LEVERAGING REDUNDANCY IN REPRESENTATIONS FOR ROBUST SYNTHETIC VIDEO DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic-video detectors are ranked on benchmarks whose default tier is raw or lightly compressed footage. However, when a video is streamed by a viewer, it is usually re-encoded by the distribution channel, often at a very different quality. Across ten prior state-of-the-art detectors, we observe that the performance and ranking changes significantly based on the quality of the encoded video. For example, on the quality-controlled SynthForensics benchmark, the existing state-of-the-art detector falls from 93.2 at pristine quality to 44.6 %AUC at medium compression, which is indistinguishable from chance. By analyzing spatial frequency sub-bands, we find that the bands driving detector decisions shift with compression. Relying on specific bands–either by design or as a consequence of training–can therefore leave a detector vulnerable to compression. To reduce this dependence, we randomly suppress frequency bands during training so that no single band is indispensable and combine features from multiple network depths to provide alternative sources of evidence. We also match compression conditions between real and synthetic training samples, preventing compression itself from serving as a shortcut for identifying synthetic videos. Trained on a single generator, our detector holds 90.0 %AUC at CRF40 on the same benchmark, moving from 99.7 on the default tier. Beyond compression, we evaluate robustness to adversarial attacks, which remains understudied in synthetic-video detection. Applying band suppression at inference mitigates most of the accuracy loss caused by a query-limited black-box attack, without impacting pristine accuracy.

1 INTRODUCTION

Real-world deployment of synthetic-video detectors has become a central question for forensic research. The EU AI Act requires providers of generative AI systems to mark synthetic outputs in machine-readable form, such as watermarks or metadata, and to make AI-generated or manipulated content detectable even when it is not deceptive. (European Union, 2024). In practice, however, synthetic video rarely reaches viewers in the form produced by the generator. Social platforms, messaging apps, and broadcast pipelines typically re-encode videos according to their own rate-control ladders. A deployed detector therefore sees compressed copies at varying quality levels, rather than pristine generator outputs. For deployment, what matters is the accuracy that survives this re-encoding across the quality range encountered in the wild.

For example, detectors on social-media platforms may encounter videos whose median compression is equivalent to CRF40 (Fig. 1, top). CRF (Constant Rate Factor) controls encoding quality in x264 and x265, with higher values indicating stronger compression and lower quality. Other platforms have their own

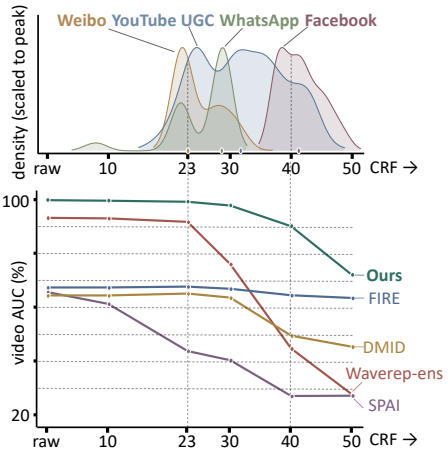


Figure 1: Public video channels deliver content at lower quality than the videos used to rank synthetic-video detectors. *Top:* Peak-normalized distribution of CRF-equivalents for videos range from CRF23 (Weibo) to CRF41 (Facebook). *Bottom:* on quality-controlled SynthForensics, prior detectors degrade as CRF rises, while ours holds 90% AUC at CRF40.

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

quality medians: Weibo CRF23 and Facebook CRF41 (Yang et al., 2020), WhatsApp CRF28 (Shullani et al., 2017), and YouTube User Generated Content CRF32 (Ying et al., 2021). Yet most synthetic-video detection benchmarks have emphasized generator diversity rather than quality coverage, with the recent SynthForensics benchmark (Leotta et al., 2026), an evolution of the quality-controlled FaceForensics++ (Rössler et al., 2019), as a notable exception. When evaluated across this "in-the-wild" quality range, existing detectors that lead on default high-quality videos can fall to chance level, for example at CRF40 (Fig. 1, bottom).

We find that compression shifts the frequency sub-bands of a video frame that carry the evidence used by a detector. As a result, a detector tied to a fixed set of frequency sub-bands, either by architecture or by training, loses accuracy as compression increases. This matters because synthetic-video detection rests on the premise that generation leaves traces in the image, and many of the most separable traces—resampling periodicities (Frank et al., 2020; Durall et al., 2020), upsampling artifacts (Tan et al., 2024b), and spectral peaks (Corvi et al., 2025)—occupy specific regions of the spectrum. A detector that cannot follow this evidence as it moves under re-encoding is left reading the wrong part of the signal.

Our insight is to make the detector rely on redundant forensic evidence across both frequency sub-bands and network depths. We introduce *band dropout*, which randomly suppresses input frequency sub-bands during training so that no single band becomes indispensable. We pair it with a *multi-depth readout* over all transformer blocks, allowing the classifier to use representations learned throughout the network rather than only the final block. Together, these interventions can vary *where* the decision can be made. We also construct the training data so that each real–synthetic pair is compressed in the same way, preventing the model from separating classes by compression alone. Across pairs, compression varies over a wide range, so no single quality level dominates training.

With these design choices, our model, trained on a *single* generator, reaches 90.0 AUC at CRF40 on SynthForensics (DFD) and 90.1 on AIGVDBench, compared with 44.6 and 83.7 for the strongest prior detector. It is the only method that remains above 90% AUC at both source quality and CRF40 on both benchmarks. This result holds cell-by-cell across all twelve AIGVDBench generators and all eight SynthForensics generators. Band dropout also provides a second form of robustness that is often overlooked in the literature: adversarial robustness to query-based attacks. Because randomly band-suppressed views can also be used at inference, the same mechanism raises the cost of a query-based adversarial attack without reducing clean accuracy.

In summary, our contributions are as follows:

- We quantify synthetic-video detector accuracy across the quality range encountered in the wild and show that compression shifts the frequency bands carrying forensic evidence.
- We systematically evaluate the impact of compression by augmenting existing benchmarks with quality ladders, enabling a granular analysis of where current models collapse.
- We propose a compression-robust training approach based on band dropout and multi-depth readout, so that redundancy in the detector's evidence supports robustness.
- We show that our approach is more robust to compression than existing detectors and, by design, also improves robustness to query-based adversarial attacks.

2 Related Work

Synthetic video detection and robustness analysis: Synthetic-video detectors often exploit cues that emerge only when frames are viewed as a video. Some model temporal consistency or long-range dynamics with state-space architectures (Chen et al., 2026; Ma et al., 2024). Others add optical-flow streams to the RGB stream (Bai et al., 2025), search for reconstruction traces (Song et al., 2024), use spatiotemporal or perceptual-geometry scores (Zhang et al., 2025; Internò et al., 2025), or preserve artifacts at native scale (Li et al., 2026; Kundu et al., 2025). Frame-based methods can also use video structure indirectly. For example, Waverep-ens (Corvi et al., 2025) is applied frame by frame, but its training batches are built from video frames. These directions are complementary to ours. Our detector makes a single-frame decision, but it is trained to induce redundancy in its representations and to enforce temporal label consistency across frames. Because our interventions act at training time, they can coexist with video-native architectures, temporal modules, and other detector designs. Robustness, however, is usually tested narrowly. Waverep-ens (Corvi et al., 2025) trains and tests between CRF16 and CRF30, while broader evaluations report sharp degradation under

distribution shift (Chandra et al., 2026; Li et al., 2025). Adversarial robustness remains even less explored (Khan et al., 2025). In Sec. 4, we therefore include a query-based attack evaluation using Square Attack (Andriushchenko et al., 2020).

Synthetic image detectors: Synthetic-image detectors are a closely related source of baselines for synthetic-video detection because they can be applied to videos frame by frame. They do not directly model motion, temporal consistency, or video-level compression behavior, so they may suffer when deployed on re-encoded video. At the same time, they remain essential comparisons: many are strong, widely studied, and designed to generalize across visual sources. One line of work learns transferable CNN features. Wang et al. (2020) showed that a CNN trained on a single generator can transfer across many generators when enough augmentation is used. Gragnaniello et al. (2021) further showed that this transfer comes largely from the augmentation strategy and the removal of early downsampling, rather than from the architecture itself. A second line reads frozen general-purpose representations, either from the final CLIP embedding (Ojha et al., 2023) or from every CLIP block (Koutlis & Papadopoulos, 2024). A third line starts from an explicit forensic signal model. NPR uses pixel correlations left by upsampling (Tan et al., 2024b), FreqNet constrains the classifier to the frequency domain (Tan et al., 2024a), and SPAI and FIRE use spectral and reconstruction residuals (Karageorgiou et al., 2025; Chu et al., 2025). These methods may not capture all video-specific failure modes, but their strength and per-frame formulation make them a vital point of comparison for our setting.

