FREQUENCY ATTRIBUTION MASK: TOWARD DETECTOR-ADAPTIVE ROBUSTNESS IN AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-generated image (AIGI) detectors report high accuracy on benchmark splits, yet collapse on unseen generators and after routine post-processing such as JPEG compression or resizing, exposing a reliance on shortcut cues that do not transfer. Existing remedies name a specific source-level bias and patch it with a deterministic operation at the data, architecture, or weight level, leaving every other bias untouched and tying the fix to a single detector setup. To address this, we propose *Frequency Attribution Mask* (FAM), an attribution-guided mask-tuning framework that intervenes at each detector's own input sensitivity, the point at which any such bias actually manifests. FAM derives a per-sample frequency-domain attribution from the detector's input gradient, masks the highest-attribution components of the input spectrum with class-asymmetric thresholds, and finetunes on the masked reconstructions. We apply FAM to two pretrained ResNet-50 detectors, one of which already carries a pixel-level alignment fix, and show that it consistently improves post-processing robustness over both base detectors and exceeds recent shortcut-mitigation baselines in balanced accuracy, with the largest gains where residual shortcut reliance is most severe.

1 INTRODUCTION

Recent advances in diffusion- and flow-based generative models (Ho et al., 2020; Rombach et al., 2022; Lipman et al., 2023) have produced near-photorealistic AI-generated images, raising societal concerns including misinformation, non-consensual deepfakes, and copyright infringement. Reliable AI-generated image (AIGI) detection has therefore become a critical research priority, and a diverse set of detectors (Wang et al., 2020; Corvi et al., 2023; Rajan et al., 2025; Ojha et al., 2023; Chen et al., 2024; Baraldi et al., 2024) now report impressive accuracy on benchmark datasets. However, when test images come from generators not seen during training, or after common post-processing operations such as JPEG compression and resizing, detector performance degrades sharply (Li et al., 2025b), exposing a reliance on shortcut cues (Geirhos et al., 2020) that correlate with one class on the training distribution but do not generalize.

Robustness research in AIGI detection has converged on a particular paradigm: each method names a specific bias, such as image format or size (Grommelt et al., 2024; Cazenavette et al., 2024), real-fake distribution mismatch (Yu et al., 2024; Chen et al., 2024; Rajan et al., 2025), real-class feature reliance (Rajan & Lee, 2025), or semantic shortcut formation (Gye et al., 2025; Zheng et al., 2024), and designs a deterministic intervention that targets that specific bias at the data, architecture, or weight level. The implicit premise is that, to make a detector robust, one must first identify which bias has shaped it. This premise is restrictive in two ways. First, each prescription operates only within its own enumerated taxonomy, leaving residual or newly introduced biases untouched. Second, the same source-level bias manifests differently inside detectors with different training data or finetuning schemes, so a deterministic prescription tuned for one detector setup does not transfer cleanly to another. Even detectors with a source-level bias-fix (Chen et al., 2024; Rajan et al., 2025) still degrade sharply under post-processing on unseen generators (Section 4), so no single source-level fix suffices.

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

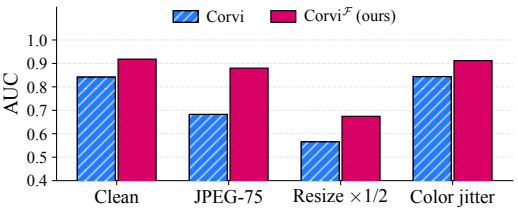


Figure 1: **Robustness of Corvi and Corvi$^{\mathcal{F}}$ on the unseen `FLUX` generator.** AUC on `FLUX` fakes against `Redcaps` reals at 256×256 resolution under four input conditions: clean inputs, JPEG-75 compression, 0.5× downscale-and-resize, and BCS color jitter at ±0.2.

Whatever the source of the bias, its effect on a trained detector ultimately manifests as the detector's *current* input sensitivity, and we argue that this manifestation is most directly diagnosed in the frequency domain. First, generative artifacts consistently concentrate in identifiable frequency bands across generator families (Tan et al., 2024b; Wang et al., 2020; Rajan et al., 2025). Second, the post-processing perturbations that destroy detection accuracy in practice, such as JPEG compression (Li et al., 2025a) as well as downsizing and low-pass filtering (Rajan & Lee, 2025), operate in the same domain, so the axis of detector sensitivity and the axis of deployment perturbation overlap. Third, a single frequency band is a global signal affecting every spatial location at once, placing such cues out of reach of pixel-wise spatial saliency. Building on these observations, we propose *Frequency Attribution Mask* (FAM), a detector-adaptive mask-tuning framework that uses each detector's own gradient as a per-sample signal of where the detector is currently most sensitive in the frequency domain, masks those frequency components in the input spectrum, and finetunes the detector on the resulting masked reconstructions. As illustrated in Figure 1 on the unseen `FLUX` generator, the resulting AUC remains markedly more stable than the base detector under post-processing.

Our main contributions are as follows: **(1)** We reframe AIGI detection robustness as a detector-adaptive problem and identify the frequency domain as the natural vantage point for diagnosing a detector's current input sensitivity. **(2)** We propose FAM, an attribution-guided mask-tuning framework that masks each detector's most-attended frequency components in the input spectrum and finetunes on the masked reconstructions. **(3)** We demonstrate that FAM consistently improves post-processing robustness on unseen generators for two pretrained detectors with and without a source-level alignment fix, and we additionally evaluate it on in-the-wild benchmarks.

2 Preliminaries

2.1 AIGI Detection and Failure Modes

We treat AIGI detection as a binary classification problem. Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ an image and $y_i \in \{0, 1\}$ its label ($y = 0$ real, $y = 1$ AI-generated). A detector $f_\theta : \mathbb{R}^{C\times H\times W} \to \mathbb{R}$ maps an image to a scalar logit, and the predicted label is $\hat{y} = \mathbb{1}[f_\theta(\mathbf{x}) > 0]$.

Two failure modes. Two failure modes motivate this work. First, detectors degrade sharply under common post-processing such as JPEG compression and resizing that real images routinely undergo when uploaded or re-encoded by content-delivery pipelines (Rajan & Lee, 2025; Li et al., 2025b). Second, they degrade under unseen-generator distribution shift, where a detector trained on one generative model produces unreliable outputs on images from a generator with a different autoencoder, sampler, or training corpus (Ojha et al., 2023; Corvi et al., 2023).

Shortcut reliance. Both failure modes share the same root cause: detectors learn shortcut features that correlate with one class on the training distribution but do not transfer (Geirhos et al., 2020). Such shortcuts include real-class artifacts such as the WEBP block grid in LSUN reals (Rajan & Lee, 2025) or JPEG artifacts in real corpora (Grommelt et al., 2024), and fake-class signatures such as generator-specific spectral fingerprints (Tan et al., 2024b; Wang et al., 2020; Corvi et al., 2023) that vanish on unseen generators. Once these cues are destroyed by post-processing or absent on a new generator, the detector loses its decision signal. Existing robustness methods (Section 5) typically target a single source-level bias with a deterministic operation at the data, architecture, or weight level. FAM instead operates on whichever cues a detector currently relies on, so that no enumeration of source-level biases is required.

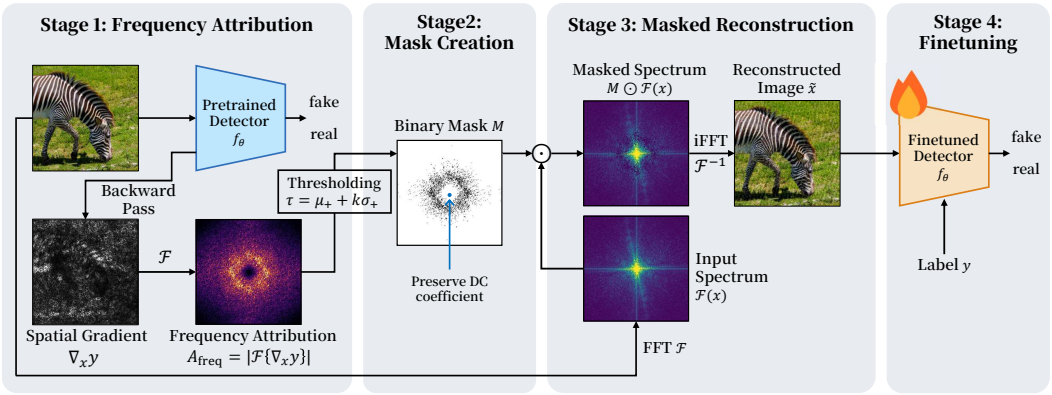


Figure 2: **Overview of the FAM pipeline. (1) Frequency Attribution** (Section 3.1): one backward pass through the *frozen* pretrained detector f_{θ_0} yields the channel-averaged FFT magnitude $\mathbf{A}_{\text{freq}}$ of the spatial gradient, computed once per training image and cached. **(2) Mask Creation** (Section 3.2): $\mathbf{A}_{\text{freq}}$ is binarized with a per-sample cutoff $\tau = \mu_+ + k\sigma_+$ and asymmetric multipliers $k_{\text{real}}, k_{\text{fake}}$, while the DC coefficient is preserved. **(3) Masked Reconstruction** (Section 3.3): the mask is applied to $\mathcal{F}(\mathbf{x})$ and inverted back to pixel space. Steps (2)–(3) are repeated per sample during **(4) Finetuning** (Section 3.3), where the full detector is updated on $(\tilde{\mathbf{x}}, y)$ pairs with binary cross-entropy.

2.2 Frequency-Domain Representation

FAM operates in the frequency domain via the 2D orthonormal discrete Fourier transform (DFT). For a single-channel image $\boldsymbol{x} \in \mathbb{R}^{H\times W}$, the forward and inverse transforms are

$$\boldsymbol{X}(u,v) = \frac{1}{\sqrt{HW}} \sum_{i,j} \boldsymbol{x}(i,j)\, e^{-2\pi\iota\left(\frac{ui}{H}+\frac{vj}{W}\right)}, \qquad \boldsymbol{x}(i,j) = \frac{1}{\sqrt{HW}} \sum_{u,v} \boldsymbol{X}(u,v)\, e^{+2\pi\iota\left(\frac{ui}{H}+\frac{vj}{W}\right)}, \tag{1}$$

where the shared $1/\sqrt{HW}$ factor makes $\mathcal{F}$ a unitary operator and ensures Parseval's identity $\|\boldsymbol{x}\|_2^2 = \|\boldsymbol{X}\|_2^2$, which is what lets the chain-rule identity in Section 3.1 hold without extra scaling. For multi-channel inputs, the DFT acts channel-wise and we average the resulting per-channel magnitudes whenever a single $H \times W$ map is required.

