practical Break* (20% ASR) and *Catastrophic Break* (50% ASR), respectively. These thresholds measure the query budget at which evasion first becomes measurable, operationally significant, or severe.

Third, we define  $Q_{95}$  to measure convergence toward the white-box reference:

$$Q_{95} = \min \{Q \in \mathbb{N} \mid ASR_{BB}(Q) \geq 0.95 \times ASR_{WB}\}.$$

Here,  $ASR_{WB}$  denotes the strongest white-box success rate measured under the same perturbation constraints. Thus,  $Q_{95}$  is the query budget required for black-box attacks to reach 95% of the measured white-box failure ceiling. If the threshold is not reached within the maximum query budget, we report the metric as right-censored, e.g.,  $Q_{95} > Q_{\max}$ .

## 4 Evaluation Protocol

This section summarizes the training, inference, dataset and attack protocols used to construct the Practical Robustness Frontier (PRF). Our goal is to make the evaluation pipeline reproducible and to reduce confounding from inconsistent preprocessing, model weights, or attack configurations. Appendix B describes the DeepfakeBench–ART integration, Appendix C details the dataset infrastructure and frame counts and Appendix D reports the full attack hyperparameters.

### 4.1 Training

We build our evaluation pipeline on *DeepfakeBench* [80], which provides standardized data loading, preprocessing, detector implementations and evaluation utilities for deepfake detection. Because public pre-trained weights can differ in preprocessing assumptions, training schedules and checkpoint quality, we retrain or fine-tune the selected detectors under a common protocol rather. Models are initialized either from scratch or from standard ImageNet/video pre-training when used by the corresponding architecture and are trained on FaceForensics++ (FF++) [59]. The resulting clean AUCs are reported in Table 1.

## 4.2 Inference

Evaluating heterogeneous detectors under the same adversarial protocol requires bridging the media-forensics pipeline with adversarial-optimization tooling. We therefore integrate DeepfakeBench with the Adversarial Robustness Toolbox (ART) [48], using a unified wrapper (Appendix B) for model inference, loss computation and attack execution.

The input format depends on the detector family. Frame-level detectors – including Naive CNNs, Spatial detectors and Frequency detectors – process RGB tensors of shape  $(3, 256, 256)$ . For video-level scoring, we aggregate frame-level outputs using a consistent protocol. Spatiotemporal video detectors – I3D, TimeSformer and XCLIP – process continuous clips of shape  $(T, 3, 256, 256)$ , where  $T$  is the temporal window length. For these models, adversarial perturbations are optimized over the full clip tensor so that the attack objective matches the model’s sequence-level inference path.

To map the PRF, we use the query-budget tiers defined in Section 3.2. White-box attacks are constrained by an  $\ell_\infty$  perturbation bound of  $\epsilon = 8/255$  unless otherwise specified. Low-query black-box evaluations are capped at  $Q_{\max} = 1,000$  target queries per sample, while high-query convergence experiments extend to  $Q_{\max} = 25,000$  for estimating convergence toward the white-box reference. For SimBA-DCT, we use  $\epsilon = 0.02$  and  $\text{freq\_dim} = 16$ . Because measured PRFs depend on attack parameterization, we report the attack basis, norm constraint, perturbation bound and query budget for each experiment in Appendix D. Where relevant, we include paired checks, such as SimBA-DCT versus SimBA-Pixel and HSJ- $\ell_\infty$  versus HSJ- $\ell_2$ , to verify that our conclusions are not artifacts of a single attack mode.

## 5 Vulnerability Analysis

### 5.1 Zero-Query Transferability

Zero-query transferability is the lowest-interaction black-box threat that detectors must confront. The attacker does not query the target detector  $f_t$  during attack construction, but is assumed to know the task domain and data distribution well enough to train or select a local surrogate  $f_s$ . The attacker then crafts adversarial examples offline against  $f_s$  and deploys them blindly against  $f_t$ . This captures deployments where target feedback is unavailable, aggressively rate-limited, or monitored by intrusion-detection systems. Because adversarial examples often transfer across models, such attacks can enable black-box evasion without target gradients or iterative target access [51, 72]. In this setting, high transferability undermines the security assumptions behind closed-API deployment by shifting the cost of optimization offline.

We evaluate all  $12 \times 12$  source→target permutations across our model suite, including source=target cases as calibration points. For each surrogate, we craft adversarial exam-

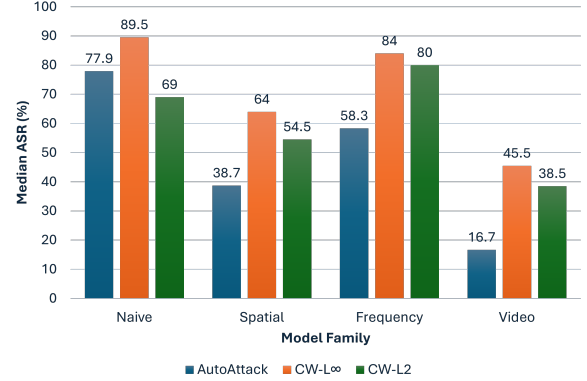


Figure 1: **Within-family zero-query transfer ASR by detector family and surrogate attack.** Bars report median ASR for attacks crafted on surrogates from the same architectural family as the target, excluding source=target calibration cases. CW- $\ell_\infty$  yields the strongest transfer overall.

ples using AutoAttack [16], CW- $\ell_\infty$  and CW- $\ell_2$  [9]. These provide complementary baselines: AutoAttack represents strong parameter-free optimization, while the CW variants test whether margin-based optimization yields more transferable examples. Hyperparameters are detailed in Appendix D.

Table 2 reports Attack Success Rate (ASR) and AUC after transfer, aggregated across the four detector families. This family-level view identifies which architectural paradigms share transferable adversarial features.

**Finding V1:** Zero-query transfer is strong enough to threaten closed-API detectors. Table 2 and Figure 1 show that attacks frequently transfer at high ASR even when the attacker never queries the target detector. CW- $\ell_\infty$  reaches at least 39% ASR for every family-level source→target pair and Catastrophic Break (50% or more) on all family-level Naive, Spatial and Frequency targets except for Naive→Frequency. Zero-query transfer is thus a significant threat: when attackers can train or select surrogates from the same data distribution, closed APIs cannot rely on opacity or rate limits alone.

**Finding V2:** Transferability is attack-dependent. Across family-level transfers, CW- $\ell_\infty$  is generally the most transferable attack, consistent with the view that  $\ell_\infty$  perturbations can distribute small changes across many input dimensions and activate shared non-robust features [27]. However, transferability does not simply follow white-box attack strength. CW- $\ell_2$  often outperforms AutoAttack in cross-family settings, such as Naive→Video, where CW- $\ell_2$  reaches 35.0% ASR compared with 19.4% for AutoAttack. Zero-query evaluation should therefore include multiple surrogate attacks rather than a single white-box baseline.

**Finding V3:** Temporal aggregation disrupts (some) transfer. Video targets are generally harder to compromise from spatial or frequency surrogates than frame-level families. For example, Spatial→Video and Frequency→Video reach 39.0%

Table 2: **Family-Aggregated Zero-Query Transferability Matrix.** Rows denote the surrogate family; columns denote the target family. Each cell reports median ASR (%) / median AUC for AutoAttack, CW- $\ell_\infty$  and CW- $\ell_2$ . Aggregations exclude identical source-target model pairs (no diagonal self-transfer).

| Surrogate $\rightarrow$ | Attack            | Target Family (Median ASR% / Median AUC) |                   |                     |                 |
|-------------------------|-------------------|------------------------------------------|-------------------|---------------------|-----------------|
| Target Family           | Algorithm         | Naive CNNs                               | Spatial Detectors | Frequency Detectors | Video Detectors |
| Naive                   | AutoAttack        | 77.9 / 0.270                             | 41.8 / 0.541      | 29.5 / 0.434        | 19.4 / 0.378    |
|                         | CW- $\ell_\infty$ | 89.5 / 0.153                             | 54.0 / 0.479      | 39.0 / 0.624        | 44.0 / 0.598    |
|                         | CW- $\ell_2$      | 69.0 / 0.342                             | 44.0 / 0.616      | 31.0 / 0.705        | 35.0 / 0.676    |
| Spatial                 | AutoAttack        | 80.0 / 0.276                             | 38.7 / 0.584      | 24.7 / 0.476        | 16.0 / 0.461    |
|                         | CW- $\ell_\infty$ | 95.0 / 0.102                             | 64.0 / 0.420      | 72.0 / 0.365        | 39.0 / 0.628    |
|                         | CW- $\ell_2$      | 75.0 / 0.288                             | 54.5 / 0.512      | 54.0 / 0.516        | 33.0 / 0.692    |
| Frequency               | AutoAttack        | 82.4 / 0.257                             | 40.5 / 0.569      | 58.3 / 0.598        | 18.0 / 0.438    |
|                         | CW- $\ell_\infty$ | 96.0 / 0.081                             | 57.0 / 0.454      | 84.0 / 0.226        | 41.0 / 0.612    |
|                         | CW- $\ell_2$      | 64.0 / 0.382                             | 45.0 / 0.608      | 80.0 / 0.227        | 36.0 / 0.670    |
| Video                   | AutoAttack        | 63.6 / 0.388                             | 35.0 / 0.576      | 24.3 / 0.490        | 16.7 / 0.571    |
|                         | CW- $\ell_\infty$ | 99.0 / 0.048                             | 68.0 / 0.392      | 63.0 / 0.412        | 45.5 / 0.578    |
|                         | CW- $\ell_2$      | 82.0 / 0.228                             | 56.0 / 0.494      | 49.0 / 0.558        | 38.5 / 0.643    |

Table 3: **Low-query failure thresholds under SimBA-DCT.** We report: query ranges (min–max across models within each family) to reach  $\text{ASR} \geq 5\%$  (leak),  $\text{ASR} \geq 20\%$  (practical break) and  $\text{ASR} \geq 50\%$  (catastrophic), plus family-median ASR and AUC-drop at representative budgets. **Note:** for Spatial and Video families,  $\text{ASR} \geq 50\%$  is reached only by a subset of models at 1k (I3D and XCLIP, respectively); other models remain below 50% within the sweep range.

| Family    | $q(\text{ASR} \geq 5\%)$ | $q(\text{ASR} \geq 20\%)$ | $q(\text{ASR} \geq 50\%)$ | Med. ASR@500 | Med. ASR@1000 | Med. AUC-drop@1000 (pp) |
|-----------|--------------------------|---------------------------|---------------------------|--------------|---------------|-------------------------|
| Naive     | 100–250                  | 500                       | 1000                      | 0.21         | 0.58          | 38.6                    |
| Spatial   | 250                      | 750                       | > 1000                    | 0.15         | 0.45          | 28.1                    |
| Frequency | 50                       | 250–500                   | 750–1000                  | 0.37         | 0.68          | 47.8                    |
| Video     | 250–500                  | 750–1000                  | > 1000                    | 0.10         | 0.21          | 10.1                    |

and 41.0% ASR under CW- $\ell_\infty$ , respectively, lower than most Spatial/Frequency target combinations. This suggests that temporal aggregation can disrupt perturbations optimized for static spatial or spectral cues. However, given that many combinations with a video target are near or even above 40% ASR, the threat of transfer attacks against video models remains significant.

#### Remark 1

Our results show that zero-query transferability is a practical threat: transfer attacks frequently reach high ASR across detector families.

## 5.2 Low-Query Attacks (1–1,000 Queries)

Most black-box robustness evaluations assume that attackers can make thousands of queries to the model, but attacks on deployed detectors are usually constrained by API rate limits, query costs and monitoring. The low-query regime tests whether detectors fail within the interaction budget most rele-

vant to constrained APIs. We use the early-break thresholds defined in Section 3.3: Leakage ( $\text{ASR} \geq 5\%$ ), Practical Break ( $\text{ASR} \geq 20\%$ ) and Catastrophic Break ( $\text{ASR} \geq 50\%$ ).

We evaluate this regime using SimBA-DCT, which is especially relevant for detectors that rely on spatial or spectral regularities. To confirm our choice of the DCT variant, we compare SimBA-DCT with pixel-space SimBA at the same query budget and find that the DCT basis yields higher ASR for most evaluated models. We report ASR and AUC degradation at fixed query milestones up to 1,000 queries. Table 3 and Figure 2 summarize the family-level breakpoints, with per-model trajectories reported in Appendix F.

**Finding V4:** *Frequency-aware detectors break earliest under low-query DCT attacks.* The Frequency family (F3Net/SPSL/SRM) reaches Leakage at roughly 50 queries and Catastrophic Break between 750 and 1,000 queries. This is the steepest early PRF among the four families. The result suggests that frequency-aware detectors can expose low-dimensional attack directions: the same spectral regularities that help distinguish real from fake media can also make low-frequency black-box search unusually query-efficient.

Table 4: **High-Query Practical Robustness by Architectural Family (HSJ- $\ell_\infty$  Only).** Results reflect the asymptotic HSJ- $\ell_\infty$  **High Budget** setting (25k queries). The table reports the **median** Attack Success Rate (ASR), **median** attacked-set AUC and the asymptotic  $Q_{95}$  convergence limit. Ranges denote the interquartile span (25th–75th percentile) across models within each family.

| Attack Configuration (by Target Family)                | HSJ- $\ell_\infty$ ASR | HSJ- $\ell_\infty$ AUC | $Q_{95}$ (Queries)     |
|--------------------------------------------------------|------------------------|------------------------|------------------------|
| <b>HSJ-<math>\ell_\infty</math> (High Budget: 25k)</b> |                        |                        |                        |
| ↪ Naive CNNs                                           | 0.82 (0.80–0.82)       | 0.262 (0.258–0.270)    | 6,812 (5,934–7,430)    |
| ↪ Spatial Detectors                                    | 0.72 (0.67–0.73)       | 0.317 (0.310–0.319)    | 10,629 (10,447–13,239) |
| ↪ Frequency Detectors                                  | 0.77 (0.77–0.79)       | 0.287 (0.281–0.291)    | 13,452 (12,611–14,430) |
| ↪ Video Detectors                                      | 0.54 (0.53–0.55)       | 0.440 (0.433–0.447)    | 23,148 (22,901–23,395) |

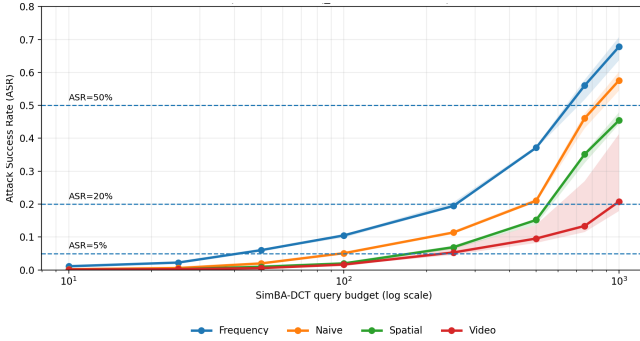


Figure 2: **Low-query Practical Robustness Frontier under SimBA-DCT** ( $\epsilon = 0.02$ , freq\_dim=16). Curves show family-median ASR versus query budget with interquartile bands; dashed lines mark Leakage, Practical Break and Catastrophic Break thresholds. Frequency-aware detectors break earliest, while Video models remain lowest overall.