Forensics-related cues in the frequency domain: It is well known that generators can leave traces in the frequency domain. Prior work has found periodic artifacts from transposed convolution, high-frequency deficits, and frequency cues that survive mild post-processing (Frank et al., 2020; Durall et al., 2020; Dzanic et al., 2020; Qian et al., 2020; Jeong et al., 2022). Corvi et al. (2025) extend this view to video by measuring the spectra of real and generated videos directly. This line of work characterizes the signal present in the data. Our question is different: which parts of the spectrum does a trained detector actually use when making its decision? The distinction matters because the strongest signal in the data and the evidence used by the model need not be the same. If a detector concentrates its decision in a narrow set of sub-bands, re-encoding can weaken or move that evidence and leave the model with no alternative source of support. This is the failure mode our method is designed to address.

Evaluation benchmarks: Multi-generator video benchmarks have grown quickly, including GenVideo (Chen et al., 2026), GenVidBench (Ni et al., 2026), AIGVDBench (Ma et al., 2026), and SynthForensics (Leotta et al., 2026). We use the last two in our experiments. Their treatment of compression, however, remains uneven, and method rankings are typically determined at each benchmark's default encoding. The closest published measurement is the SAFE Challenge post-processing track, where the winning method fell from 0.93 to 0.74 AUC (Trapeznikov et al., 2026). Deepfake detection is an important exception, since FaceForensics++ made re-encoding a standard evaluation axis through its compressed tiers (Rössler et al., 2019). But deepfake detection is not the same as synthetic-video detection. It usually detects a manipulation of authentic footage, and cues such as blending boundaries or identity inconsistency do not necessarily exist in video generated end to end. SynthForensics inherits human-centric content, and its detector roster comes largely from this face-forensics family, including UCF, RECCE, ProDet, FTCN, GenConViT, and Effort. A similar boundary separates local manipulation detection, which segments spliced or inpainted image regions (**?**Kang et al., 2025) or partially edited video (Liu et al., 2025). Our setting instead asks for a whole-clip decision on video generated entirely by a model.

Leveraging multi-depth cues from deep networks: Our multi-depth readout builds on an established idea: different layers of a deep network can carry different forensic information. RINE aggregates the class token from every block using a trainable importance estimator, and intermediate layers have been reported to separate real from generated content better than the final output (Koutlis & Papadopoulos, 2024; Huang et al., 2026). Our use of this idea is different. Prior work uses multi-depth cues mainly to improve representation quality. We use them for fault tolerance under re-encoding. Different blocks degrade unevenly as compression increases, so a classifier that reads only the last block inherits that block's failure. Reading across depths gives the classifier multiple routes to the decision.

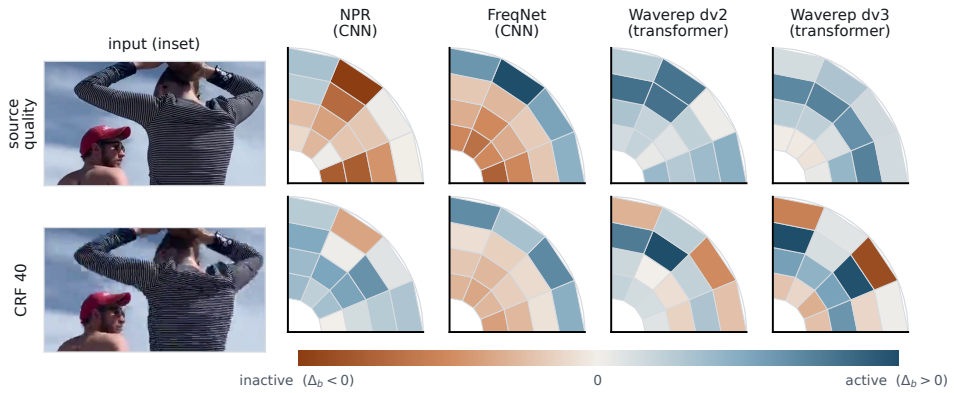


Figure 2: **Frequency sub-bands that contribute to synthetic-video detector decisions shift under compression.** For four prior detectors—two CNNs (NPR (Tan et al., 2024b), FreqNet (Tan et al., 2024a)) and two transformers (DINOv2 and DINOv3 trained under the Waverep recipe (Corvi et al., 2025))—we visualize active sub-bands over one quadrant of the frequency plane, where radius represents scale and angle represents orientation. We compare source quality (top) with CRF40 compression (bottom). Blue marks *active* sub-bands, whose removal hurts accuracy more than a structureless perturbation of the same energy, while red marks inactive sub-bands, whose removal hurts less and often improves accuracy.

3 METHOD

Our method has two parts. First, we locate the evidence current detectors use by measuring which frequency sub-bands affect their decisions. Second, we train a detector whose evidence is deliberately redundant, so that no single frequency band or transformer block becomes indispensable. We first give an intuition for why this redundancy is needed. The implementation section then describes how we enforce it with band dropout, multi-depth readout, and compression-matched training pairs.

3.1 INTUITION: LOCATING DETECTOR EVIDENCE

A video rarely reaches a viewer at the quality at which it was generated. Encoding, rescaling, and re-encoding change which frequency content survives. We therefore ask which parts of the frequency spectrum a detector actually uses for its decision, with and without compression.

Which sub-bands carry the evidence: We decompose a frame into a steerable pyramid (Simoncelli & Freeman, 1995) with $S{=}4$ scales and $K{=}4$ orientations, plus highpass and lowpass residuals, giving $B = SK + 2 = 18$ sub-bands. Let c_b be the coefficients of sub-band b, and let e_b be their energy. We measure a sub-band's impact by perturbing it and recording the change in AUC. The perturbation must preserve e_b, because an energy drop can affect accuracy by itself, independent of the information in that band. We therefore rotate c_b toward a random permutation of itself, preserving energy while disrupting within-band structure, and denote the resulting AUC drop by δ_b. To isolate the sub-band-specific effect, we subtract the AUC drop from applying the same rotation to arbitrary matched-energy coefficients, $\delta^{\text{null}}(e_b)$. This gives $\Delta_b = \delta_b - \delta^{\text{null}}(e_b)$. A positive value, $\Delta_b > 0$, indicates that the sub-band is "active" and carries evidence used by the detector.

Fig. 2 visualizes Δ_b for two CNNs and two transformers at source quality and CRF40. Across models, active sub-bands shift under compression. Thus, a detector that relies on a fixed sub-band profile can fail when compression changes where the useful evidence appears. Training on more compressed video may help for known settings, but it does not scale to the many encoders, rate-control policies, rescalings, and post-processing pipelines seen in deployment. Instead, we control what training can control: whether the detector is allowed to depend on any particular sub-band. If no sub-band is reliably available during training, no single sub-band can become indispensable at test time.

Spectral bias across transformer blocks: Deep networks exhibit spectral bias: under gradient-based training, they preferentially capture low-frequency structure (Rahaman et al., 2019; Xu et al., 2019). In vision transformers, this bias also appears across depth. Self-attention acts as a low-pass filter over the token grid, so successive blocks attenuate high-frequency spatial information in favor of smoother, more global representations (Park & Kim, 2022; Wang et al., 2022). In the limit of pure attention, token representations can collapse toward uniformity (Dong et al., 2021). Layer-wise

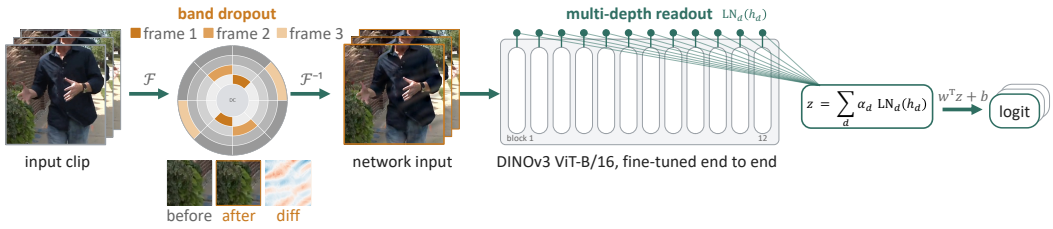


Figure 3: **Model overview.** During training, band dropout randomly suppresses polar Fourier tiles in each frame, forcing the detector to avoid reliance on any single frequency region. A DINOv3 ViT-B/16 backbone is then fine-tuned end to end with a multi-depth readout that learns a weighted mixture of class tokens from all transformer blocks.

analyses likewise find that ViT blocks differ in the locality and granularity of the information they encode (Raghu et al., 2021; Walmer et al., 2023). The final block is therefore not uniformly optimal for forensic evidence. It emphasizes semantic, lower-frequency content, while earlier blocks can preserve finer spatial detail. Aggregating multiple depths lets the detector access both regimes. We use this as our second design choice: the detector logit is a learned mixture over class tokens from all transformer blocks, rather than the last block alone.