Why the frequency domain. Three properties make the frequency domain a more suitable vantage point than pixel space for examining a detector's input sensitivity. First, generative artifacts concentrate in identifiable frequency bands across generator families: high-frequency checkerboard residue from GAN-style transposed convolutions (Tan et al., 2024b; Wang et al., 2020) and mid-frequency residue from the autoencoder decoder of diffusion models (Rajan et al., 2025). Second, the post-processing perturbations that destroy detection accuracy at deployment also act in the same domain: JPEG compression substantially diminishes the high-frequency components on which CNN-era detectors depend (Li et al., 2025a), while low-pass filtering and downsizing further degrade the same band (Rajan & Lee, 2025), so the axis of detector sensitivity and the axis of deployment perturbation overlap. Third, a single coefficient $\boldsymbol{X}(u,v)$ contributes the same complex exponential to every spatial location, so a band-localized cue is by construction distributed across the entire pixel grid; pixel-wise spatial saliency therefore cannot isolate such a cue, whereas frequency-domain attribution localizes it to a single (u,v), a gap we verify empirically with a Spatial-FAM ablation in Section 4.

3 Frequency Attribution Mask (FAM)

We present *Frequency Attribution Mask* (FAM), a detector-adaptive mask-tuning framework for AIGI detection. Figure 2 illustrates an overview of the pipeline. We first describe how the per-sample frequency-domain attribution is computed via a chain-rule identity (Section 3.1). We then turn this attribution into a per-sample binary mask that marks the dominant high-attribution bins for

removal (Section 3.2), and the masked reconstruction is fed back to the detector, which is finetuned end-to-end to find new evidence in the residual signal (Section 3.3).

3.1 Frequency Attribution

To diagnose which frequency cues a trained detector currently leans on, we require a per-sample score $\mathbf{A}_{\text{freq}}(u, v)$ that quantifies how each frequency coefficient $\boldsymbol{X}(u, v)$ of the input contributes to the detector's prediction $y = f_\theta(\boldsymbol{x})$. Spatial saliency $|\partial y / \partial \boldsymbol{x}(i, j)|$ (Simonyan et al., 2013) is the standard tool in pixel space, but a band-localized cue is by construction distributed across all spatial positions (Section 2.2), so spatial saliency cannot isolate the offending band. We therefore extend gradient-based saliency to the frequency domain.

The key observation is that, under the orthonormal DFT, the gradient of y with respect to a frequency coefficient is itself the DFT of the spatial gradient. Reparameterizing $\boldsymbol{x}$ through the inverse DFT in Equation (1) and applying the chain rule yields

$$\frac{\partial y}{\partial \boldsymbol{X}(u, v)} \;=\; \sum_{i,j} \frac{\partial y}{\partial \boldsymbol{x}(i, j)} \frac{\partial \boldsymbol{x}(i, j)}{\partial \boldsymbol{X}(u, v)} \;=\; \mathcal{F}\left(\frac{\partial y}{\partial \boldsymbol{x}}\right)(u, v), \tag{2}$$

where the second equality uses the linearity of the orthonormal DFT and treats the spatial gradient $\nabla_{\boldsymbol{x}}\, y$ as an image to be transformed; the full step-by-step derivation is given in Appendix B.2. Since $\boldsymbol{x}$ is real, the gradient is complex-valued, and we use its magnitude as our scalar attribution. Averaging across color channels yields a single frequency attribution map

$$\mathbf{A}_{\text{freq}}(u, v) \;=\; \frac{1}{C} \sum_{c=1}^{C} \big|\mathcal{F}(\nabla_{\boldsymbol{x}_c}\, y)(u, v)\big|. \tag{3}$$

Equation (2) requires only that f_θ be differentiable in $\boldsymbol{x}$, so each detector contributes its own $\mathbf{A}_{\text{freq}}$ for each input at the cost of one backward pass plus a single FFT, and the detector itself is never queried on out-of-distribution frequency-domain inputs.

3.2 Attribution-Guided Mask

The attribution distribution is heavily right-skewed: a small number of bands carry most of the mass and a long tail of near-zero values dominates the rest. A fixed top-p quantile therefore either over-masks the tail or fails to cover the dominant peak. We use a per-sample, distribution-adaptive threshold based on the first two moments of the positive part of $\mathbf{A}_{\text{freq}}$,

$$\tau = \mu_+ + k \cdot \sigma_+, \quad \mathbf{M}(u, v) = \mathbb{1}\big[\mathbf{A}_{\text{freq}}(u, v) \le \tau\big], \tag{4}$$

where $\mu_+ = \mathbb{E}[\mathbf{A}_{\text{freq}} \mid \mathbf{A}_{\text{freq}} > 0]$, $\sigma_+ = \text{Std}[\mathbf{A}_{\text{freq}} \mid \mathbf{A}_{\text{freq}} > 0]$, and $k > 0$ is a scalar multiplier that controls how aggressively the high-attribution tail is removed. Smaller k removes more coefficients; larger k leaves the input nearly intact. We additionally force $\mathbf{M}(0, 0) = 1$ so that the DC coefficient is never removed, which prevents the masked reconstruction from drifting in global brightness.

Asymmetric per-class control. Real and fake images carry different cues: real images suffer primarily from real-class spectral biases such as the WEBP block grid in LSUN (Rajan & Lee, 2025; Grommelt et al., 2024), while fake images carry generator-specific frequency fingerprints concentrated in different bands. FAM therefore allows asymmetric masking with separate multipliers $k_{\text{real}}, k_{\text{fake}}$, computed independently on each class's positive attribution distribution. The symmetric setting $k_{\text{real}} = k_{\text{fake}} = k$ recovers a single-knob version.

3.3 Finetuning with Masked Reconstructions

After the mask is built in the frequency domain, we project it back into pixel space so that the detector receives an input on the same domain on which it was trained:

$$\tilde{\mathbf{x}} = \mathcal{F}^{-1}\big(\mathbf{M} \odot \mathcal{F}(\mathbf{x})\big), \tag{5}$$

where $\odot$ denotes elementwise multiplication broadcast across channels. Because $\mathbf{M}$ is real and $\mathbf{x}$ is real, $\tilde{\mathbf{x}}$ is real up to numerical error and can be passed to the detector without architectural modification. The detector is then finetuned end-to-end on $(\tilde{\mathbf{x}}, y)$ pairs with the standard binary cross-entropy loss.

Precomputed attribution. The attribution $\mathbf{A}_{\text{freq}}$ of Equation (3) is obtained in a *single* pass over the training set with the *frozen* pretrained detector f_{θ_0} and cached. It is *never recomputed* as fine-tuning proceeds. Each training step only reads the cached map and derives the threshold τ, the mask $\mathbf{M}$, and the masked reconstruction of Equation (5) from it. Implementation details, hyperparameters, and pseudocode appear in Appendix B.

4 EXPERIMENTS

We empirically validate FAM on two representative pretrained AIGI detectors, Corvi and Rajan, which share a ResNet-50 backbone but differ in how their synthetic training data is constructed. Our experiments investigate **(1)** whether FAM improves robustness to common post-processing operations on top of base detectors that are already trained against them, one of which also received a source-level bias-fix, **(2)** whether the gains transfer to in-the-wild benchmarks, and **(3)** why FAM produces these gains, examined through faithfulness and ablation analyses. Implementation details and extended analyses appear in Appendices B–E.

4.1 EXPERIMENTAL SETUP

Datasets. For real images, we use $3{,}000$ images from Redcaps (Desai et al., 2021), a web-crawled corpus with natural variation in content, resolution, and compression. Fakes come from nine generators unseen by any base detector: Stable Diffusion (Rombach et al., 2022), Midjourney (Midjourney, 2023), Kandinsky (Razzhigaev et al., 2023), Playground (Li et al., 2024), PixArt (Chen et al., 2023), LCM (Luo et al., 2023), FLUX (Labs, 2024), Wuerstchen (Pernias et al., 2023), and aMUSEd (Patil et al., 2024). We evaluate on the post-processed (`test_processed`) split of LDM-FakeDetect (Rajan & Lee, 2025), and additionally on in-the-wild data: Chameleon (Yan et al., 2024), SynthWildX (Cozzolino et al., 2024), and the two open-platform subsets of AIGIBench (Li et al., 2025b), SocialRF and CommunityAI, whose fakes were collected from social-media platforms and AI-art communities. Further details are in Appendix C.1.

Baselines. We compare against a comprehensive set of state-of-the-art detectors spanning multiple paradigms. From CNN-based methods, we include Corvi (Corvi et al., 2023), Rajan (Rajan et al., 2025), NPR (Tan et al., 2024b), SAFE (Li et al., 2025a), AIDE (Yan et al., 2024), FerretNet (Liang et al., 2025), LaDeDa (Cavia et al., 2024), and DRCT (Chen et al., 2024). From CLIP-based approaches, we evaluate UFD (Ojha et al., 2023), C2P-CLIP (Tan et al., 2025), RINE (Koutlis & Papadopoulos, 2024), and ClipDet (Cozzolino et al., 2024). We also consider Vision-Transformer-based methods including CoDE (Baraldi et al., 2024) and SPAI (Karageorgiou et al., 2025), as well as training-free detectors AEROBLADE (Ricker et al., 2024) and WaRPAD (Choi et al., 2025). Additionally, we compare against Corvi+ and Rajan+ (Rajan & Lee, 2025), which apply the Stay-Positive algorithm to retrain only the last layer with non-negative weight constraints. Detailed descriptions of each baseline are provided in Appendix C.2.

We apply FAM to two pretrained detectors and refer to the resulting methods as Corvi$^{\mathcal{F}}$ and Rajan$^{\mathcal{F}}$. Corvi and Rajan share a ResNet-50 backbone trained on real images from LSUN and COCO but differ in how synthetic training data is constructed: Corvi uses LDM-generated images conditioned on text prompts, while Rajan uses VAE reconstructions of real images to ensure pixel-level alignment between training pairs. Rajan thus already carries a pixel-level alignment fix, yet this alignment alone is not sufficient: as Section 4.2 shows, Rajan still degrades under post-processing on unseen generators.

Metrics. We report AUC as a threshold-independent ranking measure and balanced accuracy $\text{B. Acc} = (\text{R.Acc} + \text{F.Acc})/2$ as a single, threshold-aware summary of detector reliability, where R.Acc and F.Acc denote the real and fake accuracies at a fixed decision threshold. B. Acc equals classification accuracy under a class-balanced test set.

