SPECTRAL TAIL AUXILIARY LEARNING FOR AI-GENERATED IMAGE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As generative image models evolve rapidly, the perceptual gap between generated and real images continues to narrow, making AI-generated image detection increasingly challenging. Many existing methods exploit frequency-domain cues for detection, typically described as frequency-domain artifacts or high-frequency discrepancies. However, the specific and recurring spectral regularities remain insufficiently understood and characterized. In this paper, we systematically analyze the one-dimensional radial log-power spectra of real and generated images. We find that generated images do not necessarily exhibit higher or lower energy across the entire spectrum or high-band range. Instead, their spectra deviate from the power-law decay and show an anomalous uplift in the ultra-high-frequency tail. We term this phenomenon *spectral tail uplift*. We further explain this phenomenon through nonlinear harmonic accumulation in trained generative models, suggesting that it can serve as a structural cue across generative architectures. Based on this observation, we propose Spectral Tail Auxiliary Learning (STAL), a frequency-domain auxiliary supervision framework for generalizable AI-generated image detection. STAL transfers spectral-tail cues from a tail-aware frequency teacher to a spatial detector during training, while all frequency-domain modules are discarded at inference time. Consequently, STAL introduces no inference overhead. Extensive experiments on 9 public datasets show that STAL achieves strong generalization and stability across generators, data distributions, and real-world scenarios.

1 INTRODUCTION

As generative image models continue to advance (Goodfellow et al., 2014; Rombach et al., 2022; Black Forest Labs, 2024; Midjourney, Inc.), AI-generated images have become increasingly photorealistic, raising concerns about misuse and creating an need for detectors that can generalize across generators and remain reliable after the compression, resizing, and re-encoding that images undergo in the wild. Frequency-domain cues have long been regarded as a promising route toward both goals, since generative pipelines are believed to leave traces in the spectrum that are invisible to the eye (Zhang et al., 2019; Frank et al., 2020; Chu et al., 2025; Li et al., 2024).

However, existing frequency-based detectors are limited in the following respects. First, the cue itself is loosely defined, often described as frequency-domain artifacts or high-frequency anomalies (Zhang et al., 2019; Luo et al., 2024), yet such cues do not exhibit a consistent pattern across generators. Taking high-frequency energy as an example, as shown in Fig. 1, some generated images exhibit excessive high-frequency energy, while others show insufficient energy. Detectors that rely on such cues, therefore, struggle to achieve strong generalization across generators. Meanwhile, existing explanations for the origins of these artifacts are largely architecture-specific, focusing mainly on GANs, and failing to provide mechanism applicable to diffusion models and other modern generators. Moreover, most methods directly use spectral features at inference time, making them sensitive to perturbations. These limitations raise a more fundamental question: *is there a spectral regularity that recurs across generators, can be explained by a common mechanism, and can be exploited without relying on fragile frequency-domain inputs at inference time?*

To answer this question, we systematically analyze the one-dimensional radial log-power spectra of real and generated images. We observe a phenomenon that consistently emerges across different

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.

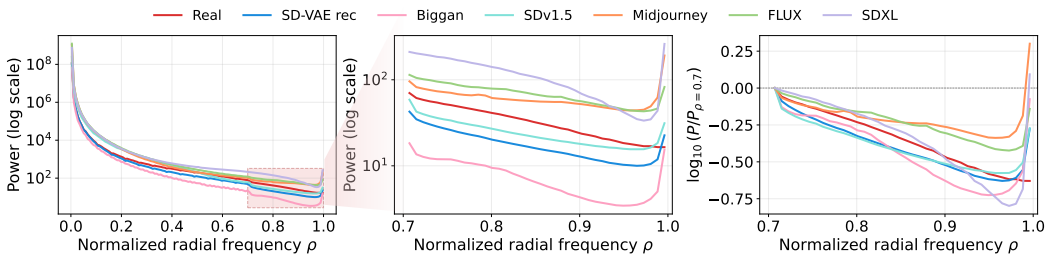


Figure 1: Radial FFT (Cooley & Tukey, 1965) power spectra of real and fake images. **Left:** spectra over the full radial frequency range. **Middle:** spectral tail over the local frequency range $\rho \in [0.7, 1]$. **Right:** normalized tail curves anchored at $\rho = 0.7$ to expose shape differences. Across generators, fakes show consistent spectral-tail deviations, revealing spectral tail uplift across architectures. The middle and right panels are included to help to understand and are not part of its definition.

generative architectures. As shown in Fig. 1, while real images typically follow an approximate power-law decay (Field, 1987; Simoncelli & Olshausen, 2001) in the radial spectrum, generated images deviate from this behavior in the ultra-high-frequency tail, exhibiting an anomalous upward shift. We term this phenomenon *spectral tail uplift*. Unlike spectral energy discrepancies, spectral tail uplift is a localized morphological cue that appears consistently across a wide range of GANs (Brock et al., 2019), diffusion models (Rombach et al., 2022; Midjourney, Inc.), autoregressive models (Han et al., 2025), and reconstructions from different VAEs (Rombach et al., 2022).