3.2 Implementation: Enforcing Redundant Evidence Search

We enforce redundancy in three ways. Band dropout discourages reliance on any one frequency region. Multi-depth readout avoids relying on a single transformer block. Compression-matched training pairs prevent compression itself from becoming a shortcut cue.

Model backbone and multi-depth readout: We use DINOv3 ViT-B/16 (Siméoni et al., 2025) as the detector backbone and fine-tune it end to end. It is self-supervised, accepts any patch-aligned input size, and exposes a class token at each of its $D{=}12$ blocks. These properties let us train at native resolution, avoid resize-induced low-pass filtering, and read evidence from multiple depths. A ViT block emits patch tokens, register tokens, and a class token. Patch tokens remain spatially local, and register tokens (Darcet et al., 2024) absorb high-norm global artifacts that would otherwise disrupt the patch grid. The class token summarizes the full patch grid into one vector and is the natural input to a linear classifier. Because every block has a class token of the same width, the stack of class tokens gives a direct way to learn a mixture over depths. Given the spectral differences across depth, this mixture also gives the classifier access to frequency information retained at different stages.

For class tokens $h_1, \ldots, h_D$, we apply a separate LayerNorm LN_d at each depth because feature scales vary across blocks. We then learn mixture weights $\theta \in \mathbb{R}^D$ over the normalized class tokens:

$$z = \sum_{d=1}^{D} \alpha_d \, \mathrm{LN}_d(h_d), \qquad \alpha = \mathrm{softmax}(\theta), \qquad \hat{y} = w^\top z + b. \tag{1}$$

We initialize θ to zero, so the readout starts as the mean of the layer-normalized class tokens.

Band dropout: Band dropout prevents any single frequency region from becoming indispensable. Like dropout (Srivastava et al., 2014), it makes part of the input unavailable during training so the model must find more than one route to the same decision. Here, the unavailable part is a frequency band rather than a neuron or spatial patch. This is closest in spirit to input-space occlusion methods such as Cutout (DeVries & Taylor, 2017), but applied in the spectral domain.

We take the 2D Fourier transform $X = \mathcal{F}(x)$ of a training frame x and divide the frequency plane into polar tiles C_i. The center disc, C_{DC}, contains the lowest frequencies and is never dropped. Around it, $N_r{=}4$ radial rings and $N_a{=}4$ angular wedges define $N_{\mathrm{dr}}{=}16$ droppable tiles indexed by scale and orientation. At each training step, independently for each frame, we apply band dropout with probability p. When applied, we sample n_{drop} tiles $\mathcal{S}$ uniformly from the 16 droppable tiles, zero them, scale the surviving droppable tiles by gain g, and leave the center disc unchanged:

$$\tilde{X} = 0 \text{ on } \bigcup_{i \in \mathcal{S}} C_i, \quad \tilde{X} = g\,X \text{ on the remaining tiles}, \quad \tilde{X} = X \text{ on } C_{\mathrm{DC}}, \quad g = \frac{N_{\mathrm{dr}}}{N_{\mathrm{dr}} - n_{\mathrm{drop}}}, \tag{2}$$

and $\tilde{x} = \mathcal{F}^{-1}(\tilde{X})$ is given to the network. We use $n_{\mathrm{drop}} = 1$ and $p = 0.5$. We apply this per frame to both the real and synthetic halves of a training pair.

Training: We train the DINOv3 backbone with multi-depth readout at native resolution. For each batch, we sample a random $R \times R$ crop, with $R \in \{480, 560, 640, 720\}$. We use Adam with a constant learning rate of 1.25×10^{-6} and a clip-grouped data sampler. At evaluation, the video logit is the average of frame-level logits.

Real videos come from Panda-70M (Chen et al., 2024), which provides per-video captions and timestamp-level desirability and caption-match scores. We clip the real videos using these annotations and use the provided caption to generate a content-aligned synthetic counterpart. Training examples are real–synthetic pairs, and pair construction matters because systematic content or quality mismatches can leak the label. We pair each real video with a synthetic video generated from its own caption. Following Corvi et al. (2025), we also add "laundered" pairs whose synthetic side is the real video round-tripped through a generator autoencoder, removing content differences and leaving only forensic cues.

Each pair is also quality-matched and compression-augmented. Rather than using source quality alone, we encode and re-encode each pair at randomly drawn CRF values, with two compressed copies per pair. Within a pair, both sides are compressed in the same way, preventing compression from serving as a label shortcut. Across pairs, CRF values vary widely, so no single quality level dominates training. Together, these choices encourage the model to focus on forensic cues rather than semantic or quality differences.

We generate the synthetic corpus with Allegro (Zhou et al., 2024), an open-weight latent video diffusion model with a 3D variational autoencoder and a spatiotemporal diffusion transformer. We choose Allegro because it produces high-quality video and is absent from our evaluation benchmarks. Thus, every benchmark column is zero-shot with respect to the training generator. We select the best model on a held-out validation set of 300 real–synthetic Panda-70M pairs. Training runs for 20 epochs. We provide additional details in App. C.

4 Results

We evaluate whether synthetic-video detectors remain accurate under deployment-time re-encoding, where video quality varies across platforms. We compare prior detectors and our method across augmented quality ladders, then analyze generator-level behavior, design ablations, and query-limited black-box attacks.

4.1 Evaluation Protocol

Benchmarks and quality ladder: We evaluate on two benchmarks using the SynthForensics re-encoding recipe: `libx264`, preset `medium`, `yuv420p`, native resolution, and a single encode from the source rung. **SynthForensics** (Leotta et al., 2026) contains people-centric videos from eight open-source generators in text-to-video and image-to-video modes. We focus on its **DFD** pool, the largest quality-controlled subset with no acquisition mismatch between classes, and call it **SF–DFD**. **AIGVDBench** (Ma et al., 2026) contains general web video from twelve generators. Details for both benchmarks are in App. A. We augment the shipped tiers with CRF10, 30, and 50 for SynthForensics, and CRF40 and 50 for AIGVDBench (Fig. 4). This gives a quality ladder closer to what detectors see in the wild (Fig. 1).

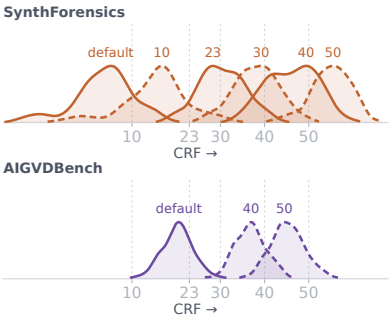


Figure 4: **Benchmarks augmented to cover in-the-wild quality.** Solid contours show shipped benchmark quality; dashed show our added rungs, together covering the range in Fig. 1. "Default" is the highest-quality tier.

Evaluation metrics: We report video-level AUC from the mean per-frame logit, with every column computed on the same video set. We also report balanced accuracy at a calibrated threshold in Tab. 1, using Platt calibration (Platt, 2000) as described in App. C.