4.2 ROBUSTNESS TO POST-PROCESSING

A practical detector must maintain reliable performance when images undergo common transformations during storage and transmission. Following the prior evaluation protocol (Rajan & Lee, 2025),

Table 1: **Robustness to post-processing on unseen generators.** Per-generator AUC and average (AUC, balanced accuracy B. Acc $=$ $(\mathrm{R.Acc}+\mathrm{F.Acc})/2$) on the post-processed split of the LDM-FakeDetect benchmark (Rajan & Lee, 2025). Methods are grouped by architecture: CLIP-based, ViT-based (V), Training-free (TF), and CNN-based detection backbones. Training-free detectors only expose AUC ("-" for B. Acc). **Bold** = best, underline = second best. FAM-finetuned variants are highlighted in blue.

	Detectors	SD	MJ	KD	PG	PixArt	LCM	FLUX	Wuerst.	aMUSEd	Avg	
		AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC	AUC	B. Acc
CLIP	UFD	59.2	42.9	56.4	53.3	54.5	57.2	25.0	73.7	85.3	56.4	50.0
	C2P-CLIP	60.9	68.5	70.6	74.6	70.6	61.0	35.5	71.7	93.0	67.4	56.1
	RINE	76.0	53.3	64.8	59.7	45.3	70.3	45.7	85.2	86.4	65.2	61.8
	ClipDet	71.9	75.2	75.4	76.9	79.3	76.7	79.8	89.2	87.8	79.1	71.3
V	CoDE	82.2	68.8	64.4	69.4	76.2	67.4	63.6	70.3	70.2	70.3	65.1
	SPAI	77.4	54.0	69.1	55.0	66.4	75.7	57.3	74.4	72.7	66.9	61.1
TF	AEROBLADE	56.1	78.0	63.7	56.7	58.2	42.5	54.6	56.6	67.0	59.3	–
	WaRPAD	64.5	59.8	65.4	60.1	57.1	52.4	35.0	82.2	91.2	63.1	–
CNN	NPR	50.3	63.0	62.7	66.1	75.6	35.9	41.9	56.0	27.9	53.3	53.2
	SAFE	53.9	40.1	48.2	38.1	36.0	47.3	48.2	51.7	57.5	46.8	50.1
	AIDE	65.8	64.9	68.3	61.1	69.8	58.3	65.8	75.6	76.1	67.3	55.9
	FerretNet	56.5	44.7	59.1	64.9	73.1	49.9	58.7	65.6	60.9	59.3	50.0
	LaDeDa	45.4	57.1	61.6	64.6	73.8	33.6	42.1	51.9	33.9	51.6	50.1
	DRCT	90.5	79.5	86.7	71.7	72.1	73.3	67.3	77.1	85.6	78.2	71.4
	Corvi	91.1	80.0	80.7	66.8	75.7	81.1	64.2	82.8	90.7	79.2	72.5
	Corvi+	95.5	85.1	89.9	88.0	93.2	91.8	84.3	**91.8**	85.7	89.5	73.9
	Corvi$^{\mathcal{F}}$	94.0	87.6	86.4	82.9	83.3	89.2	78.4	87.6	**97.0**	87.4	81.9
	Rajan	97.7	**94.5**	92.6	93.8	93.9	99.7	81.2	87.1	95.9	92.9	79.0
	Rajan+	96.1	86.1	91.6	89.3	91.4	97.3	84.7	82.1	96.6	90.6	80.7
	Rajan$^{\mathcal{F}}$	**99.0**	**94.5**	**95.6**	**97.0**	**96.8**	**99.9**	**88.0**	91.6	94.3	**95.2**	**87.8**

we construct a post-processed test set by applying its post-processing pipeline to the synthetic images, and report the results in Table 1. Both base detectors' performance degrades sharply under post-processing, with the collapse most pronounced on FLUX where each base detector loses the most. For both detectors, applying FAM improves both AUC and balanced accuracy, although both bases are already hardened against exactly these perturbations: both are trained with strong augmentation covering compression and blur, and Rajan additionally aligns real and fake in its training data. Corvi$^{\mathcal{F}}$ improves Corvi by 8.2 AUC and 9.4 B. Acc points, and Rajan$^{\mathcal{F}}$ improves Rajan by 2.3 AUC and 8.8 B. Acc points, reaching the best average AUC (95.2) and balanced accuracy (87.8) among all compared detectors. That Rajan still leaves this margin for FAM despite its pixel-level alignment is the point, since re-anchoring it onto a narrow, transferable set of frequency cues reaches bands no predefined alignment was aimed at and that post-processing destroys. FAM also outperforms Stay-Positive, the strongest post-hoc fix for these backbones, by 8.0 and 7.1 balanced-accuracy points over Corvi+ and Rajan+, and where Rajan+ gives up 2.3 AUC against its own base to gain 1.7 B. Acc, FAM raises both.

4.3 Generalization to In-the-Wild Benchmarks

We further evaluate detectors in the wild, where both the generators and the upstream transformations are partially unknown, on Chameleon, SynthWildX, and the two open-platform subsets of AIGIBench, SocialRF and CommunityAI. The full per-method results are reported in Table 5. As in the post-processed setting, Corvi$^{\mathcal{F}}$ improves Chameleon AUC by 9.4 and balanced accuracy by 16.0 points, the best balanced accuracy there, and raises CommunityAI balanced accuracy by 22.0 points to the top spot. Rajan$^{\mathcal{F}}$ takes the top AUC on SocialRF, lifts SynthWildX by 3.6 AUC and 2.6 B. Acc points, and on Chameleon trails its base's AUC (78.9 vs. 80.0) while raising balanced accuracy by 4.4 points. Both FAM variants improve over their base in balanced accuracy on all four test sets, and in AUC except for the Rajan variant on Chameleon and CommunityAI, indicating that detector-adaptive masking transfers beyond the LDM-family test set.

4.4 Additional Results

Attribution faithfulness. We test whether $\mathbf{A}_{\text{freq}}$ identifies the cues the detector depends on with the Most-Relevant-First (MoRF) input-perturbation protocol (Samek et al., 2017; Petsiuk et al., 2018): rank frequency bins by $\mathbf{A}_{\text{freq}}$, zero out the top-$K\%$ on the input spectrum, invert back to pixel space, and re-score. A faithful attribution produces a sharp AUC drop at small K. Figure 3 reports the result on 80 FLUX images. At $K = 1\%$, Corvi$^{\mathcal{F}}$ drops from AUC 0.99 to 0.64, while the same procedure barely moves Corvi base (from 0.97 to 0.94). FAM thus consolidates the detector onto a small, transferable set of high-leverage bins, whereas the base detector relies on a diffuse spectral support whose individual bins are dispensable. The same consolidation explains the AUC stability of Corvi$^{\mathcal{F}}$ under post-processing in Figure 1: a narrow support is unlikely to be erased by any single perturbation. Full MoRF curves under JPEG and resize are reported in Appendix E.3.

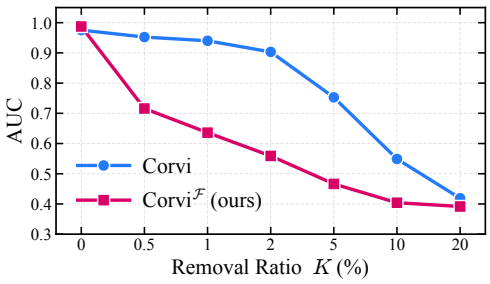


Figure 3: **Most-Relevant-First (MoRF) faithfulness on `FLUX`** (↓ lower is better)**.** A steeper drop at small K indicates a more faithful attribution that rests on a narrow set of high-leverage cues. FAM consolidates the detector onto such a narrow set, while the base detector relies on a more diffuse spectral support.

Score distribution analysis. The same consolidation also yields more overlap-resistant score distributions on the post-processed split: each base detector places a substantial fraction of fake logits on the wrong side of the decision threshold, whereas the FAM variants push the fake distribution back to the positive side and sharpen the separation from the real distribution. Figure 5 visualizes the post-processed-split score distributions for both base detectors and their FAM-finetuned counterparts on FLUX, the most challenging unseen generator in our evaluation. A detector consolidated onto a small, transferable cue can thus separate the two classes with a single calibrated decision threshold even after deployment perturbations.

Per-operation robustness. The post-processed split chains several operations together; Figure 4 applies JPEG, resizing, and WebP in isolation. Resizing produces the largest spread (+9.5 AUC on Corvi, +11.7 on Rajan), WebP a similarly large gain (+11.3 Rajan, +6.5 Corvi), and JPEG a smaller but consistent gain (+4.4 Corvi, +2.5 Rajan).

4.5 Ablation Studies

We ablate the two design choices at the heart of FAM—*where* to mask and *how* to choose what to mask—below; a sensitivity sweep over the threshold multipliers $(k_{\text{real}}, k_{\text{fake}})$ is deferred to Appendix B.4, where every grid cell is shown to improve over the base detector in balanced accuracy.

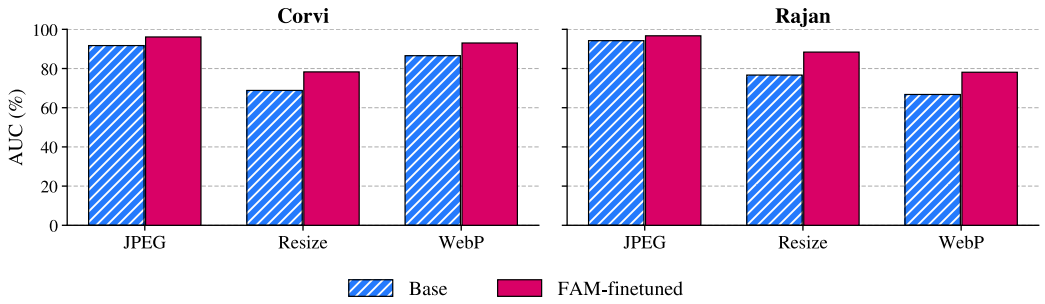


Figure 4: **Per-operation post-processing robustness.** Average AUC across the nine test generators under isolated JPEG ($q \in \{30, 40, 50\}$), resizing (scale $\in [0.25, 0.5]$), and WebP ($q \in \{30, 50, 70\}$), for each base detector and its FAM-finetuned counterpart.