Controlled experiments further show that spectral tail uplift is associated with nonlinear activations in trained generative models. Removing nonlinear activations eliminates the uplift, while replacing trained weights with random initialization substantially weakens it. We explain this behavior from the perspective of harmonic accumulation. Pointwise nonlinear activations generate new harmonic components, while the learned parameters preserve and propagate them through the generative cascade. Because nonlinear activations and learned weights are shared by virtually all mainstream generative models, this mechanism provides an architecture-agnostic explanation for spectral tail uplift, suggesting that it is a mechanistically grounded cross-generator characteristic rather than a generator- or dataset-specific artifact.

Based on this observation, we propose **S**pectral **T**ail **A**uxiliary **L**earning (**STAL**), a frequency-domain auxiliary supervision framework for generated-image detection. STAL leverages spectral tail uplift as a discriminative signal through training-time auxiliary supervision, enabling the detector to benefit from spectral-tail cues while maintaining robustness to post-processing operations at inference time. During training, STAL constructs a tail-aware frequency teacher and transfers spectral-tail cues to the representation of a spatial detector through auxiliary supervision. At inference time, all frequency-domain modules are discarded, and only the spatial detector is retained. In this way, STAL leverages tail uplift to improve detector generalization without introducing any additional inference cost. Our main contributions are summarized as follows:

- We identify *spectral tail uplift*, a local spectral regularity in the ultra-high-frequency tail that recurs consistently across GANs, diffusion models, autoregressive models and VAEs.
- We conduct a systematic analysis of spectral tail uplift and provide a cross-architecture explanation from the perspective of harmonic accumulation induced by nonlinear activations in trained generative models.
- We propose Spectral Tail Auxiliary Learning, a auxiliary supervision framework based on spectral tail uplift, and validate its effectiveness on multiple public datasets, demonstrating significant improvements over state-of-the-art methods and greater robustness to post-processing.

2 Related Works

2.1 Generative Image Models

Image synthesis technologies (Goodfellow et al., 2014; Ho et al., 2020; Rombach et al., 2022; Midjourney, Inc.; Podell et al., 2024) have advanced rapidly over the past few years. Early generative

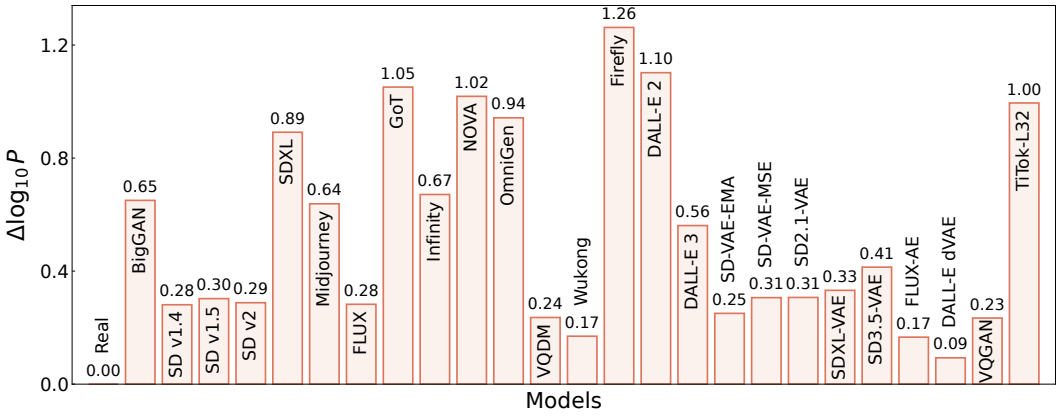


Figure 2: Spectral tail uplift across diverse generative architectures. Real images exhibit essentially no uplift, whereas images from a wide range of generators consistently show positive uplift.