4.2 Comparison with Existing Methods

We compare popular and state-of-the-art detectors on the augmented quality ladders for SynthForensics (Leotta et al., 2026) and AIGVDBench (Ma et al., 2026). We include eight image-level detectors applied per frame and averaged over a clip: CNNSpot (Wang et al., 2020), NPR (Tan et al., 2024b),

	SynthForensics (DFD)						AIGVDBench					
	%AUC			%bAcc			%AUC			%bAcc		
Detector	default	CRF40	%Δ	default	CRF40	%Δ	default	CRF40	%Δ	default	CRF40	%Δ
DMID	64.3	49.4	23.2	61.5	52.7	14.3	84.0	52.3	37.8	75.6	51.6	31.7
SPAI	65.6	27.0	58.9	57.7	37.1	35.7	74.8	34.6	53.7	66.8	42.2	36.8
FIRE	67.3	64.4	4.4	62.9	60.8	3.3	51.6	53.4	-3.5	51.0	51.6	-1.2
CNNSpot	69.2	59.3	14.3	63.2	56.5	10.6	53.7	62.5	-16.6	52.0	58.4	-12.3
UnivFD	79.1	66.6	15.8	68.9	59.7	13.4	50.7	52.8	-4.1	50.8	51.6	-1.6
NPR	93.2	34.5	63.0	85.0	49.7	41.5	69.9	59.6	14.7	64.5	56.6	12.2
FreqNet	93.7	49.6	47.0	85.3	52.0	39.0	76.9	77.1	-0.3	70.8	67.2	5.1
RINE	94.1	59.0	37.3	87.9	58.9	33.0	55.9	53.4	4.5	54.6	53.2	2.6
AIGVDet (RGB)	51.0	34.7	32.0	50.1	42.9	14.4	43.6	19.6	55.1	46.0	35.7	22.4
Waverep-ens	93.2	44.6	52.2	85.9	50.2	41.6	**99.6**	83.7	16.0	**98.0**	**81.7**	16.6
Ours	**99.7**	**90.0**	9.7	**97.0**	**78.1**	19.5	91.6	**90.1**	1.7	85.3	76.4	10.4

Table 1: **Compression changes detector rankings.** We compare detectors at default quality, where methods are usually ranked, and at CRF40, a common delivery quality for social platforms. High default accuracy does not reliably transfer to deployment-like re-encoding.

FreqNet (Tan et al., 2024a), UnivFD (Ojha et al., 2023), RINE (Koutlis & Papadopoulos, 2024), DMID (Gragnaniello et al., 2021), FIRE (Chu et al., 2025), and SPAI (Karageorgiou et al., 2025). We also include two video-level detectors, Waverep-ens (Corvi et al., 2025) and AIGVDet (Bai et al., 2025). Because deployment often involves train–test generator mismatch, we evaluate every method zero-shot using public code and checkpoints. Waverep-ens is a state-of-the-art ensemble that includes DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025) backbones.

Overall impact of varying compression levels. Tab. 1 reports %AUC and %balanced accuracy at default quality and CRF40, along with the relative drop %Δ. Fig. 5 extends this comparison across the full ladder and shows per-generator AUC at the same two tiers. Additional breakdowns are in the appendix: per-generator values (App. E), numerical results at every rung (App. E).

The main trend is that default-quality rankings do not predict deployment robustness. A saturated default-tier AUC (≥ 90) does not guarantee high accuracy after re-encoding, so a robust detector must balance peak accuracy with stability across quality levels. On SF–DFD, RINE, FreqNet, NPR, and Waverep-ens are tightly clustered at default quality, ranging from 94.1 to 93.2 %AUC. At CRF40, they fall to 59.0, 49.6, 34.5, and 44.6 %AUC. Our method moves from 99.7 to 90.0 %AUC, remaining the only detector saturated at both tiers. AIGVDBench shows the same tradeoff differently: Waverep-ens leads at default quality with 99.6 %AUC, while ours is lower but still saturated at 91.6. At CRF40, Waverep-ens drops to 83.7, while ours remains at 90.1. Some methods change little or even improve under compression, but their accuracy is near chance or below the saturated range. Thus, the robust choice is not simply the highest default score or the smallest drop, but the method that stays accurate across the ladder.

Fig. 5 a shows where collapse begins. On SF–DFD, baseline curves are mostly flat through CRF10, with a mean loss of 0.7 AUC. They begin moving by CRF23 and lose 21.8 AUC on average between CRF23 and CRF40. The steepest rung is CRF30 to CRF40, where five of the ten baselines lose the most. Our curve shifts this failure point one rung later: it loses only 2.0 points by CRF30, and its steepest segment is CRF40–50. Ranking the ten baselines by SF–DFD AUC at each tier, the Spearman correlation between default quality and CRF40 is only $+0.24$. The four best existing detectors at default quality fall to ranks four, five, nine, and seven at CRF40, while UnivFD rises from fifth to first among them and FIRE from seventh to second. Ours is first at every tier and is excluded from this baseline ranking. AIGVDBench breaks later, with baselines losing 11.2 AUC on average between default quality and CRF40 and collapsing more sharply at CRF50.

Per-generator performance analysis. The dials in Fig. 5 show the same trend without averaging across generators. On SF–DFD at default quality, several methods trace the outer ring together: our margin over the strongest baseline is modest, 99.7 against 94.1, and five detectors exceed 93 %AUC. At CRF40, ours is the only method outside the chance disc on every spoke, with 90.0 against the closest baseline at 66.6. On AIGVDBench, ours and Waverep-ens are both near-saturated at default quality, but only ours remains above 90 %AUC at CRF40. When collapse occurs, it is spread across generators rather than driven by a few outliers. On SF–DFD, Waverep-ens loses more than ten points on all 8 generators from default quality to CRF40, while ours does so on 2 (App. E).

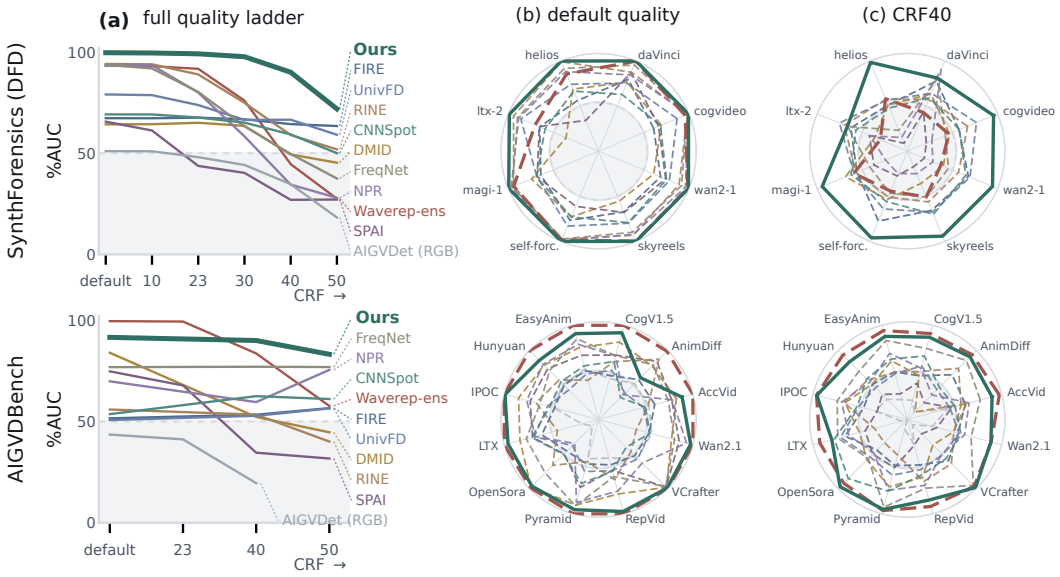


Figure 5: **Performance across quality and generators.** **(a)** %AUC across the full ladder on SynthForensics (DFD) and AIGVDBench. **(b,c)** Per-generator %AUC at default quality and CRF40, with one spoke per generator and the shaded disc below chance. Ours is green, Waverep-ens is brick red, and released baselines are dashed.

4.3 Model Design Analysis

Impact of band dropout and multi-depth readout. We start from a model with neither band dropout nor multi-depth readout, then add each intervention and measure AUC from CRF30 through CRF50. Tab. 2 shows that both choices help, with larger gains as compression increases. At CRF40, band dropout adds $+2.8$ %AUC and multi-depth readout adds $+4.5$. At CRF50, the full model gains $+9.1$ over the model with neither intervention. The components therefore help most where deployment compression makes detectors fail.

Variant	CRF 30	CRF 40	CRF 50
neither	94.5	85.7	62.8
+ band-dropout	97.4 (+2.9)	88.5 (+2.8)	67.2 (+4.5)
+ multi-depth readout	97.9 (+3.4)	90.2 (+4.5)	69.9 (+7.1)
both	97.7 (+3.2)	90.0 (+4.3)	71.9 (+9.1)

Table 2: Impact of each design choice on SF–DFD %AUC across CRF tiers.