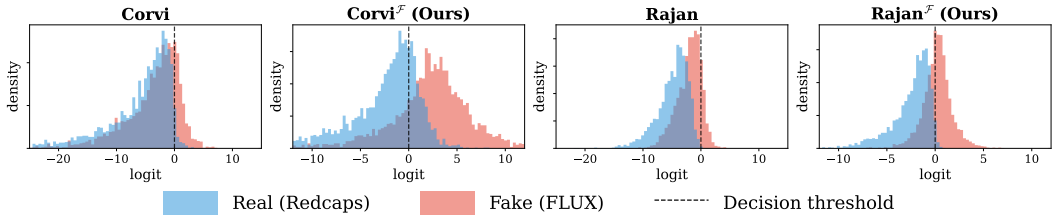


Figure 5: **Score-distribution shift after FAM finetuning on the post-processed split.** Histogram of detector logits for real (Redcaps, blue) and fake-FLUX (red) inputs, with the dashed black line marking the decision threshold. For both base detectors (Corvi and Rajan), a substantial fraction of fake logits falls below zero, whereas the corresponding FAM-finetuned variants keep the fake distributions on the positive side and well separated from the real distribution.

What to mask and where: selection vs. domain. FAM entangles two design choices: *where* to mask (frequency vs. spatial domain) and *how* to choose what to mask (gradient saliency vs. random selection at matched sparsity). To isolate each axis, we compare FAM against four controls: a spatial-domain analog (Spatial-FAM) that replaces the FFT-of-gradient attribution with a channel-averaged absolute-gradient saliency map and zeroes out the high-attribution pixels before finetuning, and two random masks that fix the same per-class sparsity as FAM but distribute the removed bins uniformly at random in either the frequency or spatial domain. Table 2 reports both axes. Holding selection fixed at saliency, the frequency mask outperforms the spatial mask by 1.7 AUC and 2.4 B. Acc points on Corvi and by 1.1 AUC on Rajan, where the two tie in B. Acc: a band-localized residue is, by definition, distributed across all spatial positions, so a pixel-wise saliency map cannot isolate it as cleanly as a frequency-domain attribution. Holding the domain fixed at frequency, gradient saliency adds another 1.8 AUC and 2.7 B. Acc points over the random frequency mask on Corvi and 2.8 AUC and 3.6 B. Acc points on Rajan, indicating that the per-sample saliency identifies a small set of bins that random selection at the same sparsity cannot recover. The two effects compose, with FAM taking or sharing the top spot in each variant.

Table 2: **Ablating the finetuning itself, the selection rule (saliency vs. random), and the domain (frequency vs. spatial).** Post-processed AUC and balanced accuracy averaged across the nine test generators. Unmasked finetuning repeats the FAM schedule on unmasked inputs. Random masks use the same per-class sparsity as FAM, so the comparison controls for the fraction of removed bins.

Method	Avg AUC	Avg B. Acc
Corvi variant		
Base (no masking)	79.2	72.5
Unmasked finetuning	80.4	74.4
Random Spatial Mask	75.5	55.1
Random Frequency Mask	85.6	79.2
Spatial Attribution Mask	85.7	79.5
Frequency Attribution Mask	**87.4**	**81.9**
Rajan variant		
Base (no masking)	92.9	79.0
Unmasked finetuning	94.2	81.6
Random Spatial Mask	92.8	87.7
Random Frequency Mask	92.4	84.2
Spatial Attribution Mask	94.1	**87.8**
Frequency Attribution Mask	**95.2**	**87.8**

5 Related Work

AI-generated image detection. Early detectors targeted GAN-generated images and exploited regular spectral artifacts left by transposed-convolution upsampling (Wang et al., 2020; Tan et al., 2024b). With the rise of diffusion-based generators (Ho et al., 2020; Rombach et al., 2022), the artifact landscape shifted toward subtler traces, and detectors emerged across CNN-based (Corvi et al., 2023; Rajan et al., 2025; Tan et al., 2024b; Li et al., 2025a; Yan et al., 2024; Liang et al., 2025; Cavia et al., 2024; Chen et al., 2024), CLIP-based (Ojha et al., 2023; Tan et al., 2025; Koutlis & Papadopoulos, 2024; Cozzolino et al., 2024), ViT-based (Baraldi et al., 2024), and training-free (Ricker et al., 2024; Choi et al., 2025) paradigms. FAM is orthogonal to this line of work: it does not propose a new detector but finetunes a pretrained detector to suppress the input-level cues the detector itself attributes its decisions to.

Mitigating shortcut learning in AIGI detection. A central recent direction commits to a specific source-level mismatch and removes it through a deterministic operation on the training distribution: SemGIR (Yu et al., 2024) and DRCT (Chen et al., 2024) align semantic content via captioned reconstruction; AlignedForensics (Rajan et al., 2025) aligns at the pixel level via a single VAE forward pass. Stay-Positive (Rajan & Lee, 2025) addresses shortcuts at the weight level by constraining the final classifier layer to non-negative weights, while patch-shuffle methods (Gye et al., 2025; Zheng et al., 2024) intervene at the input layer by destroying semantic structure. All of these interventions commit in advance to which cue is removed. FAM instead locates the offending cue adaptively, using each detector's own input gradient to score the exact contribution of every frequency, since an AIGI shortcut is a frequency band that changes every pixel at once and is therefore out of reach of any fixed, predefined input operation. Frequency is also where JPEG, resizing, and WebP act, so removing the bands a detector leans on rehearses the loss post-processing will cause. Real- and fake-class biases sit in different bands, so FAM uses class-asymmetric thresholds.

Frequency-aware detection and attribution methods. Frequency-aware detectors (Tan et al., 2024a; Kashiani et al., 2025; Karageorgiou et al., 2025) treat spectral content as the primary detection cue and learn forensic representations directly in the frequency domain. FAM takes a different view of the spectrum: rather than using frequency features as a positive signal of authenticity, it uses frequency-domain attribution as a diagnostic of the detector's existing reliance on spectral cues, and masks out exactly those bands the detector has already learned to depend on. The two perspectives are complementary, since spectrum-aware detectors and FAM target different stages of the same pipeline. Gradient-based attribution has been refined with path integration (Sundararajan et al., 2017), noise averaging (Smilkov et al., 2017), and class activation maps (Fu et al., 2020), and its faithfulness is commonly evaluated by removing the most relevant input features (Samek et al., 2017; Hooker et al., 2019). At the attribution level, we use the simplest gradient saliency (Simonyan et al., 2013); our contribution is not a new attribution rule but the chain-rule identity in Equation (2), which lifts spatial saliency into a frequency-domain attribution map at no extra cost and without ever feeding the detector an out-of-distribution input.

6 Conclusion

We proposed *Frequency Attribution Mask* (FAM) to address shortcut reliance in AIGI detection, a phenomenon where pretrained detectors latch onto narrow frequency cues that fail to transfer once test images come from unseen generators or pass through routine post-processing. Departing from the prevailing bias-enumeration paradigm, FAM intervenes at the point where any source-level bias actually manifests in a trained detector, namely its own input sensitivity. FAM derives a per-sample frequency-domain attribution directly from the detector's input gradient in a single cached pass over the training set, binarizes the attribution into a class-asymmetric mask that removes the highest-attention bands from the input spectrum, and finetunes the detector on the resulting masked reconstructions to find new evidence in the residual signal. Across two pretrained ResNet-50 detectors, with and without a source-level alignment fix, FAM delivers consistent gains on unseen generators and improves over the base detectors on in-the-wild benchmarks, with the largest improvements where residual shortcut reliance is most severe. By coupling diagnostic frequency attribution with targeted mask-tuning, FAM offers a unified post-hoc remedy for shortcut learning in AIGI detection without modifying the detector architecture or its source training distribution.

Limitations and future work. FAM targets the global spectral footprint left by an end-to-end image generator and is best suited to detectors that operate on whole-image synthesis. Local-manipulation tasks such as face-swap or face-reenactment leave a different forensic trace; a natural extension is to combine the chain-rule frequency attribution with a region-of-interest map (e.g. a face-aware spatial prior or a per-patch attribution that is then transformed into the frequency domain) so that the mask can adapt to localized artifacts. Having shown that finetuning guided by a single attribution map from the frozen base detector is already effective, an interesting direction is to recompute the map several times during finetuning, so that the mask keeps tracking the cues the updated detector comes to rely on.

AI USE STATEMENT

We used generative AI tools to assist with implementing code, conducting experiments, and analyzing experimental results. We also used these tools for polishing the manuscript and for searching recent related work in the literature review. We did not use them for the research idea or for developing and designing the method. The authors verified all technical content and take full responsibility for the entire content of the paper, including the text, code, figures, and reported results.

ETHICS STATEMENT

FAM improves AIGI detector robustness under post-processing and unseen generators, supporting reliable content authentication against misinformation, non-consensual deepfakes, and copyright infringement. We see no direct negative impact, though practitioners should assess FAM-finetuned detectors against adaptive adversaries that may exploit frequency-domain attribution for evasion.

REPRODUCIBILITY STATEMENT

The full FAM procedure is given in Algorithm 1, and the full step-by-step derivation of the chain-rule identity in Equation (2) is provided in Appendix B.2. Details of the finetuning settings, the selection of the threshold multipliers, the ablation controls, and the computational resources are provided in Appendix B, and details of the training and evaluation data and the post-processing protocol in Appendix C. The baselines, evaluated with their official implementations and pretrained checkpoints, are described in Appendix C.2. The supplementary material includes the FAM implementation, code for the frequency attribution and mask-tuning stages, evaluation scripts, and reproduction instructions.

REFERENCES

Lorenzo Baraldi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, Alessandro Nicolosi, and Rita Cucchiara. Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In *European Conference on Computer Vision*, pp. 199–216. Springer, 2024.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real-world. *arXiv preprint arXiv:2406.09398*, 2024.

George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10759–10769, 2024.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. PixArt-α: Fast training of diffusion transformer for photo-realistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Sungik Choi, Hankook Lee, and Moontae Lee. Training-free detection of AI-generated images via cropping robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Davide Cozzolino, Giovanni Poggi, Riccardo Corvi, Matthias Nießner, and Luisa Verdoliva. Raising the bar of AI-generated image detection with CLIP. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4356–4366, 2024.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Patrick Grommelt, Louis Weiss, Franz-Josef Pfreundt, and Janis Keuper. Fake or jpeg? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision Workshops*, pp. 80–95. Springer, 2024.

Seoyeon Gye, Junwon Ko, Hyounguk Shon, Minchan Kwon, and Junmo Kim. Reducing the content bias for AI-generated image detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 399–408, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18706–18717, 2025.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018.