**Finding V5: Temporal aggregation delays low-query evasion.** Video models exhibit the lowest family-median ASR through the 10–1,000 query range and remain below the Practical Break threshold for much of the sweep. This supports the hypothesis that temporal aggregation makes query-efficient evasion harder by requiring perturbations to remain effective across a sequence of frames. The wide Video-family variance in Figure 2 is driven by XCLIP, which behaves more like a frame-level model under SimBA-DCT. Thus, low-query robustness depends on the strength of temporal coupling, not merely on whether the detector consumes video.

#### Remark 2

Low-query evaluation reveals failures that clean AUC and high-budget attack ceilings hide, such as the higher risks of frequency-based detectors.

### 5.3 High-Query Convergence

While the low-query regime captures immediate operational risk, *high-query convergence* measures how quickly a black-box attacker approaches a detector’s white-box failure ceiling.

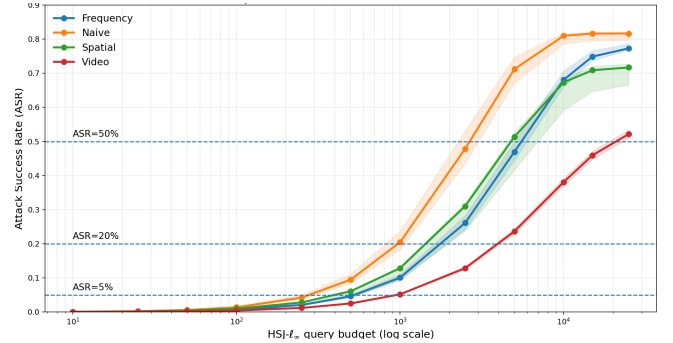


Figure 3: **High-Query Robustness Frontier under HSJ- $\ell_\infty$ .** Family-median ASR versus query budget up to 25,000 queries with interquartile bands. Video models exhibit the slowest convergence toward their white-box failure ceiling, yielding the largest  $Q_{95}$ .

We quantify this using  $Q_{95}$ : the minimum query budget required to reach 95% of the model’s white-box ASR.

We estimate this frontier using HSJ- $\ell_\infty$  with a maximum budget of 25,000 queries. We use HSJ- $\ell_\infty$  as the primary high-query stress test because it preserves the hard-label feedback assumption while producing higher ASR than HSJ- $\ell_2$  in paired comparisons at the same budget. Table 4 and Figures 3–4 show a clear separation between frame-based processing and temporally coupled video architectures, with the HSJ- $\ell_2$  comparison reported in Figure 7 in the Appendix.

**Finding V7: Temporal architectures impose the largest high-query convergence penalty.** Family-median PRF curves show that temporal models degrade much more slowly under HSJ- $\ell_\infty$  than frame-based detectors. Naive CNNs reach high ASR quickly, with a median  $Q_{95}$  of 6,812 queries. By contrast, Video models require a median of 23,148 queries to reach  $Q_{95}$ . This corresponds to roughly  $3.4\times$  the query cost of Naive CNNs,  $2.2\times$  that of Spatial detectors and  $1.7\times$  that of Frequency detectors. While temporal aggregation does not prevent eventual evasion, it substantially increases the query cost required for a black-box attacker.

**Finding V8: Similar white-box ceilings can mask very different high-query convergence.** Figure 4 compares XCLIP,

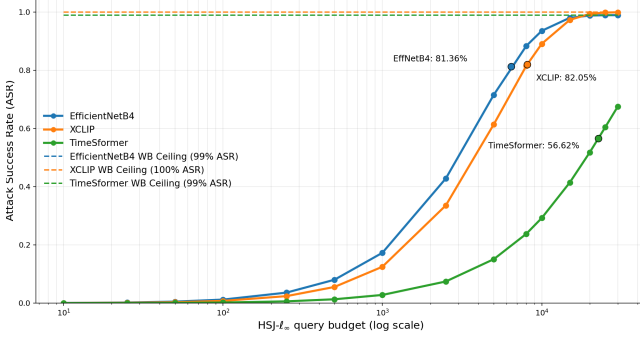


Figure 4: **XCLIP anomaly under  $\text{HSJ-}\ell_\infty$ .** XCLIP, EfficientNet-B4 and TimeSformer have similar white-box failure ceilings, but their PRF curves diverge sharply: XCLIP tracks the Naive CNN, while TimeSformer shows greater resilience.

EfficientNet-B4 and TimeSformer under  $\text{HSJ-}\ell_\infty$ . All three models approach near-total failure under white-box attacks, but their black-box convergence differs sharply: XCLIP tracks the Naive CNN baseline, while TimeSformer preserves a much flatter PRF over the same query range. This shows that white-box ASR alone obscures the query cost required to realize a detector’s vulnerability. It also shows the limitation of treating all video or multimodal detectors as structurally equivalent. Although XCLIP processes video and uses a foundation-style vision-language representation, its convergence behavior suggests that cross-frame multimodal aggregation is not enough; robustness depends on whether the architecture tightly couples evidence across time.

#### Remark 3

High-query convergence shows that while all detectors fail under white-box attacks, they differ substantially in how many black-box queries are required to approach that failure ceiling. Temporal coupling produces the slowest convergence in our evaluation.

## 5.4 Frequency Cues as Attack Surfaces

The low-query results in Section 5.2 reveal a recurring trade-off for frequency-aware detectors. Models such as F3Net, SPSL and SRM are designed to exploit spectral, phase, or high-frequency residual cues associated with synthesis pipelines [41, 43, 55]. These cues can improve detection and cross-generator generalization, but they also define a structured feature space that a black-box attacker can probe.

Figure 5 isolates this effect by comparing the three Frequency detectors against EfficientNet-B4 and I3D under SimBA-DCT. Because SimBA-DCT searches over low-frequency DCT basis directions, it tests whether frequency-structured perturbations provide a query-efficient path through

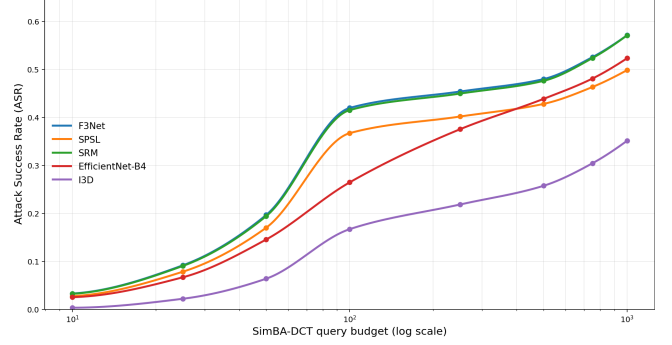


Figure 5: **Frequency cues as attack surfaces under SimBA-DCT.** ASR progression for frequency detectors vs. EfficientNet-B4 and I3D. F3Net and SRM degrade fastest, while I3D remains substantially lower across the query sweep.

the detector’s decision boundary. The frequency-aware models degrade earlier than the temporal baseline and, for F3Net and SRM, also rise more steeply than the Naive CNN baseline. At 1,000 queries, F3Net and SRM exceed 0.57 ASR, while EfficientNet-B4 is roughly 5 percentage points lower and I3D remains roughly 22 percentage points lower. SPSL follows the same overall trend but with a less severe late-query slope.

**Finding V9:** *Frequency priors can become query-efficient attack surfaces.* The same spectral structure that makes frequency-aware detectors sensitive to generative artifacts can give SimBA-DCT a compact search space for evasion. In PRF terms, frequency-aware detectors exhibit steeper early degradation because the attack basis is partially aligned with the detector’s feature prior.

This result qualifies the role of frequency analysis in deep-fake detection. Frequency cues remain useful forensic signals, but robustness should not depend on fixed artifacts that may shift across synthesis pipelines and newer generators [15].

#### Remark 4

Frequency-domain evidence is not robust. When the attack basis overlaps the detector’s feature prior, as in SimBA-DCT against frequency models, black-box evasion can become substantially more query-efficient.

## 6 Defense Analysis

The PRF analysis in Section 5 shows that detector vulnerability depends on how quickly an attacker can reach meaningful evasion under query constraints. We therefore evaluate defenses by how they shift the PRF: whether they reduce false accepts, preserve clean utility, delay black-box convergence, or withstand adaptive white-box attacks. We consider three defense families: adversarial training, boundary shaping and generative purification.

Table 5: **Unified Adversarial Training summary by architectural family.** **Left:** white-box (AutoAttack), **Right:** black-box (SimBA-DCT). Values are **family medians** across constituent models.

| Model Family | Defense | White-box (AutoAttack) |              | Black-box (SimBA-DCT) |              |
|--------------|---------|------------------------|--------------|-----------------------|--------------|
|              |         | Clean AUC              | Robust AUC   | Robust AUC            | $Q_{95}$     |
| Naive        | AT      | 0.962                  | 0.206        | 0.273                 | <b>1,796</b> |
|              | TRADES  | 0.958                  | 0.236        | 0.334                 | 1,530        |
|              | MART    | 0.947                  | 0.214        | 0.281                 | 1,536        |
|              | GradReg | 0.965                  | 0.187        | 0.246                 | 1,491        |
|              | AFSL    | <b>0.975</b>           | <b>0.321</b> | <b>0.543</b>          | 1,610        |
| Spatial      | AT      | 0.904                  | 0.184        | 0.280                 | 1,602        |
|              | TRADES  | 0.899                  | 0.212        | 0.336                 | <b>1,671</b> |
|              | MART    | 0.892                  | 0.196        | 0.284                 | 1,668        |
|              | GradReg | 0.907                  | 0.162        | 0.252                 | 1,640        |
|              | AFSL    | <b>0.976</b>           | <b>0.255</b> | <b>0.503</b>          | 1,637        |
| Frequency    | AT      | 0.956                  | 0.196        | 0.301                 | 1,461        |
|              | TRADES  | 0.952                  | 0.224        | 0.328                 | <b>1,697</b> |
|              | MART    | 0.946                  | 0.206        | 0.303                 | 1,554        |
|              | GradReg | <b>0.958</b>           | 0.173        | 0.252                 | 1,488        |
|              | AFSL    | 0.946                  | <b>0.316</b> | <b>0.485</b>          | 1,569        |
| Video        | AT      | 0.956                  | 0.241        | 0.291                 | 1,400        |
|              | TRADES  | 0.952                  | 0.262        | 0.310                 | <b>1,857</b> |
|              | MART    | 0.946                  | 0.236        | 0.298                 | 1,369        |
|              | GradReg | 0.958                  | 0.224        | 0.264                 | 1,605        |
|              | AFSL    | <b>0.962</b>           | <b>0.263</b> | <b>0.422</b>          | 1,743        |

## 6.1 Adversarial Training Is Regime Dependent

Adversarial Training (AT) [44] improves robustness by training on perturbations generated inside the learning loop. In the PRF setting, we seek to know which part of the operational frontier each variant shifts. We compare standard AT with TRADES [81], MART [77], input-gradient regularization [58] and Adversarial Feature Similarity Learning (AFSL) [32]. Summarized results are reported in Table 5.

**Finding D1:** *Adversarial training variants do not yield reliable robust detectors under direct white-box evaluation.* Table 5 summarizes family-median defense performance under AutoAttack and SimBA-DCT with  $Q_{\max} = 5,000$ . We use AutoAttack as the white-box representative because it provides a standardized parameter-free robustness evaluator and we use SimBA-DCT as the black-box representative because it is the strongest low-frequency score-based attack in our PRF analysis. All adversarial-training variants remain weak under white-box AutoAttack. Standard AT, TRADES, MART and GradReg all produce family-median robust AUCs far below 0.5 across the detector families. AFSL is the strongest variant in several rows, but even AFSL reaches only 0.321, 0.255, 0.316 and 0.263 white-box robust AUC for Naive, Spatial, Frequency and Video detectors, respectively. Thus, these defenses do not restore usable white-box robustness; they mainly change how the models fail.

**Finding D2:** *AFSL improves query-bounded robustness.*

AFSL regularizes feature representations so that clean and adversarial inputs remain close in feature space while preserving separation between real and fake classes. Table 5 shows that AFSL provides the highest robust AUC under SimBA-DCT and HSJ. This indicates that feature-similarity regularization can shift the operational PRF by making query-limited evasion less efficient. AFSL does not help much, however, against white-box evasion attacks.

**Finding D3:** *Adversarial-training variants remain vulnerable under direct semantic AutoAttack.* We further evaluate adversarially trained detectors with an adaptive AutoAttack that directly targets the deployed fake-to-real failure condition. At the standard  $\epsilon = 8/255$  setting, MART is the most resistant variant, with family-level semantic ASR ranging from 42.2–58.6%, consistent with its lower false-accept profile. TRADES improves over standard AT but remains less targeted to false-accept suppression, with ASR ranging from 48.5–65.4%. AFSL, despite strong query-bounded black-box robustness, is particularly vulnerable under the adaptive attack, with ASR ranging from 56.4–76.2%. The  $\epsilon$  ablation in Appendix G shows the same trend: all defenses degrade as the perturbation budget increases, while MART remains the strongest option for suppressing fake-to-real evasion.

#### Remark 5

No adversarial-training variant provides reliable robustness across the evaluated threat regimes. AFSL offers the strongest improvement against query-bounded SimBA-DCT, while MART better suppresses targeted fake-to-real failures; however, both remain substantially vulnerable under white-box evaluation. Adversarial training therefore shifts the degree and mode of failure rather than eliminating it.

## 6.2 Boundary-Shaping Defenses

While adversarial training implicitly shapes the decision space through examples, boundary-shaping defenses create explicit adversarial regions. We evaluate two representative approaches, **DBAT** [36] and **PadNet** [6]. Table 6 summarizes their performance using family medians across the four detector families; full per-model results are reported in Appendix Table 13. We compare clean-AUC cost, black-box robustness under SimBA-DCT at 1,500 queries and  $\text{HSJ-}\ell_\infty$  at 25,000 queries and  $Q_{95}$  under SimBA-DCT.