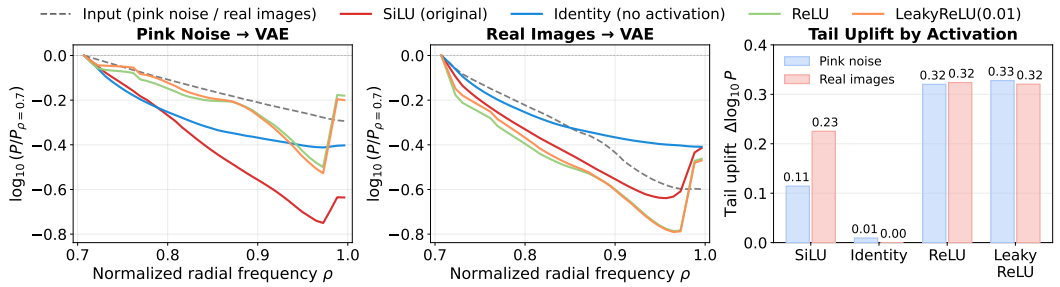


Figure 3: Activation nonlinearity is necessary for the spectral tail uplift. We replace SiLU in SD-VAE with Identity, ReLU, or LeakyReLU, then pass pink noise (**left**) and real images (**middle**) through the modified VAE. Normalized Curves show spectra on $\rho \in [0.7, 1]$. **Right:** tail uplift $\Delta \log_{10} P$, the rise from the tail's minimum to $\rho = 1$. Identity collapses the uplift to ≈ 0. ReLU and LeakyReLU strengthen it.

adversarial networks (GANs) (Goodfellow et al., 2014; Karras et al., 2019; Zhu et al., 2017; Choi et al., 2018) were already capable of producing photorealistic images, but they still suffered from issues such as mode collapse and training instability. Subsequently, Ho et al. (2020) proposed Denoising Diffusion Probabilistic Models (DDPMs), which have since become the dominant paradigm in this field. Diffusion models generate high-quality images by simulating a thermodynamic diffusion process, starting from Gaussian noise and progressively recovering images through iterative denoising. Rombach et al. (2022) further introduced Latent Diffusion Models (LDMs), which map the diffusion process into a low-dimensional latent space and employ a variational autoencoder (VAE) (Kingma & Welling, 2014) as the encoder and decoder between pixel space and latent space, thereby substantially reducing computational costs while preserving generation quality. Many mainstream generative models (Podell et al., 2024; Esser et al., 2024; Black Forest Labs, 2024) developed thereafter have also adopted this paradigm, using VAEs for encoding and decoding.

2.2 Frequency-based Generated Image Detection

Many studies (Frank et al., 2020; Tan et al., 2024a; Liu et al., 2024; Luo et al., 2024; Yan et al., 2025) have explored frequency-domain cues for AI-generated image detection. Zhang et al. (2019) showed that up-sampling operations in GANs introduce periodic artifacts in the frequency domain. Chandrasegaran et al. (2021) investigated high-frequency spectral decay discrepancies in GAN-generated images and showed that they can be largely eliminated by replacing zero-insertion-based scaling in the generator's final upsampling stage, thereby yielding spectral decay that is more consistent with that of real images. However, we study a localized non-monotonic upturn at the extreme end of the spectrum, associated with the accumulation of harmonics induced by nonlinear activations in trained generators, and remains observable across a variety of generative models whose final upsampling stage does not employ zero-insertion-based scaling. Some works exploit frequency information

directly or jointly with spatial features. FreqNet (Tan et al., 2024a) learns in frequency space to improve generalization, while MaskSim (Li et al., 2024) learns discriminative spectral masks. Luo et al. (2024) leverages high-frequency reconstruction errors, and FIRE (Chu et al., 2025) focuses on mid-frequency components that generators struggle to reconstruct. Other methods combine frequency cues with pretrained CLIP (Radford et al., 2021) features (Yan et al., 2025; Liu et al., 2024). Although prior works leverage frequency cues in various ways, they mostly rely on spectral discrepancies whose patterns vary across generators. A recurring cross-generator spectral regularity with a common mechanism therefore remains insufficiently characterized and exploited.

3 METHOD

3.1 SPECTRAL TAIL UPLIFT

We compare the azimuthally averaged 1D radial power spectra $S(\rho)$ of real and generated images, where $\rho \in [0, 1]$ denotes the normalized radial frequency. As shown in Fig. 1, real-image spectra approximately follow a power-law decay,

$$S(\rho) \propto \rho^{-\alpha}, \tag{1}$$

and decrease monotonically up to $\rho = 1$. Generated images exhibit similar behavior at low and mid frequencies, but show a characteristic decrease-to-increase transition in the ultra-high-frequency tail. We refer to this pattern as *spectral tail uplift*, which definition is as follow: given ρ_0, S exhibits spectral tail uplift if there exists $\rho \in [\rho_0, 1)$ such that $S(\rho)$ decreases on $[\rho_0, \rho]$ and increases on $[\rho^*, 1]$. We quantify its strength by

$$\Delta = \log_{10} S(1) - \min_{\rho \geq \rho_0} \log_{10} S(\rho), \tag{2}$$

where $\Delta = 0$ for a monotonically decreasing spectrum and $\Delta > 0$ when an uplift is present. We use Δ only for quantitative comparisons in Figs. 2, Figs. 3 and Appendix A. As shown in Fig. 2, this pattern recurs across GANs (Brock et al., 2019), diffusion models (Rombach et al., 2022; Midjourney, Inc.), autoregressive models (Han et al., 2025), and reconstructions from different VAEs (Rombach et al., 2022), while real images do not exist.