Interaction between band dropout and multi-depth readout. We next ask how the two interventions interact. To isolate the effect of band dropout, we compare two models trained with the same multi-depth head, only one of which uses band dropout.

For both models, we measure the evidence at each depth by evaluating %AUC from each block's class token through the trained linear head on SF–DFD (Fig. 6). Band dropout lifts the early half of the blocks from 0.507 to 0.739 AUC, while leaving the late half largely unchanged. This matters because the depth mixture α in Eq. 1 is fixed after training and cannot adapt to a compressed input. When compression weakens the blocks that carried the source-quality decision, the readout benefits from early blocks that retain usable evidence. At CRF40, early blocks retain 0.69 of their discriminative signal with band dropout, compared with 0.40 without it ($+0.29 \pm 0.085$). The same interaction appears along the frequency axis. With multi-depth readout alone, the effective sub-

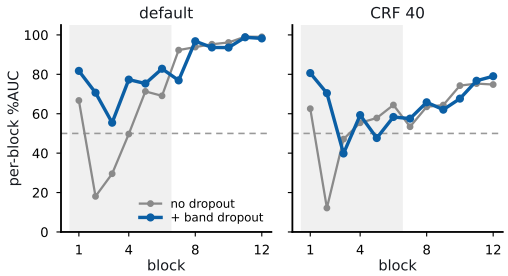


Figure 6: **Band dropout strengthens early-block evidence.** Per-block %AUC is measured by reading each block's class token through the trained head. Band dropout lifts early blocks at both default quality and CRF40, while late-block performance is largely unchanged.

band count contracts under compression, from
8.7 to 7.0. Adding band dropout reverses the direction, increasing the count from 5.6 to 10.1.

Band dropout as a defense against a query-only adversary. In deployment, an API-facing detector may expose a score without exposing its weights. We therefore evaluate a query-limited Square attack (Andriushchenko et al., 2020) and apply random band suppression at inference, using the same energy scaling as Eq. 2 (protocol in App. F). We set the perturbation budget to $\epsilon = 0.5/255$ in L_∞. This is below the $8/255$ common in ImageNet-style robustness studies because large perturbations are unlikely to survive the re-encoding setting studied here. Our sweep is over query budget, not over ϵ, and we do not claim robustness at larger perturbation budgets.

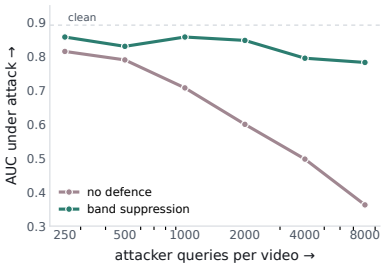


Figure 7: **Query-only robustness against attacker budget.** Under Square attack with $\epsilon = 0.5/255$, inference-time band suppression raises the attack cost rather than eliminating the attack. At 8,000 queries, undefended AUC falls to 0.362, while the defended model remains at 0.782.

Inference-time band suppression only helps when the model has been trained for it. Applied to a model without band-dropout training, it gives little protection: AUC falls from 0.869 to 0.556 under attack, compared with 0.881 to 0.583 without it. For the model trained with band dropout, the same mechanism raises the attack cost. From 250 to 8,000 queries, the undefended detector falls from 0.815 to 0.362, while the defended one moves only from 0.858 to 0.782.

Other design choices: dropout strength, training balance, backbone, and generator. We also test whether the effect depends on a narrow hyperparameter or training source. With the readout fixed, varying $p \in \{0, 0.25, 0.5\}$ raises the effective number of sub-bands at CRF40 from 7.67 to 8.21 and 8.04. Thus, the effect appears at both non-zero settings and is not tied to one precise value of p. Training-pair balance is more critical. When compressed fakes are not paired with quality-matched reals, the model loses 17.1 points from default quality to CRF40, compared with 9.2 for the matched corpus of the same size. Finally, the recipe is not specific to one generator or backbone: App. E.3 reports four single-generator training runs and a DINOv2 ViT-B/14 replication.

5 Discussions

Limitations: Our defense is designed for compression robustness and query-limited black-box attacks, not for unrestricted adversaries. Under an unrestricted white-box attacker, and under transfer from a surrogate model, no variant we trained survives, and band suppression does not change that outcome. Thus, the adversarial claim is intentionally narrow: band-level redundancy raises the cost of a low-amplitude query-only attack, but it is not a general adversarial defense. Our quality ladders approximate common delivery settings, but real platforms may introduce additional perturbations, including resizing, filtering, codec changes, and rate-control policies that are not fully captured here. Our detector is evaluated through frame-level evidence aggregated over video, so it does not exhaust the possible gains from stronger temporal modeling or from combining our training recipe with video-native architectures. Finally, our training uses a controlled set of generators and data sources. Broader deployment will require continued evaluation as generation models, codecs, and distribution channels evolve.

Conclusion: Synthetic-video detectors should be evaluated at the quality levels where they are actually deployed, not only on pristine or lightly compressed benchmark tiers. We show that re-encoding changes both detector accuracy and detector ranking, because compression shifts the frequency evidence that models use for their decisions. This makes robustness a selection criterion, not merely an extra stress test: a detector must remain accurate while its input quality changes. Our method addresses this failure mode by making the decision redundant across frequency sub-bands and transformer depths, and by preventing compression itself from becoming a shortcut during training. The result is a detector that remains accurate across deployment-like quality ladders while preserving strong source-quality performance. The broader lesson is that robustness for synthetic-video detection is not a separate add-on, but part of what makes a detector deployable.

Reproducibility Statement

Sec. 3 specifies the model architecture, input resolutions, optimizer, learning rates, weight decay, batch size, and loss. Sec. 4 defines the evaluation protocol, including benchmark tiers, re-encoding ladders, and per-video aggregation, while App. C details the training corpus, optimization, and checkpoint selection. Sec. 4 specifies the sub-band decomposition, the intervention, the matched-perturbation comparison, the null baseline, and the transform reconstruction error. It also specifies the per-block decomposition and the identity that makes it exact. Published baselines are quoted from their sources where stated, and otherwise evaluated from released checkpoints under the protocol in Sec. 4.

Ethics Statement

This work studies synthetic-video detection, motivated in part by disclosure obligations under Article 50 of the EU AI Act. Detection research is dual-use: the same attribution method that reveals which sub-bands a detector uses can also indicate which sub-bands an adversary might target, and Sec. 4 reports attacks that succeed. We judge publication to be net beneficial because these attacks are standard and already available, while the failure mode we document is less widely recognized: detectors can appear accurate at benchmark encoding yet fall toward chance after ordinary re-encoding. That failure directly affects whether deployed detection systems work. We use only established public benchmarks and released checkpoints, and we quote benchmark authors' reported numbers for detectors whose preprocessing we cannot faithfully reproduce.

Use of AI Tools

An LLM-based coding assistant was used for software-engineering support, including implementation and debugging of evaluation and attribution code, script refactoring, and figure and table preparation. It was also used for editorial assistance on the manuscript. We used the open-weight video generation model Allegro to generate the synthetic training corpus described in Sec. 3.2 and App. C. All experimental design, hypotheses, analysis decisions, and conclusions are the authors'. All reported numbers were produced by the authors' code on the authors' compute and verified by the authors against source artifacts.

References

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 (ECCV)*, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018.

Jianfa Bai, Man Lin, Gang Cao, and Zijie Lou. Ai-generated video detection via spatial-temporal anomaly learning. In *Pattern Recognition and Computer Vision (PRCV 2024)*, volume 15040 of *Lecture Notes in Computer Science*, pp. 460–470. Springer, 2025.

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 (CVPR) Workshops*, pp. 10668–10678, 2026.

Haoxing Chen, Yan Hong, Zizheng Huang, Zhuoer Xu, Zhangxuan Gu, Yaohui Li, Jun Lan, Huijia Zhu, Jianfu Zhang, Weiqiang Wang, and Huaxiong Li. DeMamba: Ai-generated video detection on million-scale GenVideo benchmark. *Science China Information Sciences*, 69(6):162103, 2026.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. FIRE: Robust detection of diffusion-generated images via frequency-guided reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12830–12839, 2025.

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

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Terrance DeVries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. In *International Conference on Machine Learning (ICML)*, 2021.

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 (CVPR)*, pp. 7890–7899, 2020.

Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

European Union. Regulation (EU) 2024/1689 (artificial intelligence act), article 50: Transparency obligations for providers and deployers of certain AI systems, 2024. URL `https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng`. Article 50 generally applies from 2 August 2026.

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*, pp. 3247–3258, 2020.

Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva. Are GAN generated images easy to detect? a critical analysis of the state-of-the-art. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2021.

Zhenhan Huang, Pin-Yu Chen, Tejaswini Pedapati, and Jianxi Gao. Intermediate representations are strong ai-generated image detectors. *arXiv preprint arXiv:2605.04358*, 2026.

Christian Internò, Robert Geirhos, Markus Olhofer, Sunny Liu, Barbara Hammer, and David Klindt. Ai-generated video detection via perceptual straightening. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-0697.

Yonghyun Jeong, Doyeon Kim, Seungjai Min, Seongho Joe, Youngjune Gwon, and Jongwon Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 48–57, 2022.

Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. Legion: Learning to ground and explain for synthetic image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18937–18947, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Naseem Khan, Tuan Nguyen, Amine Bermak, and Issa Khalil. Unmasking synthetic realities in generative ai: A comprehensive review of adversarially robust deepfake detection systems. *arXiv preprint arXiv:2507.21157*, 2025.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 394–411, 2024.

Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K. Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28050–28060, 2025.

Roberto Leotta, Salvatore Alfio Sambataro, Claudio Vittorio Ragaglia, Mirko Casu, Yuri Petralia, Francesco Guarnera, Luca Guarnera, and Sebastiano Battiato. SynthForensics: Benchmarking and evaluating people-centric synthetic video deepfakes. *arXiv preprint arXiv:2602.04939*, 2026.

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

Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, volume 38, 2025. doi: 10.52202/085713-5739.

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

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

Long Ma, Zihao Xue, Yan Wang, Zhiyuan Yan, Jin Xu, Xiaorui Jiang, Haiyang Yu, Yong Liao, and Zhen Bi. Your one-stop solution for AI-generated video detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4458–4470, 2026.

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

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Namuk Park and Songkuk Kim. How do vision transformers work? In *International Conference on Learning Representations (ICLR)*, 2022.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000.

Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred A. Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning (ICML)*, 2019.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Dasara Shullani, Marco Fontani, Massimo Iuliani, Omar Al Shaya, and Alessandro Piva. VISION: a video and image dataset for source identification. *EURASIP Journal on Information Security*, 2017(1):15, 2017. doi: 10.1186/s13635-017-0067-2.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Eero P Simoncelli and William T Freeman. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *ICIP*, 1995.

Xiufeng Song, Xiao Guo, Jiache Zhang, Qirui Li, Lei Bai, Xiaoming Liu, Guangtao Zhai, and Xiaohong Liu. On learning multi-modal forgery representation for diffusion generated video detection. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

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

Chuangchuang Tan et al. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *CVPR*, 2024b.

Kirill Trapeznikov, Gabriel Mancino-Ball, Jonathan Li, Paul Cummer, Jai Aslam, Danial Samadi Vahdati, Tai Nguyen, Matthew C. Stamm, Peter Bautista, Michael Davinroy, Laura Cassani, and Jill Crisman. Advancing reliable synthetic video detection: Insights from the SAFE challenge. *Forensic Science International: Digital Investigation*, 57:302120, 2026. doi: 10.1016/j.fsidi.2026.302120.

Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Peihao Wang, Wenqing Zheng, Tianlong Chen, and Zhangyang Wang. Anti-oversmoothing in deep vision transformers via the fourier domain analysis: From theory to practice. In *International Conference on Learning Representations*, 2022.

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

Zhi-Qin John Xu, Yaoyu Zhang, and Yanyang Xiao. Training behavior of deep neural network in frequency domain. In *Neural Information Processing (ICONIP 2019)*, volume 11953 of *Lecture Notes in Computer Science*, pp. 264–274. Springer, 2019.

Pengpeng Yang, Daniele Baracchi, Massimo Iuliani, Dasara Shullani, Rongrong Ni, Yao Zhao, and Alessandro Piva. Efficient video integrity analysis through container characterization. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):947–954, 2020. doi: 10.1109/JSTSP.2020.3008088.

Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-VQ: 'patching up' the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14019–14029, 2021. doi: 10.1109/CVPR46437.2021.01380.

Shuhai Zhang, ZiHao Lian, Jiahao Yang, Daiyuan Li, Guoxuan Pang, Feng Liu, Bo Han, Shutao Li, and Mingkui Tan. Physics-driven spatiotemporal modeling for ai-generated video detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025. doi: 10.52202/085713-5810.

Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model. *arXiv preprint arXiv:2410.15458*, 2024.

A Evaluation Benchmarks and the Quality Ladder

A.1 Benchmarks

SynthForensics. SynthForensics (Leotta et al., 2026) is people-centric video. Its synthetic split contains eight open-weight generators in text-to-video and image-to-video modes: CogVideoX, Wan2.1, SkyReels-V2, MAGI-1, Self-Forcing, LTX-2, daVinci-MagiHuman, and helios. Of these, seven of eight are 2025 releases, the oldest is August 2024, and two could not be matched to a public release date. We evaluate against the 363 DeepFakeDetection reals, the pool whose two classes carry no acquisition mismatch, and refer to this setting throughout as SF–DFD.

AIGVDBench. AIGVDBench (Ma et al., 2026) is general web video over twelve generators. Its real pool is not the same size at every rung: the source tier contains $\approx 20{,}000$ reals, while each compressed rung contains $\approx 3{,}001$. Sampling reals independently per tier would leave only 74 of 500 shared videos between the source and compressed pools, so source-to-CRF deltas would mix compression effects with a change in real videos. We instead fix one video set per column and score every detector on that set, so each delta reflects quality alone.

A.2 The Re-encoding Ladder

Both benchmarks are extended with the SynthForensics re-encoding recipe: `libx264`, preset `medium`, pixel format `yuv420p`, native resolution, and one encode from the source rung. Thus, each rung measures one encode at a chosen strength, not an accumulated chain of encodes. SynthForensics ships source quality, CRF 23, and CRF 40. We add CRF 10, 30, and 50, giving six rungs. AIGVDBench ships source quality and CRF 23. We add CRF 40 and 50.

A.3 What "Source Quality" Means

Rate as a proxy for quality. A nominal tier label describes the encode applied by a benchmark, not necessarily the quality of the delivered file. To put benchmarks, platforms, and our added rungs on one axis, we convert measured bit rate to an equivalent CRF. For a fixed encoder and content, bit rate falls approximately geometrically with CRF, $R(\mathrm{CRF}) \propto 2^{-\mathrm{CRF}/k}$. We fit k on the content-matched pair available to us, namely the same SynthForensics real clips at CRF 23 and CRF 40. This gives $k = 5.93$ CRF steps per halving, which we invert on measured bits per pixel:

$$\widehat{\mathrm{CRF}} \;=\; 23 \;-\; 5.93\,\log_2\!\big(\text{bits per pixel} \,/\, 0.043\big), \tag{3}$$

anchored on the measured rate of the CRF 23 rung. This estimate has a known bias: synthetic video is smoother and encodes to fewer bits at equal CRF, so synthetic clips can appear more compressed than they are.

Shipped benchmark tiers. On SynthForensics, the three shipped tiers of the real pool measure an effective $-11.2 \to 24.1 \to 41.3$. The sequence is monotone, spans a wide range, and lands on the nominal labels at the compressed rungs. The source tier is itself encoded, and the two classes differ there by 27–215$\times$ in rate: reals are effectively lossless, while synthetic clips arrive at about CRF 22. Encoder quality is therefore correlated with the label before any detector is trained, so a detector can partly succeed at that tier by reading quality. Re-encoding both classes to a common target removes this correlation, which is why we treat CRF 23 and CRF 40 as the informative operating points and report the source tier alongside them. One asymmetry remains at matched CRF, and we flag it rather than claim to remove it: reals have been encoded once, while synthetic clips have been encoded twice, so double-compression traces remain a

B Baseline Protocol

B.1 How the Released Detectors Are Run

Every baseline in Tab. 1 is evaluated zero-shot from its released checkpoint, using its repository's preprocessing with only the data root changed. Image-level detectors are applied per frame and averaged over the clip, matching our per-video aggregation. Every number in a column is computed on the same video set, so columns compare detectors and rows compare quality tiers. One implementation

issue is recorded explicitly: detectors whose published preprocessing includes an antialiased resize were at one point scored without it, worth $+7.7$ pooled AUC on FreqNet. The affected cells are regenerated before any number is quoted.