**Finding D4:** *Boundary shaping offers only limited defense.* Both defenses remain vulnerable at the evaluated query budgets. Table 6 shows that under SimBA-DCT at 1,500 queries, family-median ASR ranges from 44–57% for PadNet and 48–62% for DBAT, while attacked-set AUC remains below 0.5 for every family. At 25,000 queries,  $\text{HSJ-}\ell_\infty$  raises family-median ASR to 64–77% for PadNet and 68–80% for DBAT. These values are better than most AT variants, but worse than AFSL and worse than random guessing. In short, neither defense should be considered reliable against attacks with over 1,500 queries.

**Finding D5:** *PadNet provides a consistently better clean-utility/query-cost trade-off than DBAT.* Although neither defense is sufficient by itself, PadNet performs consistently better across the four detector families. Its family-median clean-AUC loss ranges from 0.8–1.5 percentage points, compared with 2.1–2.9 pp for DBAT. PadNet also reduces family-median SimBA-DCT ASR by 4–5 pp and  $\text{HSJ-}\ell_\infty$  ASR by 3–4 pp relative to DBAT. Finally, PadNet increases SimBA-DCT  $Q_{95}$  by approximately 300–400 queries over the undefended baseline, whereas DBAT increases it by only 100–200 queries. PadNet therefore shifts the PRF farther while imposing a smaller clean-utility cost, but the remaining attack success rates show that this improvement is incremental rather than sufficient for standalone defense.

**Finding D6:** *Boundary shaping is even more vulnerable when the attacker targets the boundary mechanism directly.* We further stress-test PadNet and DBAT with defense-aware adaptive attacks and report the full  $\epsilon$  sweeps in Appendix H along with detailed per-model family results. For PadNet, the pad-gradient attack often reaches the padding region but does not always cross into the real class; at the standard

$\epsilon = 8/255$  setting, family-level ASR ranges from 37.8–54.1%. The stronger rejection-aware margin attack directly optimizes  $z_{\text{real}} > \max(z_{\text{fake}}, z_{\text{pad}})$  and raises ASR to 49.2–68.4%. DBAT shows the same pattern: collapsing clean and adversarial companion classes yields 48.3–62.9% ASR, while treating companion outputs as rejection and targeting only the clean real class raises ASR to 58.9–73.6%. Overall, boundary shaping shifts the PRF by delaying and redirecting attack trajectories, but an adaptive attacker can still use boundary, rejection, or companion regions as waypoints toward the harmful real decision region.

#### Remark 6

While PadNet provides a stronger clean-utility/query-cost trade-off than DBAT, neither defense is secure against attacks with moderate query budgets.

## 6.3 Purification Defenses

Purification defenses prepend an input-transformation module  $P_\theta(\cdot)$  before the detector  $f(\cdot)$ , yielding a composed system  $g(x) = f(P_\theta(x))$ . Because these defenses are stochastic and transformation-based, they must be evaluated under both non-adaptive and adaptive threat models. We examine **DiffPure** [49] and **IVORY** [63] as representative techniques.

**Finding D7:** *Purification fails against black-box attacks.* Figure 6 aggregates performance when the adversary attacks the composed system  $g(x)$  using ( $\text{HSJ-}\ell_\infty$  high, SimBA-DCT @ 1,500 queries). Across detector families, median ASR remains high, ranging from roughly 49% to 61%. DiffPure yields slightly lower black-box ASR than IVORY, but both defenses leave substantial evasion rates under query-bounded attacks. Thus, purification should also be considered a weak defense against attacks with 1,500 queries or more.

**Finding D8:** *Adaptive evaluation reveals substantially more stable robustness for IVORY than DiffPure.* Under BPDA+EOT with  $K = 8$ , IVORY maintains low family-median attack success rates of 14.4% under adaptive AutoAttack and 13.5% under adaptive C&W. DiffPure is considerably more vulnerable under the same evaluation, reaching median ASRs of 29.4% and 27.7%, respectively. At the individual-model level, DiffPure reaches ASRs as high as 39.1% under AutoAttack and 36.4% under C&W, whereas IVORY remains below 27% across all evaluated models.

The difference becomes clearer as the EOT estimate is strengthened. Increasing the sample count from  $K = 4$  to  $K = 8$  changes IVORY’s median ASR only from 13.8% to 14.4% under AutoAttack and from 12.7% to 13.5% under C&W. In contrast, DiffPure increases from 16.1% to 29.4% and from 14.9% to 27.7%, corresponding to increases of 13.3 and 12.8 percentage points. This sensitivity suggests that DiffPure’s apparent robustness under weaker adaptive evaluation is partly influenced by stochastic optimization difficulty,

Table 6: **Boundary-shaping defenses by architectural family.** Values are family medians across three detectors.  $\Delta$ Clean reports the clean-AUC loss relative to the undefended model in percentage points; attack columns report ASR / attacked-set AUC; and  $\Delta Q_{95}$  reports the increase in SimBA-DCT queries over the undefended baseline. Lower  $\Delta$ Clean and ASR, higher AUC and larger  $\Delta Q_{95}$  are better.

| Family            | $\Delta$ Clean AUC (pp) |            | SimBA-DCT @ 1.5k  |                   | HSJ- $\ell_\infty$ @ 25k |                   | $\Delta Q_{95}$ (SimBA) |             |
|-------------------|-------------------------|------------|-------------------|-------------------|--------------------------|-------------------|-------------------------|-------------|
|                   | PadNet                  | DBAT       | PadNet            | DBAT              | PadNet                   | DBAT              | PadNet                  | DBAT        |
| Naive CNNs        | 1.5                     | 2.6        | .57 / .397        | .62 / .364        | .77 / .277               | .80 / .266        | +400                    | +200        |
| Spatial           | 0.8                     | 2.1        | .53 / .411        | .57 / .393        | .76 / .273               | .79 / .252        | +300                    | +200        |
| Frequency         | 1.2                     | 2.4        | .56 / .391        | .60 / .380        | .76 / .287               | .80 / .275        | +400                    | +100        |
| Video             | 0.9                     | 2.9        | .44 / .441        | .48 / .418        | .64 / .361               | .68 / .333        | +300                    | +100        |
| <b>All models</b> | <b>1.0</b>              | <b>2.5</b> | <b>.55 / .400</b> | <b>.59 / .384</b> | <b>.76 / .284</b>        | <b>.80 / .267</b> | <b>+400</b>             | <b>+150</b> |

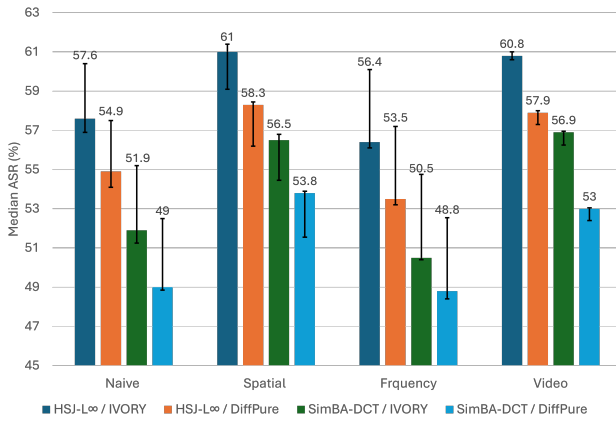


Figure 6: **Black-box purification robustness.** Family-median ASR under HSJ- $\ell_\infty$  and SimBA-DCT against purified systems. IVORY slightly outperforms DiffPure, but neither prevents high evasion rates.

whereas IVORY is substantially more stable under the adaptive attacks we evaluate.

#### Remark 7

Purification fails against black-box attacks with moderate query budgets. Partial white-box robustness against adaptive attacks provides hope that zero-query grey-box attacks may be difficult against these defenses.

## 7 Discussion and Future Directions

Our approach to cost-aware robustness motivates a shift in how deepfake detectors are evaluated and deployed.

### 7.1 Cost-Aware Robustness

Scalar metrics such as clean AUC, cross-dataset AUC, or final attack success rate summarize whether a detector fails, but

not how much interaction an attacker needs to reach that failure. This distinction is critical for deployed systems, where feedback may be limited by hidden scores, rate limits, query costs, or monitoring. We recommend that future robustness evaluations supplement scalar reporting with PRF-style measurements. At minimum, security claims should report zero-query transferability and low-query breakpoints. These metrics make attacker effort explicit and allow detectors with similar clean or white-box performance to be compared under deployment-relevant constraints [8].

### 7.2 Reproducibility as a Barrier to Evaluation

Our evaluation exposed substantial engineering barriers when combining general-purpose adversarial-robustness toolboxes with media-forensics repositories. Frameworks such as ART and DeepfakeBench use different assumptions about input formats, batching, preprocessing and model wrappers, requiring nontrivial integration before attacks can be run consistently.

A second barrier is the fragility of detector artifacts. Some public checkpoints depend on undocumented preprocessing, augmentation choices, or architecture-specific loading assumptions and small mismatches can silently degrade clean performance. Video models introduce an additional challenge: adversarial optimization over high-dimensional clip tensors can trigger memory exhaustion even when the same model is practical for ordinary inference. These issues make robustness claims difficult to reproduce unless authors release verified checkpoints, preprocessing code, attack wrappers and end-to-end evaluation scripts. Standardized threat models and reproducible pipelines should thus be treated as part of the robustness claim, not just as implementation details [54].

### 7.3 Distribution Shift

Our PRF analysis is anchored to detectors trained on FaceForensics++ and therefore primarily reflects artifacts associated with autoencoder- and GAN-era manipulations. Newer generators, including latent diffusion models and autoregressive video models, may change the forensic signal available to

detectors. In particular, diffusion-based synthesis can weaken some high-frequency artifacts associated with GAN upsampling [15], which may reduce the reliability of detectors that depend heavily on fixed spatial or spectral artifacts.

This does not imply that frequency or spatial cues should be discarded. Rather, it suggests that detectors relying on a single artifact channel may be brittle under both distribution shift and adaptive attack. Future work should map PRFs across newer synthetic-media datasets and generator families, including diffusion-based images and long-form video generation. Our results suggest that temporal consistency is a promising source of robustness, but this claim should be tested under modern video generators, where motion coherence, object persistence and cross-frame identity consistency may differ substantially from FF++-style manipulations.

## 7.4 Recommendations for Practitioners

We distill our findings into deployment-oriented recommendations, organized by operational setting.

**Model procurement and evaluation.** Practitioners should not select detectors solely by clean AUC or cross-dataset generalization. Procurement evaluations should include zero-query transfer, low-query breakpoints and high-query convergence. When latency and compute budgets permit, temporally coupled video architectures should be considered, because they produced the slowest black-box convergence in our evaluation. However, video input alone is not sufficient: architectures should be evaluated for whether they actually couple evidence across time rather than simply aggregating frame-level features like XCLIP does.

**API-limited deployments.** For services protected by rate limits, query costs, or monitored access, the key question is how quickly an attacker can reach useful evasion. Defenses such as AFSL and IVORY are best interpreted as PRF-shifting components: they can raise the cost of query-constrained attacks, but they do not remove vulnerabilities. In this setting, defenders should evaluate whether a defense delays Leakage and Practical Break under the specific feedback exposed by the deployment interface.

## 7.5 Limitations and Future Directions

First, our experiments are anchored to FaceForensics++ (c23) in order to isolate architectural differences under a common training and evaluation pipeline. Future work should map robustness frontiers under stronger distribution shift, including Celeb-DF [39], DFDC [18] and newer diffusion- or transformer-generated media.

Second, computational constraints limited the scale of adaptive evaluation. BPDA+EOT over high-dimensional video

clips is memory-intensive and some model-defense combinations were difficult to audit at large scale. More efficient adversarial-auditing infrastructure is needed before adaptive robustness can be evaluated routinely for video and multimodal detectors.

Finally, the threat landscape is shifting toward multimodal audio-visual forgery [31]. Future PRF evaluations should test whether multimodal fusion increases attacker cost or whether attackers can bypass the system by perturbing a single modality and exploiting weaknesses in the joint decision rule.

## 8 Conclusion

The central thesis of this Systematization of Knowledge is that the security of a face-forgery detector cannot be adequately measured by a single scalar metric. Clean generalization, static robust accuracy and final attack success rates each capture only one point in a larger operational landscape. By introducing the **Practical Robustness Frontier (PRF)**, we shift the evaluation focus from whether a detector can eventually be broken to how much attacker effort is required under different feedback and query constraints.

Our results show that detector inductive biases reshape the adversarial optimization problem. Frequency-aware detectors can expose efficient low-query attack directions when the attack basis aligns with their spectral feature prior. Attacking temporally coupled video architectures, by contrast, requires perturbations to remain effective across frames, which slows high-query convergence and increases the cost of black-box evasion. The XCLIP results further show that accepting video input alone is not sufficient, because robustness depends on how tightly the architecture couples evidence across time. We also find that current practical defenses unfortunately do not provide robustness against even moderate query budgets.

Deepfake detection should therefore be evaluated as a deployed security system rather than only as a clean classification benchmark. Future evaluations should report cost-aware attack frontiers, include zero-query and low-query threat models, measure high-query convergence and apply adaptive testing to stochastic or transformation-based defenses. Robustness claims should specify not only whether a detector can be broken, but how much interaction, feedback and computation the attacker requires.

## Ethical Considerations

This work is inherently dual-use. By systematically evaluating zero-query transfer, query-bounded black-box attacks and adaptive attacks against defenses, our study characterizes conditions under which deepfake detectors can be evaded. Such knowledge could potentially reduce the effort required by an adversary to identify weaknesses in forensic systems. At the same time, these weaknesses must be understood for defenders to evaluate whether detectors remain reliable under adversarial pressure rather than only under clean or cross-dataset evaluation.

The primary stakeholders include users affected by automated authenticity decisions, developers and operators of deepfake-detection systems, researchers studying multimedia forensics, individuals represented in benchmark face-forgery datasets and society more broadly. For system operators and users, the principal risk is that attack findings could facilitate evasion of detectors used in moderation, authentication, or provenance workflows. For dataset subjects, the relevant concerns involve responsible handling of facial and biometric data. Conversely, more realistic robustness evaluation can benefit these stakeholders by exposing security limitations before detectors are relied upon in higher-consequence settings.

We mitigate these risks in several ways. Our experiments target controlled research models and established benchmark datasets rather than commercial APIs, production moderation pipelines, or other live services. We do not evaluate or make security claims about specific commercial deployments. Results involving face-forgery datasets are reported in aggregate and the underlying datasets are used according to their respective access and usage conditions. We further frame the attacks as components of a defensive auditing methodology: the objective is to determine when detectors fail, how much adversarial interaction is required and whether defenses meaningfully shift the Practical Robustness Frontier (PRF). Our findings should be interpreted as evidence about architectural families and evaluated research systems under the stated threat models, not as a blanket assessment of all deployed deepfake detection services.