3.2 ANALYSIS

The observation above raises a natural question: where does the extra energy at the extreme high-frequency tail come from? We abstract a typical generative decoder as an L-layer cascade, in which each layer applies a trained convolution followed by a pointwise nonlinear activation,

$$x_\ell \;=\; \phi(H_\ell * x_{\ell-1})\,, \qquad \ell = 1, \ldots, L, \tag{3}$$

where H_ℓ denotes the convolution at layer ℓ, and ϕ is a pointwise nonlinear activation. We omit attention, normalization, residual connections, and upsampling from this abstraction, which isolates the single mechanism analyzed below. Within this cascade, the linear convolution can only rescale the amplitude of frequencies already present in its input, but cannot create new frequencies. The pointwise nonlinear activation, in contrast, introduces frequency components that were absent from its input. We now analyze how the emergence of spectral tail uplift is closely related to high-frequency harmonics generated by nonlinear activations and subsequently propagated through the learned convolutions.

Harmonic generation. Consider a bandlimited input

$$x(t) \;=\; \sum_{m=-M}^{M} \hat{x}_m\, e^{imt}, \tag{4}$$

of bandwidth M, with Fourier coefficients $\hat{x}_m$ and a nonzero top component $\hat{x}_M \neq 0$. Let $\phi(z) = \sum_{q=0}^{d} a_q z^q$ be a degree-d polynomial activation with $a_d \neq 0$ and $d \geq 2$.

Theorem 1 (Harmonic generation). The highest positive frequency of $\phi(x)$ is extended from M to dM, and the Fourier coefficient at frequency dM is exactly

$$\widehat{\phi(x)}_{dM} \;=\; a_d\,(\hat{x}_M)^d. \tag{5}$$

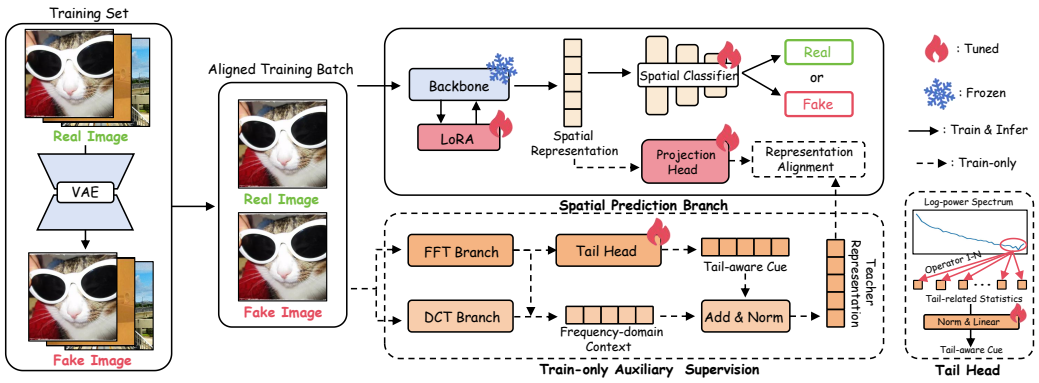


Figure 4: Overview of STAL. A tail-aware frequency teacher extracts spectral-tail cues from a frequency-preserving view and aligns them with the projected spatial feature. At inference time, all frequency modules and the projection head are discarded, leaving only the spatial detector.

This follows from the fact that x^d corresponds to a d-fold self-convolution in the Fourier domain. The full proof is given in Appendix B.1. In other words, each nonlinear activation layer introduces new frequency components beyond those present in the input, thereby extending the signal toward higher frequencies.

Natural images may also undergo nonlinear operations inside the camera image signal processing (ISP) pipeline, including gamma correction and denoising. The key difference is that these operations are constrained by optical imaging, sensor sampling, and camera processing and are not organized as a learned convolution-activation cascade. As a result, the spectra of natural images retain an approximate power-law decay. A generative decoder, by contrast, repeatedly interleaves trained convolutions with pointwise nonlinear activations. Each pointwise activation can generate new harmonic components, and the learned convolutional filters carry these harmonics toward progressively higher frequencies. This repeated synthesis provides the conditions for the sustained accumulation of high-frequency harmonics, which provides a concrete route by which extra energy reaches the spectral tail of generated images.