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*, pp. 4401–4410, 2019.

Hossein Kashiani, Niloufar Alipour Talemi, and Fatemeh Afghah. Freqdebias: towards generalizable deepfake detection via consistency-driven frequency debiasing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8775–8785. IEEE, 2025.

Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection. In *European Conference on computer vision*, pp. 394–411. Springer, 2024.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

Black Forest Labs. FLUX, 2024. Available at `https://github.com/black-forest-labs/flux`.

Daiqing Li, Aleks Kamko, Ehsan Akhgari, Ali Sabet, Linmiao Xu, and Suhail Doshi. Playground v2. 5: Three insights towards enhancing aesthetic quality in text-to-image generation. *arXiv preprint arXiv:2402.17245*, 2024.

Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 2405–2414, 2025a.

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*, 2025b.

Shuqiao Liang, Jian Liu, Renzhang Chen, and Quanlong Guan. FerretNet: Efficient synthetic image detection via local pixel dependencies. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Midjourney. Midjourney. `https://www.midjourney.com/`, 2023. Accessed: 2025-05-01.

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.

Suraj Patil, William Berman, Robin Rombach, and Patrick von Platen. amused: An open muse reproduction. *arXiv preprint arXiv:2401.01808*, 2024.

Pablo Pernias, Dominic Rampas, and Marc Aubreville. Wuerstchen: Efficient pretraining of text-to-image models. 2023.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

Anirudh Sundara Rajan and Yong Jae Lee. Stay-positive: A case for ignoring real image features in fake image detection. In *International Conference on Machine Learning (ICML)*, 2025.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *International Conference on Learning Representations (ICLR)*, 2025.

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. Kandinsky: an improved text-to-image synthesis with image prior and latent diffusion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 286–295, 2023.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

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*, pp. 10684–10695, 2022.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

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*, pp. 5052–5060, 2024a.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28130–28139, 2024b.

Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7184–7192, 2025.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8695–8704, 2020.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection. *arXiv preprint arXiv:2406.19435*, 2024.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Xiao Yu, Kejiang Chen, Kai Zeng, Han Fang, Zijin Yang, Xiuwei Shang, Yuang Qi, Weiming Zhang, and Nenghai Yu. Semgir: Semantic-guided image regeneration based method for ai-generated image detection and attribution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8480–8488, 2024.

Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems*, 37:59570–59596, 2024.

A NOTATIONS

Table 3 summarizes the notation used throughout the paper. We treat AI-generated image detection as a binary classification problem (Section 2.1), with input images $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, binary labels $y \in \{0, 1\}$, and a detector f_θ that maps an image to a scalar logit. $\mathcal{F}$ denotes the 2D orthonormal discrete Fourier transform applied per channel and $\mathcal{F}^{-1}$ its inverse.

Table 3: **Summary of notation used in the paper.**

Notation	Description
Input space and detector	
$\mathbf{x} \in \mathbb{R}^{C\times H\times W}$	Input image; default $H = W = 256$, $C = 3$.
$y \in \{0, 1\}$	Binary label (0 = real, 1 = AI-generated).
$f_\theta : \mathbb{R}^{C\times H\times W} \to \mathbb{R}$	Detector with parameters θ, mapping an image to a scalar logit.
$\hat{y} = \mathbb{1}[f_\theta(\mathbf{x}) > 0]$	Predicted label at the calibrated decision threshold.
Frequency representation	
$\mathcal{F}(\cdot)$, $\mathcal{F}^{-1}(\cdot)$	2D orthonormal forward / inverse DFT, applied per channel.
$\mathbf{X} = \mathcal{F}(\mathbf{x})$	Frequency coefficients, complex-valued on the $H \times W$ grid.
Frequency attribution	
$\nabla_{\mathbf{x}}\, y$	Spatial gradient of the detector logit with respect to the input.
$\mathbf{A}_{\text{freq}}(u, v)$	Frequency-domain attribution; channel-mean FFT magnitude of $\nabla_{\mathbf{x}}\, y$ (main paper, Eq. 3).
$\mathbf{A}_{\text{spat}}(i, j)$	Spatial attribution; channel-mean absolute spatial gradient (Spatial-FAM ablation only).
Mask construction	
μ_+, σ_+	Mean and standard deviation of the strictly positive entries of $\mathbf{A}_{\text{freq}}$.
$k_{\text{real}}, k_{\text{fake}} > 0$	Per-class threshold multipliers (selected per detector on a validation split; see Table 4).
$\tau = \mu_+ + k \cdot \sigma_+$	Per-sample threshold on $\mathbf{A}_{\text{freq}}$, with $k \in \{k_{\text{real}}, k_{\text{fake}}\}$ chosen by the input's label.
$\mathbf{M}(u, v) = \mathbb{1}[\mathbf{A}_{\text{freq}}(u, v) \le \tau]$	Binary mask (1 = keep, 0 = remove). DC is preserved: $\mathbf{M}(0, 0) = 1$.
$\tilde{\mathbf{x}} = \mathcal{F}^{-1}(\mathbf{M} \odot \mathcal{F}(\mathbf{x}))$	Masked reconstruction in pixel space.

B IMPLEMENTATION DETAILS

B.1 ALGORITHM

Algorithm 1 summarizes the two stages of FAM. **Stage 1** makes a single pass over the training set with the *frozen* pretrained detector f_{θ_0} and caches the per-image frequency attribution $\mathbf{A}_{\text{freq}}$. This is the only point where a gradient with respect to the *input* is taken. **Stage 2** finetunes the detector, deriving the threshold τ, the mask $\mathbf{M}$, and the masked reconstruction $\tilde{\mathbf{x}}$ from the cached $\mathbf{A}_{\text{freq}}$ each time a sample is drawn. Its backward pass is the ordinary one, with respect to the *parameters*.

B.2 FULL CHAIN-RULE DERIVATION

We give the complete derivation of the identity used in Eq. 2 of the main paper. The detector f_θ takes spatial pixel inputs $\mathbf{x} \in \mathbb{R}^{C\times H\times W}$, and we write $y = f_\theta(\mathbf{x})$. For notational clarity we work on a single channel. The multi-channel extension is trivial, since the orthonormal 2D DFT acts independently per channel. Reparameterizing $\mathbf{x}$ through the inverse DFT in Eq. 1 of the main paper,

$$\mathbf{x}(i, j) = \frac{1}{\sqrt{HW}} \sum_{u,v} \mathbf{X}(u, v)\, e^{+2\pi\iota\left(\frac{ui}{H} + \frac{vj}{W}\right)},$$

Algorithm 1 Frequency Attribution Mask (FAM). Stage 1 runs once over the training set and caches the attribution. Stage 2 is the finetuning loop, which reads that cache.

Require Pretrained detector f_{θ_0}, training set $\mathcal{D}$, threshold multipliers $k_{\text{real}}, k_{\text{fake}}$, optimizer $\mathcal{O}$, augmentation $\mathcal{A}$
Stage 1: frequency attribution `// once, θ0 frozen`
for each $\mathbf{x} \in \mathcal{D}$ **do**
  $\mathbf{g} \leftarrow \nabla_{\mathbf{x}} f_{\theta_0}(\mathbf{x})$
  `// frequency-domain attribution (main paper, Eq. 3), cached`
  $\mathbf{A}_{\text{freq}}[\mathbf{x}](u,v) \leftarrow \frac{1}{C}\sum_{c=1}^{C} |\mathcal{F}(\mathbf{g}_c)(u,v)|$
end for
Stage 2: mask-tuning `// reads the cache, no new input gradients`
$\theta \leftarrow \theta_0$
for a fixed number of epochs **do**
  for each minibatch $\mathcal{B} \subset \mathcal{D}$ **do**
    `// per sample (x, y) ∈ B, from the cached A_freq[x]`
    $k_y \leftarrow k_{\text{real}}$ if $y = 0$, else k_{fake}
    $\tau \leftarrow \mu_+(\mathbf{A}_{\text{freq}}[\mathbf{x}]) + k_y\, \sigma_+(\mathbf{A}_{\text{freq}}[\mathbf{x}])$
    $\mathbf{M} \leftarrow \mathbb{1}[\mathbf{A}_{\text{freq}}[\mathbf{x}] \le \tau],\ \ \mathbf{M}(0,0) \leftarrow 1$
    `// masked reconstruction, clamped to [0,1], then augmented`
    $\tilde{\mathbf{x}} \leftarrow \mathcal{A}\big(\text{clamp}_{[0,1]}\big(\mathcal{F}^{-1}(\mathbf{M} \odot \mathcal{F}(\mathbf{x}))\big)\big)$
    $\theta \leftarrow \mathcal{O}\big(\theta,\ \nabla_\theta \sum_{(\mathbf{x},y)\in\mathcal{B}} \mathcal{L}_{\text{BCE}}(f_\theta(\tilde{\mathbf{x}}), y)\big)$
  end for
end for
return finetuned detector f_θ

the partial derivative of $\mathbf{x}(i,j)$ with respect to $\mathbf{X}(u,v)$ is the corresponding term of the inverse-DFT sum:

$$\frac{\partial \mathbf{x}(i,j)}{\partial \mathbf{X}(u,v)} = \frac{1}{\sqrt{HW}}\, e^{+2\pi\iota\left(\frac{ui}{H}+\frac{vj}{W}\right)}. \tag{6}$$

Applying the chain rule to $y = f_\theta(\mathbf{x})$ yields

$$\begin{aligned}\frac{\partial y}{\partial \mathbf{X}(u,v)} &= \sum_{i,j} \frac{\partial y}{\partial \mathbf{x}(i,j)} \cdot \frac{\partial \mathbf{x}(i,j)}{\partial \mathbf{X}(u,v)} \\ &= \frac{1}{\sqrt{HW}} \sum_{i,j} [\nabla_{\mathbf{x}}\, y](i,j)\, e^{+2\pi\iota\left(\frac{ui}{H}+\frac{vj}{W}\right)},\end{aligned} \tag{7}$$

which is the forward DFT of $\nabla_{\mathbf{x}}\, y$ up to complex conjugation. Since $\mathbf{x}$ is real, $\mathcal{F}\{\nabla_{\mathbf{x}}\, y\}$ is Hermitian-symmetric, so the magnitude is preserved under complex conjugation and we conclude

$$\big|\partial y/\partial \mathbf{X}(u,v)\big| = \big|\mathcal{F}\{\nabla_{\mathbf{x}}\, y\}(u,v)\big|. \tag{8}$$

The orthonormal convention is essential, since it places the $1/\sqrt{HW}$ factor symmetrically on the forward and inverse transforms, so that the identity holds without additional scaling. The implementation reduces to a single FFT call applied to the spatial gradient.