## Open Science

We provide an anonymized reviewer artifact at:

[https://anonymous.4open.science/r/SoK\\_PRF\\_review\\_artifact-7F54/](https://anonymous.4open.science/r/SoK_PRF_review_artifact-7F54/)

The artifact contains the code developed for the evaluation and analysis presented in this work, including detector catalog and wrapper infrastructure, attack and defense module interfaces, experiment planning and command generation, aggregate metric computation, PRF construction. It also includes representative experiment configurations, preprocessing and runtime scaffolding and documentation. The accompanying README provides environment setup, configuration validation, dry-run and reproduction instructions.

The artifact intentionally does not redistribute third-party datasets, pretrained detector checkpoints, or other externally maintained resources on which our evaluation builds. In particular, resources such as FaceForensics++ and DeepfakeBench remain governed by the licenses, access conditions, and distribution mechanisms established by their respective authors. Reviewers wishing to execute the full evaluation should obtain these resources directly from their original sources and provide their local dataset and checkpoint locations as described in the artifact documentation. This preserves the original authors' distribution and licensing requirements rather than repackaging their data, implementations, or pretrained models within our artifact.

We also do not release the generated adversarial versions of the evaluation sets, saved adversarial tensors, or other reusable adversarial-media collections. These outputs can be regenerated by the evaluation pipeline when the required datasets and checkpoints are supplied, while distributing them directly would both duplicate restricted source data and unnecessarily increase the dual-use potential of the artifact. Likewise, local experiment outputs, caches, and machine-specific environment files are omitted. The included sample summaries are provided to demonstrate output schemas and analysis interfaces and should not be interpreted as substitutes for the complete experimental results reported in the paper.

## References

- [1] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE international workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.
- [2] Henry Ajder, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. The state of deepfakes: Landscape, threats, and impact. *Deeptrace*, 2019.
- [3] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pages 484–501. Springer, 2020. URL: [https://www.ecva.net/papers/eccv\\_2020/papers\\_ECCV/html/4362\\_ECCV\\_2020\\_paper.php](https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/4362_ECCV_2020_paper.php), arXiv:1912.00049.
- [4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning (ICML)*, pages 274–283, 2018.
- [5] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning (ICML)*, pages 284–293. PMLR, 2018.
- [6] Armon Barton, Matthew Wright, Shaikh Akib Shahriyar, Edgar Jatho, Mohammad Saidur Rahman, Kantha Girish Gangadhara, and Jiang Ming. PadNet: Defending neural networks against adversarial examples. *ACM Transactions on Privacy and Security*, 29(2):17:1–17:26, 2026. doi:10.1145/3799889.
- [7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, pages 813–824. PMLR, 2021.
- [8] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.
- [9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017. URL: [https://nicholas.carlini.com/papers/2017\\_sp\\_nnrobustattacks.pdf](https://nicholas.carlini.com/papers/2017_sp_nnrobustattacks.pdf), doi:10.1109/SP.2017.49.
- [10] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, pages 4724–4733, 2017. URL: [https://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Carreira\\_Quo\\_Vadis\\_Action\\_CVPR\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2017/papers/Carreira_Quo_Vadis_Action_CVPR_2017_paper.pdf).
- [11] Nuria Alina Chandra, Hannah Lee, Ryan Murtfeldt, Lin Qiu, Arnab Karmakar, Emmanuel Tanumihardja, Kevin Farhat, Ben Caffee, Changyeon Lee, Jongwook Choi, Sejin Paik, Aerin Kim, and Oren Etzioni. Deepfake-eval-2024: A multi-modal in-the-wild benchmark of deepfakes circulated in 2024. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 10668–10678, 2026. URL: [https://openaccess.thecvf.com/content/CVPR2026W/APAI/html/Chandra\\_Deepfake-Eval-2024\\_A\\_Multi-Modal\\_In-the-Wild\\_Benchmark\\_of\\_Deepfakes\\_Circulated\\_in\\_2024\\_CVPRW\\_2026\\_paper.html](https://openaccess.thecvf.com/content/CVPR2026W/APAI/html/Chandra_Deepfake-Eval-2024_A_Multi-Modal_In-the-Wild_Benchmark_of_Deepfakes_Circulated_in_2024_CVPRW_2026_paper.html).
- [12] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1277–1294, 2020. URL: <https://arxiv.org/abs/1904.02144>, doi:10.1109/SP40000.2020.00045.
- [13] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. In *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, pages 30–39, 2020.
- [14] François Chollet. Xception: Deep learning with depth-wise separable convolutions. In *CVPR*, pages 1251–1258, 2017. URL: [https://openaccess.thecvf.com/content\\_cvpr\\_2017/papers/Chollet\\_Xception\\_Deep\\_Learning\\_CVPR\\_2017\\_paper.pdf](https://openaccess.thecvf.com/content_cvpr_2017/papers/Chollet_Xception_Deep_Learning_CVPR_2017_paper.pdf).
- [15] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680*, 2022. URL: <https://arxiv.org/abs/2211.00680>.
- [16] Francesco Croce and Matthias Hein. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2206–2216. PMLR, 13–18 Jul 2020. URL: <https://proceedings.mlr.press/v119/croce20b.html>.
- [17] Xinjie Cui, Yuezun Li, Ao Luo, Jiaran Zhou, and Junyu Dong. Forensics adapter: Adapting clip for generalizable face forgery detection. In *Proceedings of the*

*Computer Vision and Pattern Recognition Conference*, pages 19207–19217, 2025.

- [18] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake detection challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*, 2020. URL: <https://arxiv.org/abs/2006.07397>, arXiv:2006.07397.
- [19] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7890–7899, 2020.
- [20] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 3247–3258, 2020. URL: <https://proceedings.mlr.press/v119/frank20a/frank20a.pdf>.
- [21] Apurva Gandhi and Shomik Jain. Adversarial perturbations fool deepfake detectors. *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2020.
- [22] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL: <https://api.semanticscholar.org/CorpusID:6706414>.
- [23] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, Feiyue Huang, and Lizhuang Ma. Spatiotemporal inconsistency learning for deepfake video detection. In *Proceedings of the 29th ACM international conference on multimedia*, pages 3473–3481, 2021.
- [24] Chuan Guo, Jacob R. Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. *arXiv preprint arXiv:1905.07121*, 2019. URL: <https://arxiv.org/abs/1905.07121>, arXiv:1905.07121.
- [25] Shehzeen Hussain, Paarth Neekhara, Malhar Jere, Farinaz Koushanfar, and Cary Ellsworth. Adversarial Threats to Deepfake Detection: A Practical Perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 83–92, 2021.
- [26] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box Adversarial Attacks with Limited Queries and Information. In *International Conference on Machine Learning (ICML)*, volume 80 of *Proceedings of Machine Learning Research*, pages 2137–2146, 2018. URL: <https://proceedings.mlr.press/v80/ilyas18a/ilyas18a.pdf>.
- [27] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html](https://proceedings.neurips.cc/paper_files/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html).
- [28] iProov. New Threat Intelligence Report Exposes the Impact of Generative AI on Remote Identity Verification. <https://www.iproov.com/press/new-threat-intelligence-report-exposes-impact-generative-ai-remote-identity-verification>, February 2024. Accessed: 2026-06-11.
- [29] Felix Juefei-Xu, Run Wang, Yihao Huang, Qing Guo, Lei Ma, and Yang Liu. Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision*, 130(7):1678–1734, 2022.
- [30] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. URL: [https://openaccess.thecvf.com/content\\_CVPR\\_2019/papers/Karras\\_A\\_Style-Based\\_Generator\\_Architecture\\_for\\_Generative\\_Adversarial\\_Networks\\_CVPR\\_2019\\_paper.pdf](https://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf).
- [31] Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL: [https://datasets-benchmarks-proceedings.neurips.cc/paper\\_files/paper/2021/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract-round2.html](https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/hash/d9d4f495e875a2e075a1a4a6e1b9770f-Abstract-round2.html).
- [32] Sarwar Khan, Jun-Cheng Chen, Wen-Hung Liao, and Chu-Song Chen. Adversarially robust deepfake detection via adversarial feature similarity learning. *arXiv preprint arXiv:2403.08806*, 2024. URL: <https://arxiv.org/abs/2403.08806>.
- [33] Linda Laurier, Ave Giulietta, Arlo Octavia, and Meade Cleti. The cat and mouse game: The ongoing arms race between diffusion models and detection methods, 2024. arXiv:2410.18866, doi:10.48550/arXiv.2410.18866.

- [34] Seth Layton, Tyler Tucker, Daniel Olszewski, Kevin Warren, Kevin Butler, and Patrick Traynor. {SoK}: The good, the bad, and the unbalanced: Measuring structural limitations of deepfake media datasets. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1027–1044, 2024.
- [35] Binh M Le, Jiwon Kim, Simon S Woo, Kristen Moore, Alsharif Abuadbba, and Shahroz Tariq. SoK: Systematization and Benchmarking of Deepfake Detectors in a Unified Framework. In *2025 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, 2025.
- [36] Matan Levi and Aryeh Kontorovich. Splitting the difference on adversarial training. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3639–3656, Philadelphia, PA, August 2024. USENIX Association. URL: <https://www.usenix.org/conference/usenixsecurity24/presentation/levi>.
- [37] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5001–5010, 2020. URL: [https://openaccess.thecvf.com/content\\_CVPR\\_2020/papers/Li\\_Face\\_X-Ray\\_for\\_More\\_General\\_Face\\_Forgery\\_Detection\\_CVPR\\_2020\\_paper.pdf](https://openaccess.thecvf.com/content_CVPR_2020/papers/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.pdf).
- [38] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *CVPR Workshops*, pages 46–52, 2019. URL: [https://openaccess.thecvf.com/content\\_CVPRW\\_2019/papers/Media%20Forensics/Li\\_Exposing\\_DeepFake\\_Videos\\_By\\_Detecting\\_Face\\_Warping\\_Artifacts\\_CVPRW\\_2019\\_paper.pdf](https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Li_Exposing_DeepFake_Videos_By_Detecting_Face_Warping_Artifacts_CVPRW_2019_paper.pdf).
- [39] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3207–3216, 2020.
- [40] Jiahao Liang, Huafeng Shi, and Weihong Deng. Exploring disentangled content information for face forgery detection. In *European Conference on Computer Vision*, pages 128–145. Springer, 2022.
- [41] 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 *CVPR*, pages 772–781, 2021. URL: [https://openaccess.thecvf.com/content/CVPR2021/papers/Liu\\_Spatial-Phase\\_Shallow\\_Learning\\_Rethinking\\_Face\\_Forgery\\_Detection\\_in\\_Frequency\\_CVPR\\_2021\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2021/papers/Liu_Spatial-Phase_Shallow_Learning_Rethinking_Face_Forgery_Detection_in_Frequency_CVPR_2021_paper.pdf).
- [42] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations (ICLR)*, 2017.
- [43] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16317–16326, 2021. URL: [https://openaccess.thecvf.com/content/CVPR2021/html/Luo\\_Generalizing\\_Face\\_Forgery\\_Detection\\_With\\_High-Frequency\\_Features\\_CVPR\\_2021\\_paper.html](https://openaccess.thecvf.com/content/CVPR2021/html/Luo_Generalizing_Face_Forgery_Detection_With_High-Frequency_Features_CVPR_2021_paper.html).
- [44] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. arXiv:1706.06083 (2017 preprint). URL: <https://arxiv.org/pdf/1706.06083>.
- [45] Yisroel Mirsky and Wenke Lee. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41, 2021.
- [46] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. URL: [https://openaccess.thecvf.com/content\\_cvpr\\_2016/html/Moosavi-Dezfooli\\_DeepFool\\_A\\_Simple\\_CVPR\\_2016\\_paper.html](https://openaccess.thecvf.com/content_cvpr_2016/html/Moosavi-Dezfooli_DeepFool_A_Simple_CVPR_2016_paper.html).
- [47] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision (ECCV)*, pages 1–18. Springer, 2022.
- [48] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Amrith Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1.0.0. *arXiv preprint arXiv:1807.01069*, 2018.
- [49] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16805–16827. PMLR, 2022. URL: <https://proceedings.mlr.press/v162/nie22a.html>.
- [50] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between

- faces and their context. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44, 2022. URL: <https://doi.org/10.1109/TPAMI.2021.3093446>.
- [51] 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*, pages 506–519, 2017.
- [52] Gan Pei, Jiangning Zhang, Menghan Hu, Zhenyu Zhang, Chengjie Wang, Yunsheng Wu, Guangtao Zhai, Jian Yang, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *ACM Computing Surveys*, 58(11):1–41, 2026.
- [53] Ivan Perov et al. DeepFaceLab: Integrated, Flexible and Extensible Face-Swapping Framework. *arXiv preprint arXiv:2005.05535*, 2020. URL: <https://arxiv.org/abs/2005.05535>.
- [54] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d’Alche Buc, Emily Fox, and Hugo Larochelle. Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. *Journal of Machine Learning Research*, 22(164):1–20, 2021. URL: <https://www.jmlr.org/papers/v22/20-303.html>.
- [55] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, pages 86–103, 2020. URL: [https://www.ecva.net/papers/eccv\\_2020/papers\\_ECCV/papers/123570086.pdf](https://www.ecva.net/papers/eccv_2020/papers_ECCV/papers/123570086.pdf).
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. URL: [https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach\\_High-Resolution\\_Image\\_Synthesis\\_With\\_Latent\\_Diffusion\\_Models\\_CVPR\\_2022\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf).
- [58] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [59] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *ICCV*, 2019. URL: [https://openaccess.thecvf.com/content\\_ICCV\\_2019/papers/Rössler\\_FaceForensics\\_Learning\\_to\\_Detect\\_Manipulated\\_Facial\\_Images\\_ICCV\\_2019\\_paper.pdf](https://openaccess.thecvf.com/content_ICCV_2019/papers/Rössler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf).
- [60] Anit Kumar Sahu, Satya Narayan Shukla, Devin Willmott, and J. Zico Kolter. Simple and efficient hard label black-box adversarial attacks in low query budget regimes. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1488–1498, 2021. doi:10.1145/3447548.3467386.
- [61] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *International Conference on Learning Representations (ICLR)*, abs/1805.06605, 2018. URL: <https://api.semanticscholar.org/CorpusID:3458858>.
- [62] SAS. Study: Deepfake fraud surges – and only 7% of organizations are firmly ready. <https://tinyurl.com/5e347nxe>, 2026.
- [63] Shaikh Akib Shahriyar, Matthew Wright, and Armon Barton. IVORY: Adversarial purification of obfuscated faces to extract soft-biometrics using diffusion transformers. In *2025 IEEE 19th International Conference on Automatic Face and Gesture Recognition*, pages 1–10. IEEE, 2025. doi:10.1109/FG61629.2025.11099354.
- [64] Shaikh Akib Shahriyar and Matthew K. Wright. Evaluating robustness of sequence-based deepfake detector models by adversarial perturbation. In *Workshop on Security Implications of Deepfakes and Cheapfakes (WDC) at ACM ASIA CCS*, pages 13–18, 2022. URL: <https://dl.acm.org/doi/10.1145/3494109.3527194>, doi:10.1145/3494109.3527194.
- [65] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018. URL: <https://openreview.net/forum?id=rJUYGxbCW>.
- [66] Fnu Suya, Anshuman Suri, Tingwei Zhang, Jingtao Hong, Yuan Tian, and David Evans. SoK: Pitfalls in evaluating black-box attacks. In *2024 IEEE Conference on*