Harmonic propagation through the cascade. In the polynomial model above, each nonlinearity extends the accessible top frequency by a factor of d. Let the initial frequency be k_0, there exists a highest-order harmonic path that evolves as $k_0 \to dk_0 \to d^2k_0 \to \cdots \to d^Lk_0$.

Theorem 2 (Harmonic-chain propagation). For a single-tone input $x_0(t) = A\cos(k_0 t)$ of amplitude A and frequency k_0, if the filter responses along this path are nonzero, i.e., $H_\ell(d^{\ell-1}k_0) \neq 0$ for all $\ell = 1, \ldots, L$, the power at the top harmonic d^Lk_0 at layer L is

$$\left|\hat{x}_L(d^Lk_0)\right|^2 \;=\; |a_d|^{\frac{2(d^L-1)}{d-1}} \left(\frac{A}{2}\right)^{2d^L} \prod_{\ell=1}^{L} \left|H_\ell\left(d^{\ell-1}k_0\right)\right|^{2d^{L-\ell+1}}. \tag{6}$$

The proof is given in Appendix B.2. The top-harmonic power thus depends jointly on the nonlinearity strength $|a_d|$, the input amplitude A, and the filter gains $|H_\ell|$ evaluated along the harmonic chain. Taking logarithms exposes the per-layer contributions more clearly:

$$\log\left|\hat{x}_L(d^Lk_0)\right|^2 \;=\; C_0 \;+\; \sum_{\ell=1}^{L} 2\,d^{L-\ell+1}\,\log\left|H_\ell\left(d^{\ell-1}k_0\right)\right|, \tag{7}$$

where C_0 absorbs the contributions of $|a_d|$ and A. The weights $2d^{L-\ell+1}$ scale exponentially with the remaining depth, so the top-harmonic path is highly sensitive to the filter gain of every layer along the chain. The nonlinearity injects new high-frequency content at each layer, and the cascaded convolutions propagate this content along the harmonic chain toward progressively higher frequencies, where it emerges as the dominant contribution at the extreme high-frequency tail and produces the uplift we observe.

Our controlled experiments support the above analysis: tail uplift consistently appears in trained generative models, while it is almost completely eliminated by removing nonlinear activations as shown

in Fig. 3, and substantially suppressed when the model uses untrained random weights. Complete experiment results are provided in Appendix A. Since nearly all mainstream generative models, including diffusion models, GANs, and autoregressive models with visual decoders, are trained and incorporate nonlinear activations, we argue that tail uplift recurs across generators, as we verify on GAN, diffusion, autoregressive models, as well as reconstructions from different VAE/AE architectures, and serves as a generator-agnostic cue.

3.3 Spectral Tail Auxiliary Learning

We propose Spectral Tail Auxiliary Learning (STAL), as shown in Fig. 4, which transfers spectral-tail cues to a spatial detector through a training-only frequency teacher. The spatial view x_s, constructed following DDA (Chen et al., 2025), is processed by the visual encoder E_θ and classifier C_s. In parallel, a frequency-preserving view x_f shares the same geometric transformations as x_s but excludes augmentations that distort spectral structure. Its radial log-power spectra $S_{\mathrm{rad}}(x_f)$ are encoded by a tail head T_ω and a context encoder F_ψ, capturing spectral-tail uplift and global spectral context, respectively. Their representations form the teacher target t_f, which is aligned with the projected spatial representation during training. At inference, all frequency-domain modules and the projection head are removed, leaving only the spatial detector.

Tail-Aware Frequency Teacher. **(i)Tail Head.** We design a tail head T_ω to characterize the pattern of spectral tail uplift. The input to T_ω is restricted to the segment of the radial spectrum $S_{\mathrm{rad}}(x_f)$ over the normalized radial frequency range $\rho \in [0.7, 1.0]$. A predefined operator set $\mathcal{P} = \{p_1, p_2, ..., p_n\}$ is employed to extract a collection of tail-related characteristics, capturing the spectral morphology from the turning point to subsequent variations in spectral slope. The resulting representation is further processed by layer normalization and a linear projection to obtain the tail embedding e_t, while a linear classifier produces an auxiliary prediction $\hat{y}_t$:

$$e_t,\ \hat{y}_t = T_\omega\Big(S_{\mathrm{rad}}(x_f)|_{\rho\in[0.7,\,1.0]} \Big)\,. \tag{8}$$