B.2 Competitors Retrained on Our Generator

Tab. 1 reports each competitor *as released*, which mixes the detector recipe with the corpus it was trained on. Following Corvi et al. (2025), we therefore retrain the leader of each family—FreqNet among image-level detectors and Waverep among video-level detectors—on a single generator and evaluate them on the same quality ladder. This makes the detector recipe the variable.

Training data. We use the source-quality Allegro split: $88{,}640$ real–synthetic pairs from our single-generator corpus with no re-encoding rung. Frames are exported at 224 px, four per video, in the `0_real`/`1_fake` layout expected by the released repositories. Allegro appears in neither evaluation benchmark, so the retrained rows remain zero-shot. We deliberately do not train competitors on the compression-matched rungs from Sec. 3.2. That pairing is part of our intervention, and using it here would credit the baseline recipe with our data design. Each method is trained with the command documented by its repository, changing only the data root and class list.

Detector	weights	default	CRF 23	CRF 30	CRF 40	CRF 50
FreqNet	released	93.7	80.3	63.3	49.6	37.5
FreqNet*	retrained	60.0	60.5	61.3	65.1	74.2
Waverep-ens	released	93.2	91.7	76.0	44.6	27.5
Waverep-ens	retrained	79.4	84.7	78.0	55.6	40.6
dv2 half	retrained	73.2	81.5	75.5	48.2	36.0
dv3 half	retrained	83.8	85.6	77.9	65.7	52.0
Ours	—	**99.7**	**99.1**	**97.7**	**90.0**	**71.9**

Table 3: **Leading image-level and video-level competitors, as released and retrained on source-quality Allegro.** Results are SF–DFD video-level AUC.

Retraining flattens both curves but costs source-quality accuracy. FreqNet gives up 33.7 points at source quality and gains 36.7 at CRF 50. The retrained Waverep ensemble gives up 13.8 points at source quality and gains 13.1 at CRF 50. This is the tradeoff described in Sec. 3.2 from the opposite direction: a single-generator corpus without compression rungs can teach cues that survive re-encoding, but at source-tier accuracy well below what the same architectures reach on their original corpora.

C Training, Checkpoint Selection and Calibration

C.1 Training Data

Every arm trains on a paired corpus along the codec diagonal. A real and a synthetic clip are drawn together at the same encoding-history rung over three rungs: uncompressed, one H.264 encode, and one H.264 re-encode. The CRF is drawn independently for the two sides, so a pair is matched in encoding history but not encoding strength, and neither identifies the label. The synthetic half comes from a single generator, Allegro (Zhou et al., 2024), with $253{,}824$ pairs. Reals come from Panda-70M (Chen et al., 2024). Each synthetic clip is generated from its real clip's caption. A subset of pairs is also "laundered" (Corvi et al., 2025): the synthetic side is the real video round-tripped through a generator autoencoder, removing the content difference and leaving only forensic cues.

C.2 Resolution and Optimization

Training uses native resolution with no resizing. We take a 720-px square crop, then a per-batch collate samples a random $R \times R$ native sub-crop, with $R \in \{480, 560, 640, 720\}$. This exposes the model to multiple patch grids and makes evaluation resolution-agnostic. For DINOv2, the ladder is $\{476, 560, 644, 714\}$ because crop sizes must divide the patch stride. That backbone also uses dynamic image size with antialiasing *off* in positional-embedding interpolation, which must match between training and evaluation or the embeddings shift. We use Adam ($\beta = (0.9, 0.999)$) with constant learning rate 1.25×10^{-6} for DINOv3 and 10^{-6} for DINOv2, and weight decay 10^{-6}. There

are two parameter groups: all backbone and head parameters in the first, and the depth mixture θ from Eq. 1 alone in the second, trained at 10^{-3} with weight decay 0. Batches contain 4 real–synthetic pairs per GPU on 32 GPUs, giving 128 pairs and 256 frames per step at 8 frames per clip. We checkpoint every 1,000 steps.

C.3 Intra-clip Consistency

Frames from one clip share a label, so a clip-grouped sampler places $K{=}8$ frames of the same clip in each batch. We add the mean within-clip logit variance:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + \lambda\, \mathbb{E}_c\big[\mathrm{Var}_{i \in c}\, \hat{y}_i\big], \qquad \lambda = 0.1. \tag{4}$$

This term is used only during training. It is present in every arm, including controls, so it is not a confound and is not one of the interventions evaluated in the paper.

C.4 Checkpoint Selection

Every arm is scored every 1,000 steps on two selection tiers of a held-out validation set of 300 real–synthetic Panda-70M pairs disjoint from training. The reported checkpoint is the argmax of the mean score across the two tiers. Selection never uses an evaluation benchmark. Every optimum is interior to the scored range, so the selected peak is bracketed rather than truncated by the training stop. The two tiers disagree about the best checkpoint for every arm, with source quality peaking consistently later. Selecting on source quality alone would move every arm later and changes our own CRF 40 number by under a point, so the comparison is not an artifact of the selection rule.

D Sub-band Attribution Details

For completeness, we give the exact sub-band perturbation used in Sec. 3.1. Let Ψ_b denote analysis of band b in the steerable pyramid (Simoncelli & Freeman, 1995), with coefficients $c_b = \Psi_b x$ and energy $e_b = \|c_b\|_2^2$. For a random permutation π and angle θ, we perturb the band by

$$R_\theta(c_b) = \cos\theta\, c_b + \sin\theta\, \pi(c_b), \qquad x^{(b,\theta)} = x - \Psi_b^\top c_b + \Psi_b^\top R_\theta(c_b). \tag{5}$$

This preserves $\|c_b\|_2$ while disrupting within-band structure. We define the sub-band excess as

$$\Delta_b = \delta_b - \delta^{\mathrm{null}}(e_b), \tag{6}$$

where δ_b is the AUC drop from perturbing band b, and $\delta^{\mathrm{null}}(e_b)$ is the drop from applying the same rotation to arbitrary matched-energy coefficients. To summarize how concentrated the positive evidence is, we compute

$$p_b = \frac{\max(\Delta_b, 0)}{\sum_{b'} \max(\Delta_{b'}, 0)}, \qquad \mathrm{ESC} = \exp\Big(- \sum\nolimits_b p_b \log p_b \Big). \tag{7}$$

This effective sub-band count equals the number of equally used bands when evidence is spread uniformly, and approaches 1 when one band dominates.

E Additional Results Across Quality and Generators

This appendix gives the numerical values behind the quality-ladder and per-generator summaries in Fig. 5. Tab. 4 reports aggregate SF–DFD performance at every CRF rung. Tab. 5 reports the per-generator comparison between our method and Waverep-ens, the strongest released baseline at default quality. We omit the older per-cell figure because Fig. 5 already shows the generator-level trend, and the table below gives the exact values.

E.1 Every SynthForensics Rung

Tab. 1 reports two of the six SynthForensics rungs. Tab. 4 gives all six, with Pd@5% alongside the two metrics used in the main text. Rows are ordered ours, then video-level, then image-level. * marks a detector reported at its best-case sign, fixed once per benchmark at the source tier and applied unchanged at every lower-quality rung. CRF 10 is omitted from the main table because no detector moves more than 0.5 AUC between it and the source rung; CRF 23 is the first rung where the ranking begins to change.

	%AUC by CRF						%bAcc by CRF				Pd@5% by CRF			
Detector	default	10	23	30	40	50	default	23	40	50	default	23	40	50
Ours	99.7	99.5	99.1	97.7	90.0	71.9	97.0	95.2	78.1	60.9	98.4	96.0	68.6	33.1
Waverep-ens	93.2	93.0	91.7	76.0	44.6	27.5	86.0	77.5	50.2	50.0	78.8	63.5	3.8	2.9
FIRE	67.3	67.3	67.6	66.8	64.4	63.4	62.9	64.3	60.8	60.9	19.9	20.5	23.5	28.7
DMID	64.3	64.3	65.0	63.5	49.4	45.3	61.5	62.1	52.7	50.3	0.6	0.8	7.7	4.1
SPAI	65.6	61.2	43.7	40.3	27.0	27.1	57.7	44.4	37.1	32.4	9.9	8.1	3.6	1.4
RINE*	94.1	93.9	88.9	74.9	59.0	51.9	87.9	78.7	58.9	52.1	79.4	60.3	10.3	9.8
UnivFD*	79.1	78.7	73.8	66.5	66.6	59.1	68.9	62.7	59.7	54.4	42.7	35.0	28.3	24.2
NPR*	93.2	93.5	80.0	58.3	34.5	28.0	85.0	59.5	49.7	50.0	80.0	58.0	7.8	3.4
FreqNet*	93.7	91.9	80.3	63.3	49.6	37.5	85.3	70.0	52.0	50.6	67.5	16.0	4.1	8.1
CNNSpot*	69.2	69.2	67.6	65.1	59.3	50.0	63.2	61.3	56.5	50.1	11.1	10.5	6.2	3.9

Table 4: **SF–DFD performance across the full quality ladder.** AUC is reported at six tiers, while balanced accuracy and Pd@5% are reported at the four tiers where those metrics were computed. Pd@5% is the detection rate at 5% false-positive rate, making it the strictest metric because it depends on the tail of the real distribution rather than only on ranking.