- Secure and Trustworthy Machine Learning*, pages 387–407, 2024. doi:10.1109/SaTML59370.2024.00026.
- [67] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. URL: <https://arxiv.org/abs/1312.6199>.
- [68] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, volume 97 of *PMLR*, pages 6105–6114, 2019. URL: <https://proceedings.mlr.press/v97/tan19a/tan19a.pdf>.
- [69] Shahroz Tariq, Sangyup Lee, and Simon S. Woo. One detector to rule them all: Towards a general deepfake attack detection framework. In *Proceedings of the Web Conference 2021*, pages 3625–3637, 2021. doi:10.1145/3442381.3449809.
- [70] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection. *Information Fusion*, 64:131–148, 2020.
- [71] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1633–1645, 2020.
- [72] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, 2018.
- [73] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL: <https://openreview.net/forum?id=SyxAb30cY7>, arXiv:1805.12152.
- [74] Luisa Verdoliva. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):910–932, 2020. URL: <https://arxiv.org/abs/2001.06564>, doi:10.1109/JSTSP.2020.3002101.
- [75] Ngan Hoang Vo, Khoa D. Phan, Anh-Duy Tran, and Duc-Tien Dang-Nguyen. Adversarial attacks on deepfake detectors: A practical analysis. In *MultiMedia Modeling*, pages 318–330. Springer, 2022. doi:10.1007/978-3-030-98355-0\_27.
- [76] Tianyi Wang, Xin Liao, Kam Pui Chow, Xiaodong Lin, and Yinglong Wang. Deepfake detection: A comprehensive survey from the reliability perspective. *ACM Computing Surveys*, 57(3):58:1–58:35, 2025. doi:10.1145/3699710.
- [77] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanguan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. URL: <https://openreview.net/forum?id=rkl0g6EFwS>.
- [78] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. DIRE for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023. URL: [https://openaccess.thecvf.com/content/ICCV2023/html/Wang\\_DIRE\\_for\\_Diffusion-Generated\\_Image\\_Detection\\_ICCV\\_2023\\_paper.html](https://openaccess.thecvf.com/content/ICCV2023/html/Wang_DIRE_for_Diffusion-Generated_Image_Detection_ICCV_2023_paper.html).
- [79] Ying Xu, Philipp Terhörst, Marius Pedersen, and Kiran Raja. Analyzing fairness in deepfake detection with massively annotated databases. *IEEE Transactions on Technology and Society*, 5(1):93–106, 2024.
- [80] Zhiyuan Yan, Yong Zhang, Xinhang Yuan, Siwei Lyu, and Baoyuan Wu. Deepfakebench: A comprehensive benchmark of deepfake detection. In *NeurIPS 2023 Datasets and Benchmarks Track*, 2023. URL: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets\\_and\\_Benchmarks.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/0e735e4b4f07de483cbe250130992726-Paper-Datasets_and_Benchmarks.pdf).
- [81] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482, 2019. URL: <https://proceedings.mlr.press/v97/zhang19p/zhang19p.pdf>.

## Appendix

### A Detection Architectures

**Naive CNNs.** Naive frame-level CNN detectors treat deepfake detection as binary image classification, without explicitly encoding where manipulation evidence should appear. We include three such CNNs with different capacity and design choices. **Xception** [14, 59] is a common baseline in deepfake detection and represents depthwise-separable convolutional architectures. **EfficientNet-B4** [68] represents compound-scaled CNN design. **Meso4Inception** [1] provides a compact, lower-capacity baseline, allowing us to test whether attack behavior changes when the detector has fewer parameters and weaker clean performance.

**Spatial Detectors.** Spatial detectors encode the assumption that manipulation evidence is localized in image space, for example around blending regions, warping artifacts, or identity/content inconsistencies. **FWA** [38] targets affine face-warping artifacts and blending boundaries. **IID** [40] represents disentanglement-based approaches that attempt to separate manipulation-relevant cues from identity or content features. We also include a **CLIP**-based forensic detector [17, 56] in this family, because it operates as a frame-level detector rather than a frequency or temporal model. Although it does not encode a hand-designed spatial artifact prior, its inclusion lets us test whether broad pretrained image representations behave differently from conventional spatial CNNs under transfer and query-bounded attacks.

**Frequency Detectors.** Frequency detectors encode the assumption that synthesis pipelines leave spectral or phase-domain traces that may be difficult to capture from RGB pixels alone [20]. **F3Net** [55] uses learnable frequency-aware decomposition and local frequency statistics. **SPSL** [41] targets spatial-phase artifacts associated with generative upsampling. **SRM** [43] uses high-frequency residual filters inspired by steganalysis. Although these models still use neural classifiers downstream, their frequency-domain preprocessing allows us to test whether explicit spectral feature extraction changes transferability and query efficiency.

**Video Detectors.** Video detectors aggregate evidence across frames, allowing us to test whether temporal modeling changes the cost of adversarial evasion. We instantiate this family with three representative video architectures fine-tuned for binary deepfake classification. **I3D** [10] uses 3D convolutions over temporal volumes. **TimeSformer** [7] uses divided space-time attention over video clips. **XCLIP** [47] adapts vision-language pretraining to video through cross-frame communication. This family is especially important for the PRF analysis because video models change the attack objective from fooling isolated frames to fooling a sequence-level

decision. As our results later show, however, video-family models differ substantially in how tightly they couple evidence across time.

### B DeepfakeBench–ART Integration

To perform our experiments, we addressed three significant differences between DeepfakeBench and ART.

First, DeepfakeBench routes inputs through a heavily structured `data_dict` dictionary schema (mapping metadata, labels and image tensors), whereas ART’s `PyTorchClassifier` strictly expects contiguous `numpy` arrays or raw `PyTorch` tensors. To bridge this divide, we engineered a differentiable wrapper module that dynamically unpacks ART’s raw tensor perturbations, injects them into the DeepfakeBench schema and seamlessly propagates the classification logits back to the ART optimizer without shattering the computational graph required for gradient calculation.

Second, generating high-dimensional adversarial video datasets induces catastrophic memory exhaustion (Out-Of-Memory, OOM) on standard GPU hardware. ART’s default `.generate()` function attempts to load the entire target dataset into VRAM to construct the adversarial batch. To circumvent this and enable at-scale generation, we implemented a custom “superbatching” infrastructure. This system explicitly partitions the dataset and feeds bounded batches (e.g.,  $N = 32$ ) sequentially to the ART optimizer, aggressively invoking Python’s garbage collector and clearing the CUDA cache between forward passes to stabilize VRAM utilization on our NVIDIA A100 and H100 hardware.

Third, we resolved a difference between binary deepfake classification and expectations of some attacks. Most detection models use a binary output space (Real vs. Fake), often represented by a single activated logit or two mutually exclusive symmetric logits. Meanwhile, several prominent adversarial optimizations within ART, specifically DeepFool [46] and the Difference of Logits Ratio (DLR) loss component of AutoAttack [16], mathematically assume a multi-class distribution of  $\geq 3$  classes. To ensure compatibility, we modified our AutoAttack ensemble to bypass the incompatible DLR loss, relying instead on Cross-Entropy-based Adaptive PGD (APGD-CE) and gradient-free Square Attacks.

### C Dataset Infrastructure and Isolation Schema

To manage the massive scale of the generated adversarial data and prevent catastrophic data overwriting, we engineered a rigidly isolated, hierarchical storage schema. DeepfakeBench utilizes Lightning Memory-Mapped Databases (LMDB) to accelerate I/O during baseline training. However, our systemic integration of the Adversarial Robustness Toolbox (ART) required direct, deterministic access to spatial tensors. There-

fore, while baseline training leverages the LMDB structures, our adversarial generation pipeline operates directly on the uncompressed RGB frame directories.

Following the standardized DeepfakeBench extraction protocol on the FF++ c23 dataset, our processed dataset yields deterministic frame counts: **114,884 training frames**, **22,354 validation frames** and **22,388 testing frames**. The foundational directory is structured as follows:

```

datasets/
├── lmdb/
│   ├── FaceForensics++_lmdb/
│   │   ├── data.mdb
│   │   └── lock.mdb
│   └── rgb/
│       ├── FaceForensics++/
│       │   ├── original_sequences/
│       │   │   └── youtube/c23/
│       │   │       ├── videos/*.mp4
│       │   │       ├── frames/*.png
│       │   │       └── masks/*.png
│       │   └── manipulated_sequences/
│       │       ├── Deepfakes/c23/
│       │       │   ├── videos/*.mp4
│       │       │   └── frames/*.png
│       │       ├── Face2Face/
│       │       ├── FaceSwap/
│       │       └── NeuralTextures/
└──
```

Adversarial payloads are isolated not only by dataset and attack method but strictly by the target *detector model* that the attack was optimized against. This ensures absolute zero-contamination during cross-architecture transferability evaluations:

```

adversarial_datasets/
├── FaceForensics++/
│   ├── efficientnetb4/ - Target Model
│   │   ├── PGD/ - Attack Optimizer
│   │   │   ├── frames/
│   │   │   │   ├── real/ - Adv. perturbed Real
│   │   │   │   └── fake/ - Adv. perturbed Fake
│   │   └── AutoAttack/
│   └── i3d/
└── f3net/
```

## D Attack Hyperparameters

This section reports the attack settings used to construct the PRF curves and defense evaluations. Unless otherwise specified, image intensities are normalized to  $[0, 1]$ .

**White-Box Attacks.** White-box attacks are constrained by an  $\ell_\infty$  perturbation bound of  $\epsilon = 8/255$ . For PGD-style optimization inside AutoAttack [16], we use a maximum of

Table 7: Attack settings used in the PRF evaluation.

| Attack     | Setting            | Value                           |
|------------|--------------------|---------------------------------|
| AutoAttack | norm / $\epsilon$  | $\ell_\infty, 8/255$            |
| AutoAttack | components         | APGD-CE, Square                 |
| C&W        | norms              | $\ell_2, \ell_\infty$           |
| C&W        | iterations / lr    | 100 / 0.01                      |
| SimBA-DCT  | $\epsilon$ , basis | 0.02, freq_dim=16               |
| SimBA      | query cap          | 1,000 main; 1,500 & 5,000 check |
| HSJ        | norms              | $\ell_2, \ell_\infty$           |
| HSJ        | high-query cap     | 25,000                          |
| BPDA+EOT   | Sample Count $K$   | 4, 8                            |

50 iterations with step size  $\alpha = 0.01$ . Because binary deepfake detectors can be incompatible with attack losses that assume three or more classes, we use the AutoAttack components that are well-defined for binary classification, including APGD-CE and Square Attack. Carlini–Wagner (C&W) [9] is evaluated in both  $\ell_2$  and  $\ell_\infty$  variants with confidence parameter  $c = 0$ , initial learning rate 0.01 and 100 maximum iterations.

**Black-Box Attacks.** For score-based black-box evaluation, we use SimBA [24] in both pixel and DCT bases. The SimBA-DCT configuration uses  $\epsilon = 0.02$  and freq\_dim=16. Low-query evaluations are capped at  $Q_{\max} = 1,500$  target queries per sample. High-query convergence experiments extend selected attacks to  $Q_{\max} = 25,000$ . For decision-based evaluation, we use HopSkipJump (HSJ) [12] under both  $\ell_2$  and  $\ell_\infty$  constraints and report ASR/AUC at fixed query checkpoints.

**Adaptive Attacks for Stochastic Defenses.** For defenses involving stochastic purification or non-differentiable operations, we evaluate the composite system  $g(x) = f(P_\theta(x))$  using EOT [5] and BPDA [4]. We use Monte Carlo EOT sample counts  $K \in \{4, 8\}$  and increase  $K$  when gradient estimates remain unstable. BPDA is applied only to operations that are non-differentiable or effectively non-differentiable in the backward pass, such as hard clipping, quantization, or discrete scheduler operations. Additional purifier-specific adaptive settings are reported in Appendix I.

## E Choice of Hyperparameters for SimBA-DCT Frequency Dimension

Before running the full PRF evaluation, we calibrated the two SimBA-DCT hyperparameters that most directly affect query efficiency: the per-step perturbation magnitude  $\epsilon$  and the DCT frequency dimension freq\_dim. This experiment was designed to choose a fixed SimBA-DCT configuration for all subsequent experiments.

We use a fixed, balanced 100-video subset of the FF++ c23 test split, consisting of 50 real videos and 50 fake videos. The fake videos are stratified by manipulation type and the same

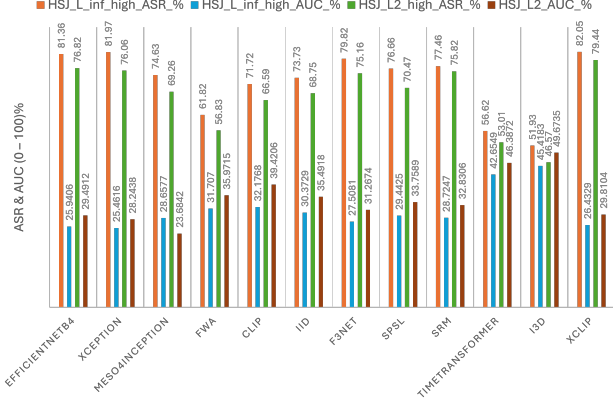


Figure 7: **HSJ- $\ell_\infty$  vs. HSJ- $\ell_2$  at 25,000 queries.** HSJ- $\ell_\infty$  yields higher ASR than HSJ- $\ell_2$  across all evaluated models, showing that norm choice materially affects high-query PRF estimates.

video IDs are reused across all  $\epsilon$ ,  $\text{freq\_dim}$  and detector settings. We evaluate five representative detectors: EfficientNet-B4, F3Net, SRM, I3D and XCLIP. These models cover a standard CNN detector, two frequency-aware detectors and two video detectors, which is sufficient for selecting an attack hyperparameter without running the full 12-model set.