B.3 Hyperparameters

Table 4 lists the finetuning settings. The optimizer, learning rate, and batch size are inherited unchanged from the base detectors' own training configuration, which Corvi and Rajan share, for the same reason that the augmentation pipeline is inherited, so that the comparison against the base isolates the effect of masking. Every FAM run finetunes for a fixed 11 epochs and keeps the final checkpoint. The threshold multipliers $(k_{\text{real}}, k_{\text{fake}})$ are selected per detector by a grid search, keeping the setting with the highest balanced accuracy on the validation split on which the base detectors select their own checkpoints. The grids are $k_{\text{real}} \in \{2.5, 2.75, 3.0\} \times k_{\text{fake}} \in \{5.5, 6.0, 6.5, 7.0\}$ for Corvi and $k_{\text{real}} \in \{6.0, 6.5, 7.0\} \times k_{\text{fake}} \in \{6.0, 6.5, 7.0, 7.5\}$ for Rajan. The unmasked-finetuning control of Table 2 of the main paper uses the same optimizer, learning rate, batch size, and training data as the corresponding FAM run, and differs only in that the inputs are left unmasked.

Table 4: **Finetuning settings for FAM.** The optimizer, learning rate, and batch size are inherited unchanged from each base detector's own training configuration. The last five rows are common to both variants.

Setting	Corvi$^{\mathcal{F}}$	Rajan$^{\mathcal{F}}$
k_{real}	3.0	6.5
k_{fake}	5.5	7.5
Optimizer	Adam	Adam
Learning rate	10^{-5}	10^{-5}
Batch size	64	64
Loss	Binary cross-entropy	
Random seed	0	
DC preservation	Yes	
Epochs	11 (final checkpoint)	
$(k_{\text{real}}, k_{\text{fake}})$ selection	Validation B. Acc	

Training augmentation. Each base detector is finetuned under its own original augmentation pipeline, so the comparison against the base isolates the effect of masking rather than of a changed input distribution. For the Corvi- and Rajan-style pipelines this is the augmentation of Corvi et al. (2023), applied in the following order to the masked reconstruction $\tilde{\mathbf{x}}$: random resized crop to 256×256 (scale $[0.08, 1]$, aspect ratio $[0.75, 1.33]$, probability 0.2); color jitter (probability 0.8); grayscale conversion (probability 0.2); cutout and additive Gaussian noise (probability 0.2 each); Gaussian blur with $\sigma \in [0, 3]$ (probability 0.5); JPEG compression with quality sampled from $[30, 100]$, through either the OpenCV or the PIL encoder (probability 0.5); a random multiple of 90° rotation; and a random horizontal flip (probability 0.5). The result is finally cropped to 96×96 for training, while evaluation runs at the native resolution of each test image. Compression and blur therefore appear in both base detectors' training distributions, which is what makes the post-processed split of Section 4.2 a test of residual rather than untrained robustness.

B.4 Sensitivity to the Threshold Multipliers

The multipliers $(k_{\text{real}}, k_{\text{fake}})$ control the fraction of the input spectrum that is removed, with smaller values inducing more aggressive masking. Figure 6 shows how the behavior of Corvi$^{\mathcal{F}}$ and Rajan$^{\mathcal{F}}$ on the benchmark of Section 4.2 changes across the selection grids of Appendix B.3. The figure reports test results for every cell only to show this sensitivity; the outlined cells are the settings selected on the validation split, and no test result entered that choice.

Two things follow. First, the gain over the base is consistent across the range examined. For both detectors, all twelve combinations exceed the base as well as the Stay-Positive finetuning baseline on that backbone: the lowest cell is 77.8 for Corvi$^{\mathcal{F}}$ against 72.5 (Corvi) and 73.9 (Corvi+), and 81.9 for Rajan$^{\mathcal{F}}$ against 79.0 (Rajan) and 80.7 (Rajan+). Second, on Rajan, whose grid contains the symmetric settings, the best asymmetric combination (89.1) exceeds every symmetric one (at most 87.0), so giving the two classes separate multipliers is more effective than masking both with a single value. Real and fake inputs carry their shortcut evidence in different bands (Section 3.2), and the two therefore require different masking strengths.

B.5 Ablation Controls

Table 2 of the main paper compares FAM against four controls. All of them share the finetuning schedule, optimizer, and augmentation of the corresponding FAM run and differ only in how the training input is produced.

- *Unmasked finetuning* feeds the unmodified image, so the arm measures what the extra finetuning alone buys.
- *Spatial Attribution Mask* replaces the frequency attribution with the channel-averaged absolute input gradient $\mathbf{A}_{\text{spat}}(i, j) = \frac{1}{C} \sum_c |\nabla_{\mathbf{x}_c} y|$, applies the same per-sample cutoff

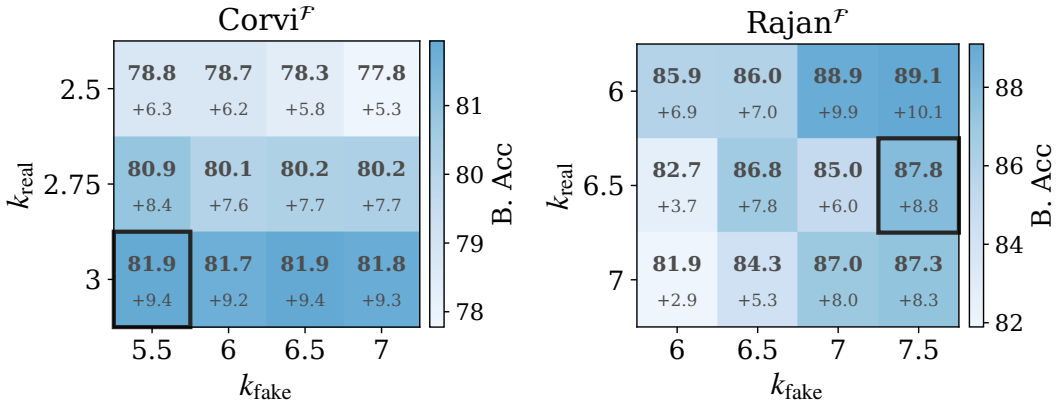


Figure 6: **How benchmark accuracy varies with the threshold multipliers.** Average balanced accuracy (B. Acc) over the nine post-processed test generators of Table 1 of the main paper for every setting of the selection grids of Appendix B.3, each finetuned for the same 11 epochs: Corvi$^{\mathcal{F}}$ (left) and Rajan$^{\mathcal{F}}$ (right). The smaller number in each cell is the gain over the base detector (Corvi 72.5, Rajan 79.0). The outlined cells are the settings selected on the validation split (Table 4). Every combination stays above its base and above Stay-Positive (Corvi+ 73.9, Rajan+ 80.7).

$\tau = \mu_+ + k\sigma_+$ to it, and zeroes the selected *pixels* instead of frequency bins. The DC coefficient has no spatial analogue, so no position is protected.

- *Random Frequency Mask* keeps the frequency domain but replaces the attribution ranking with a uniform draw over bins, preserving the DC coefficient as FAM does.
- *Random Spatial Mask* draws pixels uniformly at random.

Both random masks remove exactly as many entries as the corresponding FAM mask for that image and class, so the comparison holds sparsity fixed and isolates the value of the attribution ranking. Sparsity is determined by $(k_{\text{real}}, k_{\text{fake}})$ and by the shape of the attribution distribution.

B.6 Implementation Notes

Three details of the implementation are not visible in Equation (5) of the main paper and are recorded here for exact reproduction.

Gradients are taken with respect to the normalized input. The detectors consume ImageNet-normalized tensors, so the spatial gradient in Equation (3) of the main paper is $\partial y/\partial \mathbf{x}_{\text{norm}}$ rather than $\partial y/\partial \mathbf{x}_{\text{raw}}$. The two differ by a per-channel constant $1/\sigma_c$, which rescales the channel contributions to $\mathbf{A}_{\text{freq}}$ but leaves the ranking within a channel unchanged. Masking is unaffected: because normalization is an affine per-channel map and Equation (4) of the main paper preserves the DC coefficient, masking in the normalized domain and inverting yields exactly the same image as masking the raw spectrum.

The reconstruction is clamped and quantized. After the inverse DFT, $\tilde{\mathbf{x}}$ is clamped to $[0, 1]$ and written as an 8-bit image before augmentation. The clamp is the only non-linear step in the pipeline and affects the small number of pixels that the masking pushes outside the valid range.

The attribution target is the raw logit. We backpropagate $\sum_b y_b$ rather than $\sum_b |y_b|$, so the gradient always points in the "more fake" direction, including on real inputs. On a real image $\mathbf{A}_{\text{freq}}$ therefore measures which frequencies the detector would have to change to be pushed toward the fake side. This is a deliberate choice: it gives both classes a common reference direction, which is what makes the class-asymmetric multipliers of Section 3.2 comparable.

B.7 Computational Resources

All experiments were run on a single NVIDIA H200 NVL GPU (141 GB HBM3e, driver 580.105.08) hosted on a node with two AMD EPYC 9355 32-core CPUs and 512 GB of system memory, running Ubuntu 24.04.4 LTS with Linux kernel 6.8. The implementation uses PyTorch 2.9.1 built against CUDA 12.8, together with torchvision 0.24.1, NumPy 1.25.2, SciPy 1.9.3, scikit-learn 1.1.3, OpenCV 4.6.0.66, albumentations 1.3.1, and Captum 0.8.0 on Python 3.11. Experiment configuration is managed with Hydra 1.3.2. The approximate execution times for each component are as follows:

- Attribution computation, Stage 1 (one-time, per base detector, full training set): $\sim$ 1 hour. It is reused unchanged by every $(k_{\text{real}}, k_{\text{fake}})$ setting of that detector.
- FAM finetuning, Stage 2 (per detector, per multiplier setting): $\sim$ 2 hours.

These times are per model training instance and may vary slightly depending on the specific environment.

C Experimental Setup Details

C.1 Datasets

We describe the datasets used for training and evaluation in our experiments.

Training Data. Following the established protocol of prior work (Corvi et al., 2023; Rajan et al., 2025; Rajan & Lee, 2025), we use real images from LSUN (Yu et al., 2015) and COCO (Lin et al., 2014), totaling 90k images per source. The two base detectors use two distinct generation strategies that match their original training pipelines:

- **Corvi-style training**: 180k LDM (Rombach et al., 2022) fakes generated by text prompts that correspond to the semantic content of the real images.
- **Rajan-style training**: 180k VAE encoder–decoder reconstructions of the same real images, providing pixel-level alignment between real and fake training pairs (the AlignedForensics fix).