E.2 Per-generator Behavior

		default		CRF 23		CRF 40	
Generator	rel.	Ours	W-ens†	Ours	W-ens†	Ours	W-ens†
CogVideoX	2024	99.3	96.8	98.9	98.7	91.0	55.2
Wan2.1	2025	99.7	99.8	97.7	95.1	89.2	40.6
SkyReels-V2	2025	99.9	99.9	99.0	94.1	92.1	43.4
MAGI-1	2025	98.4	98.9	98.8	97.5	93.3	67.6
Self-Forcing	2025	99.9	99.9	99.7	99.6	92.5	56.1
LTX-2	2025	99.2	87.7	98.7	85.5	87.9	48.5
daVinci-MagiHuman	–	100.0	100.0	99.9	91.6	97.9	44.4
helios	–	99.9	77.0	99.9	88.4	93.4	44.9

Table 5: **Per-generator SF–DFD AUC for ours and Waverep-ens.** At default quality, both methods are near-saturated on most generators. At CRF 40, Waverep-ens drops sharply across all eight generators, while our method remains high on every generator. This confirms that the aggregate CRF 40 gap is not driven by a small number of outliers. †Waverep-ens cells are provisional pending a re-score on the full pool.

Table 5 gives the exact values behind the per-generator discussion in Fig. 5. Our spread remains narrow at every rung, 98.4–100.0 at source quality and 87.9–97.9 at CRF 40, so the pooled result is representative rather than driven by a few strong cells. By contrast, Waverep-ens spans 40.6–67.6 at CRF 40 and loses more than ten points on every generator.

E.3 Generator and Backbone Checks

We include two checks that the recipe is not tied to one training generator or one backbone.

Alternate backbone. We transplant the recipe onto DINOv2 ViT-B/14, a second self-supervised backbone with different pretraining, patch size, and positional-embedding convention. The DINOv2 arm uses dynamic image size and crops $\{476, 560, 644, 714\}$, while DINOv3 ViT-B/16 uses $\{480, 560, 640, 720\}$. The ladders differ because crop sizes must divide the patch stride.

Backbone	band dropout	default	CRF 23	CRF 40
DINOv2 ViT-B/14	no	96.7	98.3	93.6
DINOv3 ViT-B/16 (ours)	yes	**99.7**	**99.1**	90.0

Table 6: **Recipe on a second backbone.** SF–DFD video-level AUC is reported at the matched step 16,000. The DINOv2 arm uses multi-depth readout without band dropout, so this is a transfer check rather than a controlled backbone comparison.

The recipe transfers: the DINOv2 arm holds 93.6 AUC at CRF 40, where the released detectors in Tab. 1 are near chance. The reading is bounded by two caveats. First, the DINOv2 ladder is not

monotone across training: over 20 scored steps, CRF 40 ranges from 88.3 to 95.0 with no trend, so we report the matched step rather than an argmax. Second, this arm does not include band dropout, so it is not a like-for-like comparison with the detector we ship. The attribution pattern nevertheless agrees with the DINOv3 arms at rank $+0.78$ to $+0.83$, its highpass contribution moves $+0.039$ to -0.001 under CRF 40, and it retains $15/17$ contributing bands.

Choice of training generator. We also vary only the generator used for the synthetic half of training, keeping the architecture, readout, dropout setting, and pairing rule fixed across four single-generator corpora.

	SF–DFD		AIGVD proxy
Training generator	default	CRF 40	default
Allegro	97.1	91.8	88.9
Cosmos-Predict1	97.6	93.1	71.8
Wan2.1	92.4	93.4	82.2
Pyramid-Flow	82.3	90.3	99.1

Table 7: **One recipe, four training generators.** All four arms share the plain recipe and are selected at their best scored step. The detector reported in the main paper is the full recipe on Allegro and is not one of these rows. The AIGVD column is a five-generator proxy subset, not the full benchmark.

All four training corpora land between 90.3 and 93.4 at CRF 40 on SF–DFD, a 3.1-point spread compared with a 15-point spread at source quality. Thus, compression behavior follows the recipe more than the particular training generator, while source-tier accuracy is more corpus-sensitive. No generator is best on both benchmarks: Pyramid-Flow is strongest on the AIGVD proxy and weakest on SF–DFD, while Allegro and Cosmos-Predict1 show the reverse pattern.

F Adversarial Evaluation: Protocol and Failure Cases

F.1 Threat Model

We evaluate the adversary most relevant to a deployed detector. A detector behind an API exposes a score, not its weights, so the attacker is *query-only*: it can submit a video and read the returned score. We use Square Attack (Andriushchenko et al., 2020), a score-based black-box attack, with perturbations bounded in ℓ_∞ at ϵ =0.5/255 and a sweep over query budget. This is not chosen because it is the attack we survive; Sec. F.4 reports the white-box case precisely because we do not.

Why this ϵ. 0.5/255 in ℓ_∞ is an order of magnitude below the $8/255$ common in ImageNet robustness work. The reason is central to this paper: perturbations large enough to matter at $8/255$ are unlikely to survive delivery-time re-encoding. An attacker working on delivered video therefore operates at low amplitude, or not at all. We report this single budget and make no claim at larger ones. Our sweep is over the attacker's query count, not over ϵ.

F.2 The Defense Is the Training-time Mechanism Applied at Inference

Band dropout defines a distribution over inputs, so it can also be applied at test time. We average the detector score over several randomly band-suppressed views of the same frame, using the same energy-preserving gain as Eq. 2. This makes the detector stochastic, so any gain could in principle come from a noisier attack objective rather than real robustness. We test this in two ways.

Against a randomized defense of matched cost. At 2,000 queries, inference-time band suppression leaves 0.846 adversarial AUC against 0.893 clean. The undefended detector falls from 0.889 to 0.619, while a randomized-resolution defense of matched cost falls from 0.863 to 0.671. Band suppression absorbs more of the attack's effect without a clean-accuracy penalty, and the matched randomized alternative does not reproduce the gain. Thus, the benefit is not simply that the detector is stochastic. Because the defenses are stochastic, each clean number is its own condition's baseline, so these are within-condition drops rather than drops from a shared clean reference.

Against the mechanism being absent from training. Applying band dropout only at inference to a detector not trained with it gives little protection: $0.869 \rightarrow 0.556$ under attack, compared with $0.881 \rightarrow 0.583$ with no defense. The inference-time defense therefore depends on the training-time intervention, rather than merely being compatible with it. This is the comparison Fig. 7 reports across the query sweep.

F.3 Where the Attack Concentrates

The attack does not spread its budget evenly over the spectrum. 47% of the ℓ_∞ attack energy lands in the highpass residual, and in the finest oriented bands its share exceeds those bands' share of image energy by 15–$18\times$. This is the same region of the decomposition that detectors rely on and that compression damages first. That overlap helps explain why the same band-level mechanism addresses both compression and query-only attacks.

F.4 What Does Not Hold

With full gradient access, ℓ_∞ PGD with Expectation over Transformation (Athalye et al., 2018) drives every arm we tested to 0.000 AUC at every $\epsilon \geq 1/255$, with or without band-dropout training. No training-time intervention we tried changes this. Perturbations also transfer between checkpoints of the same architecture at efficiency 0.99–1.02. An attacker with a same-architecture surrogate is therefore not meaningfully slowed by lacking the exact weights. The scope of our claim is accordingly narrow: band-level redundancy raises the cost of a query-only adversary at a low perturbation budget, but does not provide white-box robustness or resistance to surrogate transfer.