We sweep

$$\epsilon \in \{0.01, 0.02, 0.05, 0.10\}, \quad \text{freq\_dim} \in \{4, 8, 16, 32\}, \quad (1)$$

and report ASR at 500 and 1,000 queries. We use two metrics. First, we report complete untargeted ASR over the balanced subset:

$$\text{ASR}_{\text{all}}(Q) = \frac{1}{|\mathcal{D}_{\text{real}}| + |\mathcal{D}_{\text{fake}}|} \sum_{x_i \in \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{fake}}} \mathbb{I}[f(A(x_i, Q)) \neq y_i]. \quad (2)$$

This measures whether a configuration makes SimBA-DCT stronger as a general black-box attack. Second, we report fake-to-real semantic ASR on the fake videos:

$$\text{ASR}_{\text{F2R}}(Q) = \frac{1}{|\mathcal{D}_{\text{fake}}|} \sum_{x_i \in \mathcal{D}_{\text{fake}}} \mathbb{I}[f(A(x_i, Q)) = \text{real}]. \quad (3)$$

This is the operationally harmful evasion direction for deepfake detection. Unless otherwise stated, both metrics are computed on clean-correct samples so that pre-existing detector errors are not counted as attack successes.

As shown in Table 8, across all tested  $\epsilon$  values,  $\text{freq\_dim} = 16$  provides the strongest query efficiency among the evaluated basis sizes. A very small number of frequency dimensions,  $\text{freq\_dim} = 4$ , is too restrictive and under-explores useful attack directions. Increasing to  $\text{freq\_dim} = 8$  substantially improves both complete ASR and fake-to-real ASR, confirming that the low-frequency channel is not specific to a single exact basis size. Increasing further to  $\text{freq\_dim} = 32$

Table 8: **Effects of SimBA-DCT step size and frequency dimension on a balanced 100 sample FF++ c23 test subset.** The subset contains 50 fake and 50 real videos. Cells report the mean across five representative detectors: EfficientNet-B4, F3Net, SRM, I3D and XCLIP.  $\text{ASR}_{\text{F2R}}$  is fake-to-real semantic ASR on fake samples;  $\text{ASR}_{\text{all}}$  is complete untargeted ASR on the balanced subset.

| $\epsilon$ | $\text{freq\_dim}$ | F2R ASR@500 | All ASR@500 | F2R ASR@1000 | All ASR@1000 |
|------------|--------------------|-------------|-------------|--------------|--------------|
| 0.01       | 4                  | 18.5        | 18.3        | 29.3         | 28.9         |
| 0.01       | 8                  | 25.6        | 26.5        | 36.2         | 35.6         |
| 0.01       | 16                 | 31.4        | 31.5        | 40.9         | 41.2         |
| 0.01       | 32                 | 29.0        | 29.5        | 39.4         | 38.5         |
| 0.02       | 4                  | 32.9        | 33.6        | 45.3         | 46.0         |
| 0.02       | 8                  | 45.0        | 46.3        | 58.9         | 59.0         |
| 0.02       | 16                 | 54.8        | 55.3        | 66.4         | 65.9         |
| 0.02       | 32                 | 51.6        | 52.7        | 63.7         | 63.8         |
| 0.05       | 4                  | 46.0        | 46.3        | 60.4         | 60.7         |
| 0.05       | 8                  | 60.5        | 61.2        | 75.4         | 75.4         |
| 0.05       | 16                 | 71.6        | 72.4        | 85.9         | 84.9         |
| 0.05       | 32                 | 66.4        | 67.3        | 79.7         | 80.9         |
| 0.10       | 4                  | 51.1        | 53.0        | 67.1         | 66.8         |
| 0.10       | 8                  | 66.0        | 67.1        | 83.0         | 83.9         |
| 0.10       | 16                 | 79.5        | 78.4        | 93.2         | 92.3         |
| 0.10       | 32                 | 72.9        | 74.6        | 88.3         | 88.4         |

does not consistently improve early-query success because the larger basis expands the coordinate search space.

The different  $\epsilon$  values provide a second calibration axis. As expected, larger  $\epsilon$  values increase ASR, but they also correspond to more aggressive per-step perturbations. We therefore select  $\epsilon = 0.02$  and  $\text{freq\_dim} = 16$  as the default configuration for the full PRF evaluation: this setting provides strong query efficiency while remaining less aggressive than  $\epsilon = 0.05$  or  $\epsilon = 0.10$ .

## F Extended Low-Query Threat Analysis Data

To ensure absolute transparency and to facilitate future reproducibility of the Practical Robustness Frontier (PRF) formalized in Section 5.2, we provide the granular data of our low-query evaluations.

The tables below detail the exact evasion trajectories of all 12 deepfake detection architectures when subjected to the query-efficient **SimBA-DCT** black-box attack [24]. Because operational threat modeling relies on precise capacity planning (e.g., setting API rate limits based on exact failure thresholds), aggregating this data into median bounds within the main text (see Table 3) obscures critical architectural outliers.

## G Adaptive Attacks on Adversarial-Training Variants

This section describes the adaptive evaluation used to stress-test the adversarial-training defenses in Section 6.1. We evaluate standard adversarial training (AT), TRADES, MART, input-gradient regularization (GradReg) and AFSL under a

Table 9: **Extended SimBA-DCT progression (ASR% / AUC-drop pp)**. Each cell reports **cumulative ASR (%) / AUC-drop (pp)** at the specified query budget  $Q$ . This joint view highlights early-break vulnerability (rapid ASR growth with steep AUC degradation) and the persistent low-budget resilience of pure 3D convolutions (I3D).

| Architectural Target |                        | ASR(%) / AUC-drop(pp) at Query Budget $Q$ |           |             |             |             |             |             |             |
|----------------------|------------------------|-------------------------------------------|-----------|-------------|-------------|-------------|-------------|-------------|-------------|
| Family               | Model                  | $Q=10$                                    | $Q=25$    | $Q=50$      | $Q=100$     | $Q=250$     | $Q=500$     | $Q=750$     | $Q=1000$    |
| Naive                | <b>EfficientNet-B4</b> | 2.58/2.51                                 | 6.71/6.42 | 14.59/13.81 | 26.49/25.58 | 37.55/30.16 | 43.87/33.12 | 48.10/36.33 | 52.33/39.54 |
|                      | <b>Xception</b>        | 2.16/2.30                                 | 5.06/6.16 | 13.87/13.69 | 26.24/26.13 | 38.87/30.75 | 44.38/34.50 | 49.19/37.90 | 53.00/40.29 |
|                      | <b>Meso4Inception</b>  | 1.40/1.87                                 | 4.91/3.06 | 10.18/7.69  | 20.09/14.53 | 31.69/17.05 | 36.41/19.19 | 41.68/22.53 | 45.94/25.86 |
| Spatial              | <b>FWA</b>             | 0.93/3.25                                 | 2.87/8.41 | 6.56/17.25  | 14.74/35.37 | 16.89/36.53 | 18.51/37.40 | 24.81/40.74 | 31.11/44.07 |
|                      | <b>CLIP</b>            | 1.62/1.30                                 | 5.31/4.24 | 12.55/9.97  | 28.94/22.94 | 34.02/26.91 | 37.86/29.91 | 43.88/33.86 | 49.89/37.80 |
|                      | <b>IID</b>             | 3.13/2.89                                 | 8.43/7.74 | 17.78/16.22 | 37.38/33.96 | 39.79/35.90 | 41.61/37.46 | 46.61/40.85 | 51.61/44.23 |
| Frequency            | <b>F3Net</b>           | 3.34/3.27                                 | 9.20/8.93 | 19.70/18.98 | 41.98/40.21 | 45.40/43.18 | 47.99/45.43 | 52.56/48.93 | 57.12/52.43 |
|                      | <b>SPSL</b>            | 2.80/2.89                                 | 7.87/7.75 | 17.05/16.30 | 36.73/34.21 | 40.19/36.35 | 42.82/37.96 | 46.34/40.99 | 49.86/44.01 |
|                      | <b>SRM</b>             | 3.28/3.49                                 | 9.07/9.29 | 19.45/19.43 | 41.53/40.55 | 44.97/42.78 | 47.62/44.47 | 52.36/48.25 | 57.10/52.03 |
| Video                | <b>I3D</b>             | 0.38/0.12                                 | 2.24/1.17 | 6.43/3.64   | 16.74/9.92  | 21.88/13.39 | 25.77/16.01 | 30.46/19.45 | 35.15/22.89 |
|                      | <b>TimeSformer</b>     | 0.44/0.33                                 | 2.63/1.74 | 7.18/4.83   | 18.09/12.06 | 24.37/15.94 | 28.92/18.52 | 31.40/18.34 | 33.90/21.78 |
|                      | <b>XCLIP</b>           | 3.62/3.45                                 | 9.90/9.38 | 21.06/19.90 | 44.64/42.06 | 47.98/45.06 | 50.51/47.33 | 55.46/50.09 | 60.41/52.84 |

Table 10: Mode sensitivity across attacks (family means). We report mean **ASR (%)** and mean **attacked-set AUC (%)** at matched budgets. SimBA-DCT and HSJ- $\ell_\infty$  consistently yield higher ASR and lower attacked-set AUC than their Pixel/ $\ell_2$  counterparts.

| Attack                         | Model Family  | ASR % (DCT/ $\ell_\infty$ ) | ASR % (Pixel/ $\ell_2$ ) | AUC % (DCT/ $\ell_\infty$ ) | AUC % (Pixel/ $\ell_2$ ) |
|--------------------------------|---------------|-----------------------------|--------------------------|-----------------------------|--------------------------|
| SimBA-DCT vs Pixel             | Naive Avg     | 67.21                       | 57.83                    | 36.57                       | 43.36                    |
| SimBA-DCT vs Pixel             | Spatial Avg   | 54.05                       | 43.62                    | 41.13                       | 48.60                    |
| SimBA-DCT vs Pixel             | Frequency Avg | 65.58                       | 55.34                    | 37.90                       | 45.58                    |
| SimBA-DCT vs Pixel             | Video Avg     | 68.50                       | 62.43                    | 34.19                       | 38.18                    |
| HSJ- $\ell_\infty$ vs $\ell_2$ | Naive Avg     | 80.00                       | 75.40                    | 26.62                       | 27.81                    |
| HSJ- $\ell_\infty$ vs $\ell_2$ | Spatial Avg   | 69.09                       | 64.06                    | 31.42                       | 36.96                    |
| HSJ- $\ell_\infty$ vs $\ell_2$ | Frequency Avg | 77.98                       | 73.82                    | 28.56                       | 32.62                    |
| HSJ- $\ell_\infty$ vs $\ell_2$ | Video Avg     | 54.28                       | 49.79                    | 44.04                       | 48.03                    |

direct fake-to-real adaptive attack. Unlike PadNet and DBAT, these defenses do not introduce explicit padding, rejection, or adversarial-companion outputs. Therefore, an adaptive attack does not need to reason about rejection states; it only needs to maximize the harmful semantic decision that a fake input is classified as real.

## G.1 Threat Model and Attack Objective

Let  $z_{\text{real}}(x)$  and  $z_{\text{fake}}(x)$  denote the detector logits for the real and fake classes. For each fake input  $x$ , the adaptive objective directly maximizes the real-vs-fake margin:

$$J_{\text{AT}}(x) = z_{\text{real}}(x) - z_{\text{fake}}(x), \quad \|\delta\|_\infty \leq \epsilon. \quad (4)$$

An attack is counted as successful only when the final perturbed fake input is classified as real. Since there is no rejection state for these defenses, the reported correct robust rate is simply the complement of semantic ASR.

We use AutoAttack to reduce dependence on manually tuned attack hyperparameters [16]. In our binary Real/Fake

setting, the DLR component is not well-defined for the two-class output used by several detectors, so we use the binary-compatible subset described in Appendix D: APGD-CE together with Square Attack. We sweep  $\epsilon \in \{2/255, 4/255, 6/255, 8/255, 12/255\}$  and report the standard comparison point at  $\epsilon = 8/255$ .

## G.2 Family-Level Results

Table 14 reports family-mean semantic ASR across the  $\epsilon$  sweep. The results show three consistent patterns. First, MART provides the lowest fake-to-real ASR across all families at  $\epsilon = 8/255$ , matching its role as the most false-accept-focused adversarial-training variant. Second, TRADES improves over standard AT while preserving a more utility-oriented trade-off. Third, GradReg behaves differently across architectures: it does not provide the strongest frame-level robustness, but it gives the lowest video-family ASR among the evaluated variants at  $\epsilon = 8/255$ , consistent with its PRF benefit for temporally coupled models. AFSL shows the op-

Table 11: Adversarial purification under **white-box gradient attacks on the base detector  $f$**  (non-adaptive to  $P_\theta$ ). AutoAttack and CW adversarial examples are generated against  $f$ , then purified, then evaluated. Robust performance is restored close to the clean (purified) baseline.

| Model     | Defense  | CleanAcc | CleanAUC | AA Acc | AA AUC | AA ASR% | CW Acc | CW AUC | CW ASR% |
|-----------|----------|----------|----------|--------|--------|---------|--------|--------|---------|
| EffNetB4  | IVORY    | 0.925    | 0.970    | 0.886  | 0.915  | 11.4    | 0.894  | 0.923  | 10.6    |
| EffNetB4  | DiffPure | 0.925    | 0.970    | 0.859  | 0.893  | 14.1    | 0.874  | 0.905  | 12.6    |
| Xception  | IVORY    | 0.932    | 0.976    | 0.892  | 0.914  | 10.8    | 0.902  | 0.925  | 9.8     |
| Xception  | DiffPure | 0.932    | 0.976    | 0.864  | 0.897  | 13.6    | 0.881  | 0.907  | 11.9    |
| Meso4Inc  | IVORY    | 0.813    | 0.786    | 0.728  | 0.653  | 27.2    | 0.819  | 0.661  | 18.1    |
| Meso4Inc  | DiffPure | 0.813    | 0.786    | 0.713  | 0.631  | 28.7    | 0.801  | 0.646  | 19.9    |
| FWA       | IVORY    | 0.657    | 0.940    | 0.791  | 0.861  | 20.9    | 0.799  | 0.873  | 20.1    |
| FWA       | DiffPure | 0.657    | 0.940    | 0.759  | 0.836  | 24.1    | 0.774  | 0.847  | 22.6    |
| CLIP      | IVORY    | 0.824    | 0.840    | 0.831  | 0.891  | 16.9    | 0.841  | 0.903  | 15.9    |
| CLIP      | DiffPure | 0.824    | 0.840    | 0.799  | 0.866  | 20.1    | 0.814  | 0.881  | 18.6    |
| IID       | IVORY    | 0.877    | 0.917    | 0.809  | 0.864  | 19.1    | 0.821  | 0.879  | 17.9    |
| IID       | DiffPure | 0.877    | 0.917    | 0.786  | 0.839  | 21.4    | 0.799  | 0.856  | 20.1    |
| F3Net     | IVORY    | 0.938    | 0.978    | 0.901  | 0.921  | 9.9     | 0.909  | 0.929  | 9.1     |
| F3Net     | DiffPure | 0.938    | 0.978    | 0.876  | 0.903  | 12.4    | 0.891  | 0.915  | 10.9    |
| SPSL      | IVORY    | 0.880    | 0.929    | 0.801  | 0.841  | 19.9    | 0.811  | 0.851  | 18.9    |
| SPSL      | DiffPure | 0.880    | 0.929    | 0.776  | 0.819  | 22.4    | 0.789  | 0.834  | 21.1    |
| SRM       | IVORY    | 0.918    | 0.965    | 0.874  | 0.901  | 12.6    | 0.884  | 0.909  | 11.6    |
| SRM       | DiffPure | 0.918    | 0.965    | 0.851  | 0.879  | 14.9    | 0.867  | 0.891  | 13.3    |
| TimeTrans | IVORY    | 0.924    | 0.969    | 0.941  | 0.956  | 5.9     | 0.949  | 0.964  | 5.1     |
| TimeTrans | DiffPure | 0.924    | 0.969    | 0.919  | 0.941  | 8.1     | 0.931  | 0.947  | 6.9     |
| I3D       | IVORY    | 0.821    | 0.968    | 0.756  | 0.766  | 24.4    | 0.764  | 0.874  | 23.6    |
| I3D       | DiffPure | 0.821    | 0.968    | 0.724  | 0.741  | 27.6    | 0.739  | 0.854  | 26.1    |
| XCLIP     | IVORY    | 0.912    | 0.959    | 0.886  | 0.919  | 11.4    | 0.896  | 0.931  | 10.4    |
| XCLIP     | DiffPure | 0.912    | 0.959    | 0.861  | 0.894  | 13.9    | 0.876  | 0.909  | 12.4    |

posite profile: it improves query-bounded black-box behavior in the main PRF analysis, but remains highly vulnerable to direct white-box semantic optimization.