(ii)Full-Spectrum Context. Since spectral tail uplift is a localized spectral pattern, its abnormality is more effectively characterized with reference to the overall trend across the remaining frequency bands. We therefore employ a lightweight frequency-domain encoder F_ψ to encode the complete spectrum $S_{\mathrm{rad}}(x_f)$ together with the local DCT statistics $D_{\mathrm{loc}}(x_f)$ into a contextual representation:

$$h_f = F_\psi(S_{\mathrm{rad}}(x_f), D_{\mathrm{loc}}(x_f))\,. \tag{9}$$

Here, $D_{\mathrm{loc}}(x_f)$ preserves local frequency-distribution information that may be suppressed by radial averaging, thereby complementing the global spectral trend represented by $S_{\mathrm{rad}}(x_f)$. **(iii)Teacher Target.** We combine the contextual representation h_f with the tail embedding e_t to construct the frequency-domain supervision target:

$$t_f = \mathrm{LN}(h_f + \beta e_t)\,, \tag{10}$$

where $\mathrm{LN}(\cdot)$ denotes layer normalization, and β controls the contribution of the tail embedding. The two components of the teacher thus play complementary roles. The tail head captures the localized decay profile of the spectral tail, whereas the frequency-domain encoder provides the global spectral trend against which this local pattern is interpreted. Consequently, t_f represents the shape of the spectral tail relative to the global spectral trend.

Frequency-to-spatial auxiliary learning. The frequency teacher is designed to transfer tail-aware frequency cues to the spatial representation through training-time auxiliary supervision. The main spatial branch extracts an image representation from the spatial view x_s and produces the classification prediction:

$$\mathbf{v}_s = E_\theta(x_s), \qquad \hat{y}_s = C_s(\mathbf{v}_s), \tag{11}$$

where E_θ denotes the visual encoder of the spatial detector, $\mathbf{v}_s$ is the spatial image representation, and C_s is the spatial binary classifier. To receive supervision from the frequency teacher, we use a projection head g_η to map the spatial representation into the teacher feature space:

$$\mathbf{p}_s = g_\eta(\mathbf{v}_s). \tag{12}$$

The projected spatial feature $\mathbf{p}_s$ is used only for training-time representation alignment and is discarded at inference time. We then align $\mathbf{p}_s$ with the stop-gradient frequency teacher target:

$$\mathcal{L}_{\text{align}} = \frac{1}{2} \sum_{c \in \{0,1\}} \mathbb{E}_{i:y_i=c} \left[1 - \cos(\mathbf{p}_{s,i}, \text{sg}(\mathbf{t}_{f,i})) \right]. \tag{13}$$

Here, $y_i \in \{0, 1\}$ denotes the authenticity label of the i-th sample, $\mathbb{E}_{i:y_i=c}$ denotes averaging over samples of class c in the current mini-batch, and $\text{sg}(\cdot)$ denotes stop-gradient operation. This loss computes the alignment error separately for real and generated images and assigns equal weight to the two classes, preventing the alignment from being dominated by one class. The stop-gradient operation fixes the teacher-side target and prevents the alignment loss from modifying the frequency supervisory signal. By imposing this representation constraint only during training, STAL transfers spectral-tail cues to the spatial detector without requiring any frequency-domain input or teacher branch at inference time. In addition to representation alignment, we apply auxiliary classification losses to the frequency teacher to ensure that it remains discriminative:

$$\hat{y}_f = C_f(\mathbf{h}_f), \qquad \mathcal{L}_{\text{freq}} = \text{BCE}(\hat{y}_f, y), \qquad \mathcal{L}_{\text{tail}} = \text{BCE}(\hat{y}_t, y). \tag{14}$$

Here, C_f is an auxiliary classifier applied to the frequency context representation, and $\hat{y}_f$ is the prediction from the frequency context branch. $\mathcal{L}_{\text{freq}}$ encourages $\mathbf{h}_f$ to preserve task-relevant frequency context, while $\mathcal{L}_{\text{tail}}$ encourages the tail head to learn spectral-tail structures that are useful for generated-image detection. The frequency auxiliary objective is defined as

$$\mathcal{L}_{\text{aux}} = \lambda_{\text{freq}} \mathcal{L}_{\text{freq}} + \lambda_{\text{align}} \mathcal{L}_{\text{align}} + \lambda_{\text{tail}} \mathcal{L}_{\text{tail}}, \tag{15}$$