FAM attribution is precomputed once on the full training set for each base detector, using that detector's frozen pretrained checkpoint, and cached. The thresholding, masking, and inverse DFT that turn it into a masked reconstruction are then applied per sample during finetuning (Appendix B.1).

Evaluation Data. For evaluation, we use real images from the Redcaps dataset (Desai et al., 2021), which contains diverse web-crawled images with natural variations in content, resolution, and compression. The synthetic evaluation set spans two regimes.

Generator-specific evaluation. We evaluate on images from nine generators that span the dominant generation paradigms:

- **Diffusion-based models**: Stable Diffusion (SD) (Rombach et al., 2022), FLUX (Labs, 2024), Kandinsky (Razzhigaev et al., 2023), Playground (Li et al., 2024), and PixArt (Chen et al., 2023).
- **Latent consistency models**: LCM (Luo et al., 2023), which enables fast sampling through consistency distillation.
- **Alternative architectures**: Wuerstchen (Pernias et al., 2023), employing a two-stage latent diffusion approach, and aMUSEd (Patil et al., 2024), an autoregressive masked image model.
- **Commercial systems**: Midjourney (Midjourney, 2023), a proprietary text-to-image service known for high aesthetic quality.

None of these generators is seen during the training of any base detector.

In-the-wild evaluation. We additionally evaluate on images collected in the wild, from three sources:

- **Chameleon** (Yan et al., 2024): A curated collection of challenging AI-generated images from online AI-art communities, in which the synthetic samples are visually difficult for humans.
- **SynthWildX** (Cozzolino et al., 2024): Text-to-image fake images posted on the open web. We report the mean across the DALL·E 3, Firefly, and Midjourney v5 splits.
- **AIGIBench** (Li et al., 2025b), open-platform subsets: *SocialRF*, fakes harvested from social-media platforms (X, Facebook, Reddit) using AI-related hashtags; and *CommunityAI*, fakes harvested from AI-art communities (ArtStation, Civitai, Liblib). Real images for each subset are sampled in equal numbers from FFHQ (Karras et al., 2019), CelebA-HQ (Karras et al., 2018), and Open Images V7 (Kuznetsova et al., 2020), yielding a one-to-one balance between real and fake. Of the AIGIBench groups, only these two consist of fakes harvested in the wild; the other groups contain images produced by running specific generators, so we do not include them in the in-the-wild evaluation.

Chameleon and SynthWildX images have additionally undergone unknown post-processing during upload and sharing.

C.2 Baseline Methods

We compare against a comprehensive set of state-of-the-art AI-generated image detectors. Below we provide a brief description of each baseline, categorized by their underlying architecture and approach. For all baselines we follow the official implementations and pretrained checkpoints released by the authors, and evaluate every detector under the identical evaluation protocol described in Section 4.

CNN-based Detectors.

- **Corvi** (Corvi et al., 2023): A ResNet-50 detector trained on real images from LSUN and COCO paired with LDM-generated fakes from corresponding text prompts. The first downsampling operation is removed to preserve low-level forensic traces, and aggressive data augmentation is applied during training.
- **Rajan** (Rajan et al., 2025): Uses the same ResNet-50 architecture and real images as Corvi, but trains on VAE reconstructions of the real images rather than prompt-based generation. This alignment ensures that real and fake training pairs differ only in decoder artifacts.
- **Corvi+ / Rajan+** (Rajan & Lee, 2025): Apply the Stay-Positive algorithm to the Corvi and Rajan detectors. The backbone is frozen and the final linear layer is retrained with non-negative weight constraints, so that the detector cannot use the absence of generative artifacts as positive evidence of realness. These are our most direct comparison points among existing source-level fixes.
- **NPR** (Tan et al., 2024b): Feeds a ResNet-50 classifier with a residual image computed as the difference between the original image and its bilinearly interpolated version, exploiting upsampling artifacts that recur across generators.
- **SAFE** (Li et al., 2025a): A lightweight ResNet-50 detector that replaces downsampling with crop operations, applies ColorJitter and RandomRotation augmentations, and uses patch-based random masking to enforce local awareness.
- **AIDE** (Yan et al., 2024): A ConvNeXt-based detector that combines multiple experts to extract visual artifacts and noise patterns. The same paper introduces the Chameleon benchmark.
- **FerretNet** (Liang et al., 2025): A lightweight detector with depthwise separable and dilated convolutions that operates on median-filter residuals.
- **LaDeDa** (Cavia et al., 2024): A locally aware deepfake detector that scores 9×9 patches and pools the patch-level scores to produce the final image score.
- **DRCT** (Chen et al., 2024): A ConvNeXt detector trained with diffusion-reconstruction contrastive learning on semantically aligned training pairs reconstructed via DDIM inversion.

ViT-based Detectors.

- **CoDE** (Baraldi et al., 2024): A ViT trained with contrastive embeddings that capture global and local image similarities.
- **SPAI** (Karageorgiou et al., 2025): A spectral learning approach that detects AI-generated images at arbitrary resolutions by extracting frequency-domain features.

CLIP-based Detectors.

- **UFD** (Ojha et al., 2023): A linear probe on the frozen CLIP ViT-L/14 vision encoder.
- **C2P-CLIP** (Tan et al., 2025): Injects category-common prompts into the CLIP framework while keeping the vision encoder frozen.
- **RINE** (Koutlis & Papadopoulos, 2024): A trainable importance estimator that aggregates representations from intermediate transformer blocks of a frozen CLIP encoder.
- **ClipDet** (Cozzolino et al., 2024): An ensemble that aggregates predictions from multiple CLIP backbones pretrained on different datasets.

Training-free Detectors.

- **AEROBLADE** (Ricker et al., 2024): Detects AI-generated images by measuring the LPIPS reconstruction error after passing each image through pretrained LDM autoencoders, without any detector training.
- **WaRPAD** (Choi et al., 2025): Measures the sensitivity of DINOv2 features to high-frequency Haar-wavelet perturbations.

For training-free methods we report only AUC in Table 1 of the main paper, since they do not expose a calibrated decision threshold.

C.3 Post-Processing Protocol

For the post-processed evaluation reported in Section 4.2, each test image is independently subjected to JPEG compression (quality factor sampled from $\{50, 65, 75, 90\}$), image resizing (scale sampled uniformly from $[0.5, 1.0]$), and color jitter on brightness, contrast, and saturation (each sampled uniformly from $[-0.2, 0.2]$), applied in random order. This protocol matches Rajan & Lee (2025) and approximates the kind of degradation that real images undergo during a single round of upload and sharing.

D Additional Results

We provide full per-method results on the in-the-wild benchmarks, together with qualitative visualizations of the FAM pipeline on training images. The per-operation robustness breakdown and the score distributions appear in the main paper.

D.1 In-the-Wild Benchmarks

Table 5 reports the full per-method results on the four in-the-wild test sets: Chameleon, SynthWildX, and the SocialRF and CommunityAI subsets of AIGIBench. The summary of the trends is provided in Section 4.3.

D.2 Extended Visualization

We provide two additional sets of qualitative results to complement the quantitative analysis in the main paper.

Figure 7 compares mean frequency attribution before and after FAM finetuning for both real and fake inputs. In the Corvi panel, attribution mass increases over the inner and middle frequencies and decreases over the outer, high-frequency region. The Rajan panel reverses this, with mass leaving

Table 5: **Detection performance on the in-the-wild benchmarks.** Methods are grouped as in Table 1 of the main paper. *Chameleon* (Yan et al., 2024): a curated collection of challenging AI-generated images from online art communities. *SynthWildX* (Cozzolino et al., 2024): text-to-image fakes posted on the open web. We report the mean across the DALL·E 3, Firefly, and Midjourney v5 splits. *SocialRF* and *CommunityAI*: the two open-platform subsets of AIGIBench (Li et al., 2025b), with fakes harvested from social-media platforms and from AI-art communities, respectively. We report AUC and balanced accuracy (B. Acc $= (\mathrm{R.Acc}+\mathrm{F.Acc})/2$). Bold = best, underline = second best. FAM-finetuned variants are highlighted in blue.

	Detector	Chameleon		SynthWildX		SocialRF		CommunityAI	
		AUC	B. Acc	AUC	B. Acc	AUC	B. Acc	AUC	B. Acc
CLIP	UFD	55.4	50.0	46.1	50.0	53.4	50.0	58.5	50.0
	C2P-CLIP	44.2	51.1	68.4	57.1	63.0	56.4	47.1	51.0
	RINE	39.2	42.1	69.2	57.9	61.3	56.9	35.6	40.2
	ClipDet	52.7	51.5	81.7	68.4	81.8	69.7	67.9	59.2
V	CoDE	81.6	73.9	80.7	71.9	65.6	61.9	70.6	63.4
TF	AEROBLADE	75.4	–	81.9	–	69.1	–	**87.5**	–
	WaRPAD	55.2	–	66.1	–	60.0	–	60.9	–
CNN	SAFE	57.1	52.5	49.5	50.0	60.9	58.4	57.9	54.5
	FerretNet	42.9	52.5	47.4	49.9	56.4	50.6	43.6	49.9
	LaDeDa	51.0	52.7	58.3	50.9	53.0	50.1	49.5	50.0
	NPR	52.2	53.4	51.0	49.9	58.0	54.8	47.7	55.2
	Corvi	73.3	59.3	89.3	76.0	73.9	61.1	68.2	53.3
	Corvi+	**83.6**	59.1	88.6	73.0	78.8	62.3	70.4	52.7
	Corvi$^{\mathcal{F}}$	82.7	**75.3**	92.9	**83.6**	79.1	**72.2**	81.6	**75.3**
	Rajan	80.0	63.0	85.3	77.9	77.3	65.7	79.2	55.3
	Rajan+	75.4	64.7	**93.9**	81.8	79.7	66.6	66.9	57.8
	Rajan$^{\mathcal{F}}$	78.9	67.4	88.9	80.5	**82.0**	72.1	79.1	60.6