The adaptive evaluation clarifies the regime-dependent nature of adversarial training. MART is the strongest option when the operational goal is suppressing fake-to-real false accepts: it has the lowest semantic ASR for every detector family at  $\epsilon = 8/255$  and remains the lowest across the  $\epsilon$  sweep. TRADES provides a more balanced profile, improving over standard AT while retaining the clean accuracy. GradReg does not dominate under direct white-box semantic optimization for frame-level or frequency-aware detectors, but it remains comparatively stronger for the video family than for the frame-level families, which could be interpreted that gradient smoothing is most useful when the model aggregates evidence across time. AFSL has the weakest adaptive white-box performance among the stronger defenses at  $\epsilon = 8/255$ , specially on frequency-aware detectors, which is consistent with our PRF result that AFSL primarily shifts query-bounded black-box behavior rather than eliminating white-box vulnerability. Overall, these results reinforce Remark 5 that adversarial training is not a single defense profile and the best variant depends on whether the deployment objective is lower false accepts, clean utility, slower black-box convergence, or resistance to direct semantic white-box attack.

## H Adaptive Attacks on Boundary-Shaping Defenses

This section describes the adaptive evaluation used to stress-test the boundary-shaping defenses in Section 6.2. Our goal is not only to measure whether PadNet and DBAT reduce aggregate attack success, but also to determine whether their explicit boundary regions can be bypassed by an attacker that knows the defense design. To check sensitivity to attack strength, we swept the main attack hyperparameters, including perturbation budget  $\epsilon$ , optimization steps, restarts and step size. Specifically, we use  $\epsilon \in \{2/255, 4/255, 6/255, 8/255, 12/255\}$ , steps  $\in \{50, 100, 200\}$ , restarts  $\in \{5, 10\}$  and step sizes  $\in \{1/255, 2/255, 4/255\}$ . Unless otherwise stated, the reported tables use the standard white-box perturbation budget  $\epsilon = 8/255$  with 200 optimization steps, 10 restarts and step size  $2/255$ . We use this as the main setting because it gives the strongest stable first-sweep attack configuration: 200 steps and 10 restarts reduce under-optimization, while a  $2/255$  step size provides more reliable convergence than the smaller  $1/255$  setting without the overshooting risk of the more aggressive  $4/255$  setting. We additionally report the  $\epsilon$  sweep for PadNet in Figure 8 to show how semantic attack success changes as the perturbation budget increases.

Table 12: Adversarial purification under **query-bounded black-box attacks**. Unlike the white-box results, robustness gains here are modest: query-efficient attacks can exploit directions that survive purification, particularly low-frequency / semantically consistent perturbations.

| Model     | Defense  | CleanAcc | CleanAUC | HSJ Acc | HSJ AUC | HSJ ASR% | SimBA Acc | SimBA AUC | SimBA ASR% |
|-----------|----------|----------|----------|---------|---------|----------|-----------|-----------|------------|
| EffNetB4  | IVORY    | 0.925    | 0.970    | 0.424   | 0.401   | 57.6     | 0.481     | 0.442     | 51.9       |
| EffNetB4  | DiffPure | 0.925    | 0.970    | 0.451   | 0.420   | 54.9     | 0.510     | 0.472     | 49.0       |
| Xception  | IVORY    | 0.932    | 0.976    | 0.438   | 0.413   | 56.2     | 0.494     | 0.460     | 50.6       |
| Xception  | DiffPure | 0.932    | 0.976    | 0.467   | 0.432   | 53.3     | 0.513     | 0.480     | 48.7       |
| Meso4Inc  | IVORY    | 0.813    | 0.786    | 0.368   | 0.346   | 63.2     | 0.415     | 0.397     | 58.5       |
| Meso4Inc  | DiffPure | 0.813    | 0.786    | 0.399   | 0.376   | 60.1     | 0.440     | 0.411     | 56.0       |
| FWA       | IVORY    | 0.657    | 0.940    | 0.382   | 0.356   | 61.8     | 0.429     | 0.396     | 57.1       |
| FWA       | DiffPure | 0.657    | 0.940    | 0.414   | 0.385   | 58.6     | 0.460     | 0.432     | 54.0       |
| CLIP      | IVORY    | 0.824    | 0.840    | 0.428   | 0.401   | 57.2     | 0.476     | 0.445     | 52.4       |
| CLIP      | DiffPure | 0.824    | 0.840    | 0.459   | 0.424   | 54.1     | 0.507     | 0.474     | 49.3       |
| IID       | IVORY    | 0.877    | 0.917    | 0.390   | 0.362   | 61.0     | 0.435     | 0.406     | 56.5       |
| IID       | DiffPure | 0.877    | 0.917    | 0.417   | 0.384   | 58.3     | 0.462     | 0.431     | 53.8       |
| F3Net     | IVORY    | 0.938    | 0.978    | 0.436   | 0.407   | 56.4     | 0.495     | 0.456     | 50.5       |
| F3Net     | DiffPure | 0.938    | 0.978    | 0.465   | 0.434   | 53.5     | 0.512     | 0.482     | 48.8       |
| SPSL      | IVORY    | 0.880    | 0.929    | 0.362   | 0.335   | 63.8     | 0.410     | 0.386     | 59.0       |
| SPSL      | DiffPure | 0.880    | 0.929    | 0.391   | 0.364   | 60.9     | 0.437     | 0.404     | 56.3       |
| SRM       | IVORY    | 0.918    | 0.965    | 0.442   | 0.416   | 55.8     | 0.497     | 0.465     | 50.3       |
| SRM       | DiffPure | 0.918    | 0.965    | 0.471   | 0.437   | 52.9     | 0.520     | 0.485     | 48.0       |
| TimeTrans | IVORY    | 0.924    | 0.969    | 0.388   | 0.361   | 61.2     | 0.430     | 0.402     | 57.0       |
| TimeTrans | DiffPure | 0.924    | 0.969    | 0.421   | 0.392   | 57.9     | 0.469     | 0.436     | 53.1       |
| I3D       | IVORY    | 0.821    | 0.968    | 0.392   | 0.360   | 60.8     | 0.431     | 0.401     | 56.9       |
| I3D       | DiffPure | 0.821    | 0.968    | 0.419   | 0.391   | 58.1     | 0.470     | 0.447     | 53.0       |
| XCLIP     | IVORY    | 0.912    | 0.959    | 0.396   | 0.372   | 60.4     | 0.444     | 0.412     | 55.6       |
| XCLIP     | DiffPure | 0.912    | 0.959    | 0.433   | 0.402   | 56.7     | 0.482     | 0.451     | 51.8       |

## H.1 Threat Model and Metrics

We evaluate fake-to-real evasion. For each fake input, an attack is counted as semantically successful only if the final prediction is mapped to the real class. Rejection outcomes are not counted as semantic success. We report six metrics:

We evaluate fake-to-real evasion. For each fake input, an attack is counted as semantically successful only if the final prediction is mapped to the real class. Rejection outcomes are not counted as semantic success. We report three final-outcome metrics:

- **Semantic ASR**: the percentage of fake inputs classified as real.
- **Padding/rejection rate**: the percentage of fake inputs classified into a padding, rejection, or adversarial-companion rejection state at the end of the attack.
- **Correct robust rate**: the percentage of fake inputs that remain classified as fake.

For PadNet, the padding region is the explicit padding/barrier class. For DBAT, there is no literal padding class; therefore, we use “padding-entry” and “padding-crossing” to refer to entry into the adversarial

companion classes. Under the rejection-aware DBAT setting, both adversarial companion outputs are treated as rejection.

For PadNet, the padding region is the explicit padding/barrier class. For DBAT, there is no literal padding class; therefore, the rejection column refers to adversarial-companion outputs when they are treated as rejection. Under the semantic-companion DBAT setting, clean and adversarial companion classes are collapsed back to the same semantic label, so the final rejection rate is zero.

## H.2 PadNet Adaptive Attacks

PadNet introduces a padding class between ordinary class regions. Let  $z_{\text{real}}(x)$ ,  $z_{\text{fake}}(x)$  and  $z_{\text{pad}}(x)$  denote the logits of the real, fake and padding classes. We evaluate two defense-aware attacks.

**Pad-gradient AutoAttack.** The first attack follows the adaptive strategy used in the PadNet paper: it uses the padding class as the optimization signal. In our binary deepfake setting, this objective can be written as a padding-entry margin,

$$J_{\text{pad}}(x) = z_{\text{pad}}(x) - z_{\text{fake}}(x). \quad (5)$$

This encourages the perturbation to move from the fake class toward the padding region. However, the attack is only

Table 13: **PadNet vs. DBAT across 12 deepfake detectors.** Columns report clean AUC for the baseline and defended models; budgeted black-box robustness under SimBA-DCT at 1,500 queries and HSJ- $\ell_\infty$  at the high-query setting; and  $Q_{95}$  under SimBA-DCT. In this evaluation, PadNet generally incurs a smaller clean-AUC penalty, produces lower attacked ASR and yields larger  $Q_{95}$  gains than DBAT.

| Model       | Clean AUC |        |       | SimBA-DCT @ 1.5k |                  |                  |                  | HSJ- $\ell_\infty$ (High) |                  |                  |                  | $Q_{95}$ (SimBA) |        |      |
|-------------|-----------|--------|-------|------------------|------------------|------------------|------------------|---------------------------|------------------|------------------|------------------|------------------|--------|------|
|             | Base      | PadNet | DBAT  | ASR <sub>P</sub> | AUC <sub>P</sub> | ASR <sub>D</sub> | AUC <sub>D</sub> | ASR <sub>P</sub>          | AUC <sub>P</sub> | ASR <sub>D</sub> | AUC <sub>D</sub> | Base             | PadNet | DBAT |
| EffNetB4    | 0.970     | 0.960  | 0.944 | 0.58             | 0.386            | 0.62             | 0.364            | 0.78                      | 0.277            | 0.80             | 0.266            | 1.2k             | 1.6k   | 1.3k |
| Xception    | 0.976     | 0.961  | 0.949 | 0.57             | 0.397            | 0.62             | 0.352            | 0.77                      | 0.275            | 0.82             | 0.262            | 1.3k             | 1.7k   | 1.5k |
| Meso4Inc    | 0.786     | 0.771  | 0.761 | 0.52             | 0.412            | 0.56             | 0.406            | 0.71                      | 0.316            | 0.74             | 0.277            | 1.5k             | 1.9k   | 1.7k |
| FWA         | 0.939     | 0.932  | 0.918 | 0.55             | 0.389            | 0.60             | 0.378            | 0.80                      | 0.273            | 0.82             | 0.250            | 1.4k             | 1.8k   | 1.6k |
| CLIP        | 0.839     | 0.830  | 0.814 | 0.51             | 0.423            | 0.57             | 0.393            | 0.75                      | 0.305            | 0.79             | 0.252            | 1.4k             | 1.7k   | 1.5k |
| IID         | 0.917     | 0.909  | 0.898 | 0.53             | 0.411            | 0.56             | 0.396            | 0.76                      | 0.266            | 0.79             | 0.268            | 1.5k             | 1.8k   | 1.7k |
| F3Net       | 0.978     | 0.966  | 0.954 | 0.56             | 0.403            | 0.60             | 0.380            | 0.76                      | 0.291            | 0.78             | 0.286            | 1.2k             | 1.6k   | 1.4k |
| SPSL        | 0.929     | 0.919  | 0.904 | 0.57             | 0.374            | 0.61             | 0.351            | 0.77                      | 0.281            | 0.81             | 0.275            | 1.3k             | 1.7k   | 1.4k |
| SRM         | 0.965     | 0.952  | 0.944 | 0.55             | 0.391            | 0.58             | 0.387            | 0.75                      | 0.287            | 0.80             | 0.250            | 1.0k             | 1.3k   | 1.1k |
| XCLIP       | 0.959     | 0.955  | 0.944 | 0.62             | 0.351            | 0.68             | 0.332            | 0.80                      | 0.250            | 0.82             | 0.250            | 1.4k             | 1.7k   | 1.5k |
| TimeSformer | 0.969     | 0.954  | 0.940 | 0.44             | 0.441            | 0.48             | 0.418            | 0.64                      | 0.361            | 0.68             | 0.333            | 2.2k             | 2.6k   | 2.4k |
| I3D         | 0.968     | 0.959  | 0.931 | 0.38             | 0.492            | 0.41             | 0.478            | 0.52                      | 0.401            | 0.55             | 0.413            | 2.7k             | 3.0k   | 2.8k |

counted as semantically successful if the final prediction is real, not pad. This distinction matters because a pad-targeted trajectory may stop in rejection rather than crossing into the harmful class.