where λ_{freq}, λ_{align}, and λ_{tail} are the weights of corresponding loss weights. The spatial branch is optimized with a classification loss and a supervised contrastive loss, following the loss configuration as DDA (Chen et al., 2025):

$$\mathcal{L}_{\text{cls}} = \text{BCE}(\hat{y}_s, y), \qquad \mathcal{L}_{\text{spatial}} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{con}} \mathcal{L}_{\text{con}}, \tag{16}$$

where $\mathcal{L}_{\text{con}}$ denotes the supervised contrastive loss, and λ_{cls} and λ_{con} are the corresponding loss weights. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{spatial}} + w_{\text{aux}}(\tau) \mathcal{L}_{\text{aux}}, \tag{17}$$

where $w_{\text{aux}}(\tau)$ is a curriculum weight that controls the strength of the frequency auxiliary terms according to the training progress τ. The frequency auxiliary supervision primarily shapes the spatial representation in the early and middle stages of training. As training proceeds, its weight gradually decreases, reducing the risk that the final detector becomes overly dependent on the training-time frequency teacher.

Inference. At inference time, we discard all frequency-domain modules, including the frequency branch, the tail head, and the projection head used for representation alignment. The final prediction is produced solely by the spatial detector:

$$\hat{y} = C_s(E_\theta(x)). \tag{18}$$

Thus, STAL uses spectral tail uplift only as a training-time auxiliary supervisory signal, introducing no frequency-domain input, teacher branch, or additional computational cost at inference time.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate STAL on nine public benchmarks, including six standard benchmarks (GenImage (Zhu et al., 2023), DRCT-2M (Chen et al., 2024), Synthbuster (Bammey, 2024), EvalGEN (Chen et al., 2025), AIGCDetectionBenchmark (Zhong et al., 2023), and ForenSynths (Wang et al., 2020)) and three in-the-wild datasets (Chameleon (Yan et al., 2025), SynthWildx (Cozzolino et al., 2024), and WildRF (Cavia et al., 2024)). These datasets cover diverse real image sources and synthetic images produced by GANs, diffusion models, and autoregressive models, while the in-the-wild datasets are collected from open-world scenarios, which involve unknown generators and post-processing operations.

Table 1: Overall comparison across 9 benchmarks. Balanced accuracy (%) is reported. DDA* denotes DDA Chen et al. (2025) with DINOv3-H+ Siméoni et al. (2025) backbone. The best and second-best results are highlighted in **bold** and underline. Numbers below each dataset name indicate the number of generators, where G, D, and AR denote GANs, diffusion models, and autoregressive models, respectively.

Method	Standard Benchmarks						In-the-Wild Datasets			Avg
	GenImage	DRCT-2M	EvalGEN	Synthbuster	ForenSynths	AIGCDetection Benchmark	Chameleon	SynthWildx	WildRF	
	1G + 7D	13D	3D + 2AR	9D	13G	7G + 10D	Unknown	3D	Unknown	
NPR (CVPR'24) Tan et al. (2024b)	51.5 ± 6.3	30.4 ± 2.7	2.9 ± 2.7	50.0 ± 2.6	47.9 ± 22.6	53.1 ± 12.2	59.9	49.8 ± 10.0	63.5 ± 13.6	45.4 ± 18.4
UnivFD (CVPR'23) Ojha et al. (2023)	64.1 ± 10.8	62.6 ± 9.5	15.4 ± 14.2	67.8 ± 14.4	77.7 ± 16.1	72.5 ± 17.3	50.7	52.3 ± 11.3	55.3 ± 5.7	57.6 ± 18.2
FatFormer (CVPR'24) Liu et al. (2024)	62.8 ± 10.4	50.1 ± 3.6	45.6 ± 33.1	56.1 ± 10.7	90.0 ± 11.8	85.0 ± 14.9	51.2	52.1 ± 8.2	58.9 ± 8.0	61.3 ± 15.7
SAFE (KDD'25) Li et al. (2025)	50.3 ± 1.2	50.4 ± 1.3	1.1 ± 0.6	46.5 ± 20.8	49.7 ± 2.7	50.3 ± 1.1	59.2	49.1 ± 0.7	57.2 ± 18.5	46.0 ± 17.3
C2P-CLIP (AAAI'25) Tan et al. (2025)	74.4 ± 8.4	58.9 ± 10.8	38.9 ± 31.2	68.5 ± 11.4	92.0 ± 10.1	81.4 ± 15.6	51.1	57.1 ± 4.2	59.6 ± 7.7	64.7 ± 16.2
AIDE (ICLR'25) Yan et al. (2025)	61.2 ± 11.9	61.2 ± 8.6	19.1 ± 11.1	53.9 ± 18.6	59.4 ± 24.6	63.6 ± 13.9	63.1	48.8 ± 0.8	58.4 ± 12.9	54.3 ± 14.0
DRCT (ICML'24) Chen et al. (2024)	84.7 ± 2.7	94.4 ± 1.8	77.8 ± 5.4	84.8 ± 3.6	73.9 ± 13.4	81.4 ± 12.2	56.6	55.1 ± 1.8	50.6 ± 3.5	73.3 ± 15.5
AlignedForensics (ICLR'25) Rajan et al. (2025)	79.0 ± 22.7	95.1 ± 6.5	68.0 ± 20.7	77.4 ± 25.0	53.9 ± 7.1	66.6 ± 21.6	71.0	78.8 ± 17.8	80.1 ± 10.3	74.4 ± 11.4
DDA (NeurIPS'25) Chen et al. (2025)	91.7 ± 7.8	98.0 ± 1.4	97.2 ± 4.2	90.1 ± 5.6	81.4 ± 13.9	87.8 ± 12.6	82.4	90.9 ± 3.1	90.3 ± 3.5	90.0 ± 5.7
DDA*	97.4 ± 0.4	99.7 ± 0.1	**99.3 ± 0.8**	**99.8 ± 0.1**	83.8 ± 6.7	92.4 ± 7.0	92.1	92.0 ± 0.3	96.3 ± 0.7	94.2 ± 5.8
STAL (ours)	**98.6 ± 0.9**	**99.8 ± 0.1**	98.9 ± 1.8	99.7 ± 0.4	**94.7 ± 5.0**	**96.9 ± 4.2**	**92.2**	**95.2 ± 0.4**	**97.9 ± 1.0**	**97.0 ± 2.6**