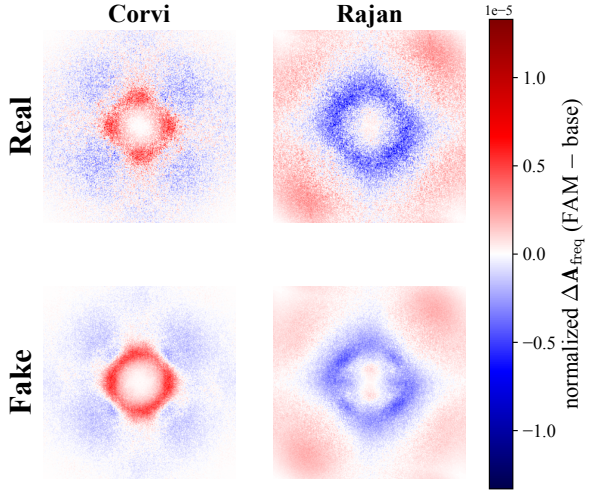


Figure 7: **Attribution shift after FAM finetuning.** Each panel shows the signed difference $\Delta\mathbf{A}_{\mathrm{freq}} = \mathbf{A}_{\mathrm{freq}}^{FAM} - \mathbf{A}_{\mathrm{freq}}^{\mathrm{base}}$ between mean frequency-domain attributions, with the DC coefficient shifted to the center. Top row uses Redcaps real images, bottom row uses fake images averaged over the seven post-processed test generators (100 samples per generator). Red marks frequencies where attribution mass increases after finetuning, blue marks frequencies where it decreases.

a middle annulus and accumulating toward the outer corners of the spectrum. The reversal reflects the different starting point of the Rajan base detector. In each case FAM takes attribution mass away from the region the detector already leans on and redistributes it onto the complementary bands, so

the direction of the shift follows the dominant region rather than a fixed band. The pattern is also consistent across the real and fake rows, indicating that FAM redistributes attribution mass across both input classes.

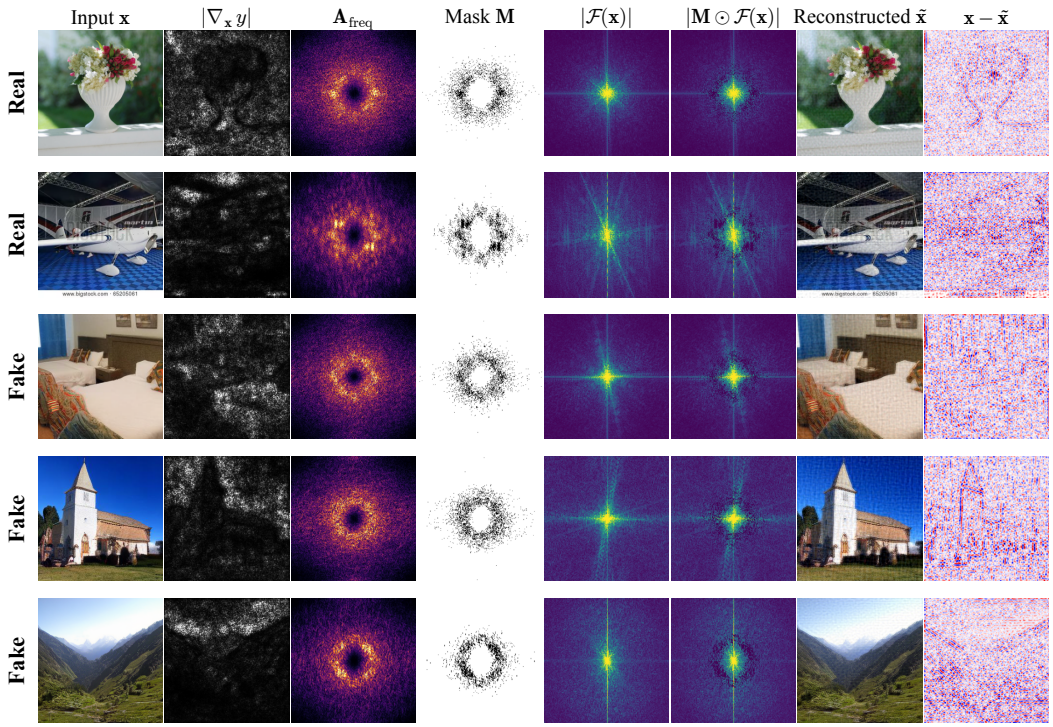


Figure 8: **Step-by-step visualization of the FAM masking pipeline on five training images** (two real, three fake), with attribution computed from the Corvi backbone. From left to right: original image $\boldsymbol{x}$, channel-mean spatial gradient $|\nabla_{\boldsymbol{x}} y|$, frequency-domain attribution $\mathbf{A}_{\text{freq}}$, binary mask $\boldsymbol{M}$ with the DC coefficient preserved, input spectrum $|\mathcal{F}(\boldsymbol{x})|$, masked spectrum $|\boldsymbol{M} \odot \mathcal{F}(\boldsymbol{x})|$, reconstructed image $\tilde{\boldsymbol{x}} = \mathcal{F}^{-1}(\boldsymbol{M} \odot \mathcal{F}(\boldsymbol{x}))$, and spatial residual $\boldsymbol{x} - \tilde{\boldsymbol{x}}$. Across all rows, $\tilde{\boldsymbol{x}}$ is visually almost indistinguishable from $\boldsymbol{x}$ even though the residual reveals localized perturbations where the detector's dominant frequency cues concentrate.

Figure 8 traces the FAM pipeline of Section 3 step by step on representative training images, illustrating how the per-sample frequency-domain attribution maps into a band-localized binary mask whose pixel-space inversion changes the original image only minimally.

E MoRF Faithfulness Analysis

We give the full protocol for the MoRF diagnostic of Section 4.4 and Figure 3 of the main paper, explain how the resulting curve should be read, and repeat the sweep on post-processed inputs.

E.1 Diagnostic Protocol

For a given detector f_θ and an input image $\boldsymbol{x}$, we (i) compute the per-sample frequency attribution $\mathbf{A}_{\text{freq}}$ (Eq. 3 of the main paper), (ii) rank the bins by $\mathbf{A}_{\text{freq}}$ in descending order, (iii) zero out the top-$K\%$ of bins on the input spectrum $\mathcal{F}(\boldsymbol{x})$ while preserving the DC coefficient, (iv) reconstruct the masked input via the inverse DFT, and (v) re-evaluate f_θ on the masked input. A bin here is a single entry (u, v) of the 256×256 frequency grid, so $K = 1\%$ corresponds to roughly 650 of the 65,536 coefficients. We sweep $K \in \{0.5, 1, 2, 5, 10, 20\}\%$ together with the unperturbed $K = 0$ reference, and report AUC against the held-out Redcaps real set on 80 fake images per generator at 256×256 resolution. A faithful attribution should produce a sharp drop in AUC at small K, since removing the bins the detector is most sensitive to should immediately destroy its decision signal.

E.2 Reading the MoRF Curve

The sweep is a test of $\mathbf{A}_{\text{freq}}$ itself. If the ranking carried no information, removing the bins it places highest would change little. Instead, deleting the top 1% collapses Corvi$^{\mathcal{F}}$ from an AUC of 0.99 to 0.64 (Figure 3 of the main paper). The ranking therefore singles out the coefficients that carry the decision, and what $\mathbf{A}_{\text{freq}}$ shows can be read as a property of the detector rather than an artifact of the attribution.

What finetuning changed is visible from two sides. The attribution-shift maps of Appendix D.2 show *where* the support moved: mass leaves the region the base detector leaned on and accumulates on the complementary bands. The MoRF curves show *how concentrated* the support is. Under the same top-1% removal, with bins ranked by each detector's own attribution, Corvi base slips only from 0.97 to 0.94, its decision spread across many individually dispensable bins, while Corvi$^{\mathcal{F}}$ collapses because its decision rides on the far smaller set that was just removed. FAM relocates the decision support and concentrates it.

A support spread over many bins does not protect the base. Corvi's AUC on `FLUX` still falls sharply once the inputs are post-processed (Figure 1 of the main paper), since the deployment operations act on whole bands rather than on individual bins (Section 2.2 of the main paper).

The concentrated support, by contrast, survives perturbation and transfers. When the sweep is repeated on JPEG-compressed and downsized inputs the sharp drop is unchanged (Appendix E.3), so the decision remains concentrated under the operations themselves. Both FAM variants improve over their base on both AUC and balanced accuracy under the full post-processing protocol (Table 1 of the main paper) and, on the four in-the-wild test sets, improve balanced accuracy on every set and AUC on all but two (Appendix D.1). The robustness reported there thus rests on a support that stays concentrated under the very operations that erase the base detector's cues.

E.3 Faithfulness Under Post-Processing

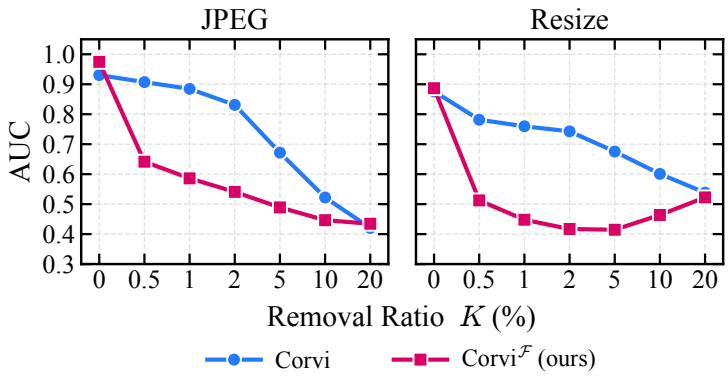


Figure 9: **MoRF faithfulness after post-processing.** AUC on 80 `FLUX` images against the Redcaps real set as the top-$K\%$ of bins ranked by $\mathbf{A}_{\text{freq}}$ are removed, with the input JPEG-compressed (left) or downsized (right) beforehand. The pattern of Figure 3 of the main paper holds under both operations.

The curve of Figure 3 of the main paper is measured on clean inputs, which leaves open whether the concentration it reveals is a property of the finetuned detector or an artifact of undegraded images. Figure 9 repeats the sweep on inputs that have already been JPEG-compressed and on inputs that have already been downsized, the two operations of the post-processing protocol of Appendix C.3 that redistribute energy across the spectrum.

The pattern is unchanged. Under JPEG, Corvi$^{\mathcal{F}}$ falls from AUC 0.98 to 0.64 once the first 0.5% of bins are removed, while Corvi base moves from 0.93 to 0.91 over the same range. Under downsizing the corresponding figures are 0.89 to 0.51 for Corvi$^{\mathcal{F}}$ and 0.89 to 0.78 for Corvi base. The

concentration of the decision support is therefore carried by the detector rather than by the condition of the input.