**Rejection-aware margin AutoAttack.** The second attack directly optimizes the harmful operational condition. It treats the padding class as a competing rejection class and maximizes the real logit over both fake and pad:

$$J_{\text{RA}}(x) = z_{\text{real}}(x) - \max\{z_{\text{fake}}(x), z_{\text{pad}}(x)\}. \quad (6)$$

This objective is stronger for the deployed binary detector because it does not treat reaching pad as progress unless the trajectory later crosses into real. Table 15 reports the standard  $\epsilon = 8/255$  operating point, while Figure 8 shows the full  $\epsilon$  sweep. Across all families, the rejection-aware attack produces higher semantic ASR than the pad-gradient attack and the gap widens as the perturbation budget increases.

### H.3 DBAT Adaptive Attacks

DBAT reshapes the binary decision problem by splitting each semantic class into clean and adversarial companion classes. Let the four logits be  $z_{\text{real,clean}}$ ,  $z_{\text{real,adv}}$ ,  $z_{\text{fake,clean}}$  and  $z_{\text{fake,adv}}$ . We evaluate two operational interpretations.

**Semantic-companion AutoAttack.** In the first interpretation, clean and adversarial companion classes share the same semantic label at inference time. A fake input is counted as real if either real companion class wins. The attack objective is

$$J_{\text{sem}}(x) = \log\left(e^{z_{\text{real,clean}}(x)} + e^{z_{\text{real,adv}}(x)}\right) - \log\left(e^{z_{\text{fake,clean}}(x)} + e^{z_{\text{fake,adv}}(x)}\right). \quad (7)$$

This is the natural semantic evaluation if DBAT companion classes are collapsed back to the original real/fake labels. In this case, there is no operational rejection class, so the final rejection rate is zero. Companion-entry and companion-crossing are still recorded as trajectory diagnostics.

**Rejection-aware clean-real AutoAttack.** In the second interpretation, adversarial companion classes are treated as rejection states. This is stricter from a deployment perspective: a fake input is semantically successful only if it is classified as *real clean*. The attacker therefore maximizes

$$J_{\text{clean-real}}(x) = z_{\text{real,clean}}(x) - \max\{z_{\text{fake,clean}}(x), z_{\text{fake,adv}}(x), z_{\text{real,adv}}(x)\}. \quad (8)$$

This directly answers whether the harmful target class beats both the correct class and the rejection/companion classes. Table 15 reports the standard  $\epsilon = 8/255$  operating point, while Figure 9 shows the full  $\epsilon$  sweep. Across all families, the rejection-aware clean-real objective produces higher semantic ASR than the semantic-companion objective and the gap persists as the perturbation budget increases.

The adaptive results support three conclusions. First, boundary shaping changes attack dynamics: both PadNet and DBAT redirect a non-trivial fraction of attacks into padding, rejection, or adversarial-companion states. Second, rejection-aware objectives are stronger than boundary-entry

Table 14: **Adaptive AutoAttack evaluation of adversarial-training variants by detector family.** Cells report family-mean fake-to-real semantic ASR (%) under the adaptive attack.

| Defense | Family     | 2/255 ASR | 4/255 ASR | 6/255 ASR | 8/255 ASR | 12/255 ASR | Correct @ 8/255 |
|---------|------------|-----------|-----------|-----------|-----------|------------|-----------------|
| AT      | Naive CNNs | 23.4      | 41.8      | 55.7      | 67.4      | 76.6       | 32.9            |
| AT      | Spatial    | 17.6      | 35.2      | 49.4      | 60.7      | 71.8       | 39.3            |
| AT      | Frequency  | 28.1      | 46.7      | 61.3      | 71.5      | 80.7       | 28.5            |
| AT      | Video      | 7.6       | 26.4      | 40.9      | 52.1      | 62.8       | 47.9            |
| TRADES  | Naive CNNs | 17.2      | 35.7      | 50.4      | 61.1      | 71.2       | 38.9            |
| TRADES  | Spatial    | 11.9      | 31.4      | 44.2      | 56.3      | 66.9       | 43.7            |
| TRADES  | Frequency  | 21.3      | 40.1      | 54.3      | 65.4      | 75.6       | 34.9            |
| TRADES  | Video      | 4.6       | 23.1      | 37.4      | 48.5      | 58.9       | 51.5            |
| MART    | Naive CNNs | 8.7       | 26.6      | 41.8      | 52.7      | 63.2       | 47.3            |
| MART    | Spatial    | 3.4       | 22.8      | 36.6      | 49.1      | 59.6       | 50.9            |
| MART    | Frequency  | 14.2      | 32.7      | 45.8      | 58.6      | 68.5       | 41.4            |
| MART    | Video      | 2.6       | 17.4      | 30.6      | 42.2      | 52.8       | 57.8            |
| GradReg | Naive CNNs | 26.4      | 45.9      | 58.8      | 70.9      | 80.2       | 29.1            |
| GradReg | Spatial    | 20.1      | 37.6      | 52.1      | 64.7      | 74.3       | 35.3            |
| GradReg | Frequency  | 30.6      | 47.7      | 63.5      | 73.8      | 83.6       | 26.2            |
| GradReg | Video      | 6.1       | 18.2      | 32.9      | 44.7      | 54.6       | 55.3            |
| AFSL    | Naive CNNs | 25.8      | 44.9      | 58.3      | 69.8      | 79.7       | 30.2            |
| AFSL    | Spatial    | 21.6      | 38.7      | 53.2      | 65.1      | 74.7       | 34.9            |
| AFSL    | Frequency  | 31.4      | 51.0      | 63.9      | 76.2      | 84.5       | 23.8            |
| AFSL    | Video      | 10.6      | 30.7      | 43.9      | 56.4      | 65.7       | 43.6            |

or semantic-companion objectives because they optimize the deployed fake-to-real failure condition directly. For PadNet, the rejection-aware margin attack increases family-level semantic ASR by roughly 11–14 percentage points at  $\epsilon = 8/255$  compared with the pad-gradient attack. For DBAT, treating adversarial companions as rejection and targeting only the clean real class increases family-level semantic ASR by roughly 10–13 percentage points at  $\epsilon = 8/255$  compared with the semantic-companion objective. The  $\epsilon$  sweeps in Figures 8 and 9 show that these gaps persist across perturbation budgets. Third, the same architectural trend seen in the main PRF evaluation persists: frequency-aware detectors remain the easiest family to attack, while temporally coupled video models have lower semantic ASR. Thus, adaptive boundary evaluation reinforces the main conclusion of Section 6.2: boundary shaping shifts the PRF and increases attacker effort, but it should not be interpreted as a complete defense against a defense-aware adversary.

## I Adaptive Evaluation of Purification Defenses

Purification defenses introduce a stochastic, multi-step transformation  $P_\theta(\cdot)$  before the detector  $f(\cdot)$ , yielding the composed system  $g(x) = f(P_\theta(x))$ . Because this transformation can make gradients noisy, discontinuous, or uninformative,

robustness estimates based on standard gradient attacks can overstate the defense’s effectiveness [4]. We therefore evaluate purifiers with adaptive attacks that optimize against the composed purifier–detector system rather than the base detector alone. Specifically, we use Expectation over Transformation (EOT) to estimate the attack objective over multiple purification samples and Backward Pass Differentiable Approximation (BPDA) for operations that are non-differentiable or difficult to differentiate through in the backward pass [5, 8, 71]. In our experiments, we report BPDA+EOT results for  $K \in \{4, 8\}$  samples to test whether conclusions are stable as the adaptive attack receives a stronger expectation estimate.

PadNet adaptive attacks: family-median semantic ASR vs. perturbation budget

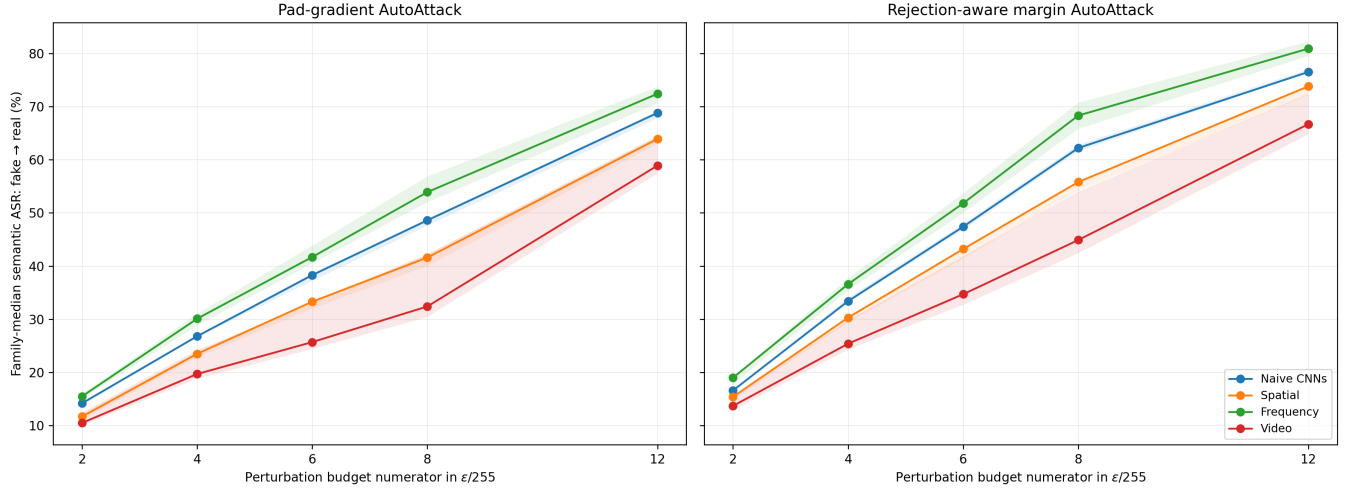


Figure 8: **PadNet adaptive attack sensitivity to perturbation budget.** Curves show family-median fake-to-real semantic ASR as  $\epsilon$  varies from 2/255 to 12/255 under the main optimizer setting (200 steps, 10 restarts and step size 2/255). Shaded bands show the interquartile range across models within each family. The rejection-aware margin objective consistently produces higher semantic ASR than the pad-gradient objective, especially at larger perturbation budgets.

Table 15: **Adaptive PadNet results by detector family.** Results use  $\epsilon = 8/255$ , 200 steps, 10 restarts and step size 2/255.

| Defense | Attack                     | Family     | Semantic ASR | Pad/Rej. | Correct Robust |
|---------|----------------------------|------------|--------------|----------|----------------|
| PadNet  | Pad-gradient               | Naive CNNs | 48.3         | 28.7     | 23.0           |
| PadNet  | Pad-gradient               | Spatial    | 41.4         | 31.1     | 27.5           |
| PadNet  | Pad-gradient               | Frequency  | 54.1         | 24.8     | 21.1           |
| PadNet  | Pad-gradient               | Video      | 37.8         | 32.6     | 29.6           |
| PadNet  | Rejection-aware margin     | Naive CNNs | 61.6         | 15.1     | 23.3           |
| PadNet  | Rejection-aware margin     | Spatial    | 54.8         | 18.5     | 26.7           |
| PadNet  | Rejection-aware margin     | Frequency  | 68.4         | 10.6     | 21.0           |
| PadNet  | Rejection-aware margin     | Video      | 49.2         | 22.7     | 28.1           |
| DBAT    | Semantic-companion         | Naive CNNs | 55.7         | 0.0      | 44.3           |
| DBAT    | Semantic-companion         | Spatial    | 49.4         | 0.0      | 50.6           |
| DBAT    | Semantic-companion         | Frequency  | 62.9         | 0.0      | 37.1           |
| DBAT    | Semantic-companion         | Video      | 48.3         | 0.0      | 51.7           |
| DBAT    | Rejection-aware clean-real | Naive CNNs | 67.0         | 13.6     | 19.4           |
| DBAT    | Rejection-aware clean-real | Spatial    | 62.3         | 15.9     | 21.8           |
| DBAT    | Rejection-aware clean-real | Frequency  | 73.6         | 9.6      | 16.8           |
| DBAT    | Rejection-aware clean-real | Video      | 58.9         | 19.7     | 21.4           |

DBAT adaptive attacks: family-median semantic ASR vs. perturbation budget

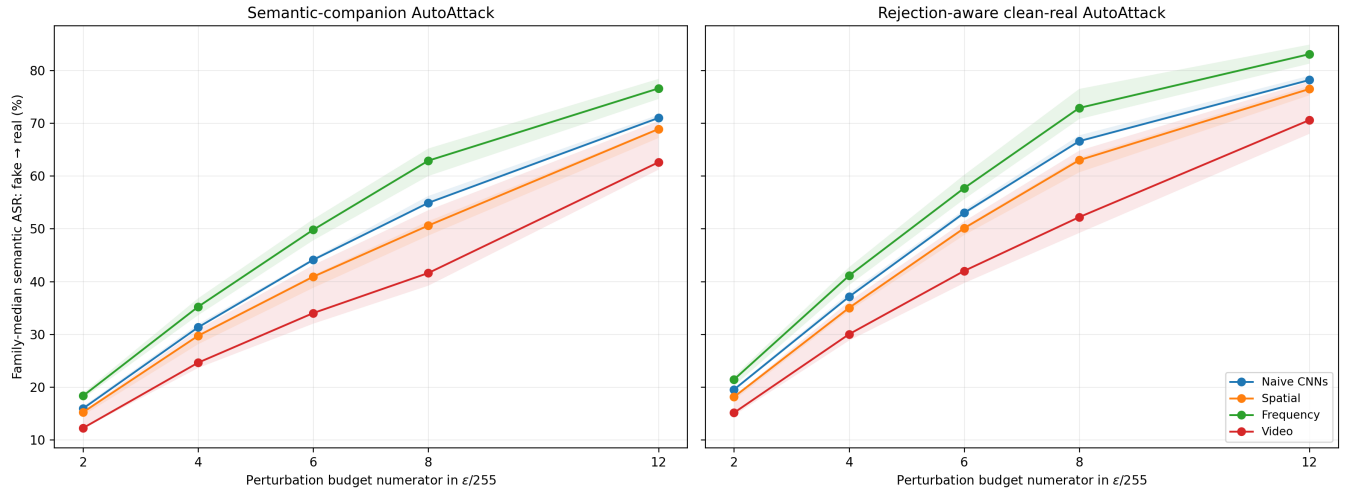


**Figure 9: DBAT adaptive attack sensitivity to perturbation budget.** Curves show family-median fake-to-real semantic ASR as  $\epsilon$  varies from  $2/255$  to  $12/255$  under the main optimizer setting (200 steps, 10 restarts and step size  $2/255$ ). Shaded bands show the interquartile range across models within each family. The rejection-aware clean-real objective consistently produces higher semantic ASR than the semantic-companion objective because it directly targets the deployed harmful outcome while treating adversarial companion outputs as rejection.