Evaluation Metrics and Comparative Methods. We report balanced accuracy (BAL), defined as the average of real image and generated image accuracies. For benchmarks with multiple subsets, we report the arithmetic mean over subsets. We compare STAL with existing representative methods, including NPR (Tan et al., 2024b), UnivFD (Ojha et al., 2023), FatFormer (Liu et al., 2024), C2P-CLIP (Tan et al., 2025), SAFE (Li et al., 2025), AIDE (Yan et al., 2025), DRCT (Chen et al., 2024), AlignedForensics (Rajan et al., 2025), and DDA (Chen et al., 2025), using their publicly released model weights. We also include DDA*, a DDA variant trained under the same setting as STAL with DINOv3-H+ (Siméoni et al., 2025) as the backbone.

Implementation Details. During training, we use real images from MSCOCO (Lin et al., 2014) and their reconstructed images as the training data. For the spatial input, we follow DDA (Chen et al., 2025) to construct aligned training samples. In parallel, we keep the original input view for the frequency auxiliary branch to avoid disrupting frequency-domain characteristics. We train the model for one epoch with a batch size of 16 per GPU. The spatial detector uses DINOv3-H+ (Siméoni et al., 2025) as the backbone and adopts LoRA (Hu et al., 2022) with a rank of 8 and $\alpha = 1.0$ for parameter-efficient fine-tuning. The input resolution is 336×336. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-4} and a weight decay of 0.005. More implementation details are provided in Appendix C.

4.2 Cross-Dataset Comparison

Table 1 presents the results on nine benchmarks. Compared with existing methods, STAL achieves the best average BAL of 97.0% , ranks first on 7 out of 9 datasets, and shows stronger stability across datasets. Compared with DDA*, STAL improves the average BAL by 2.8 percentage points and reduces the cross-dataset standard deviation from 5.8 to 2.6. The gains are especially clear on ForenSynths, AIGCDetectionBenchmark, and SynthWildx, suggesting that the improvement is not merely due to a stronger backbone.

4.3 Robustness Analysis

Fig. 5 evaluates robustness on GenImage under JPEG compression, resizing, and Gaussian blur. STAL ranks first across all perturbation settings. Its balanced accuracy drops by only 2.38 points when JPEG quality decreases from $Q = 100$ to $Q = 60$, by 1.35 points under strong downsampling ($\alpha = 0.5$), and by at most 3.10 points under Gaussian blur. The margins over the second-best method remain large under severe perturbations, reaching 7.73 points at $Q = 60$, 9.84 points at $\alpha = 0.5$, and 10.73 points at $\sigma = 2.0$.

4.4 Ablation Analysis

Table 2 shows the impact of frequency-cue usage, frequency context and tail, and backbone choice. Using the frequency branch only for training-time auxiliary supervision achieves 97.0%, outperforming both the spatial-only baseline and the dual-branch inference setting. For frequency branch

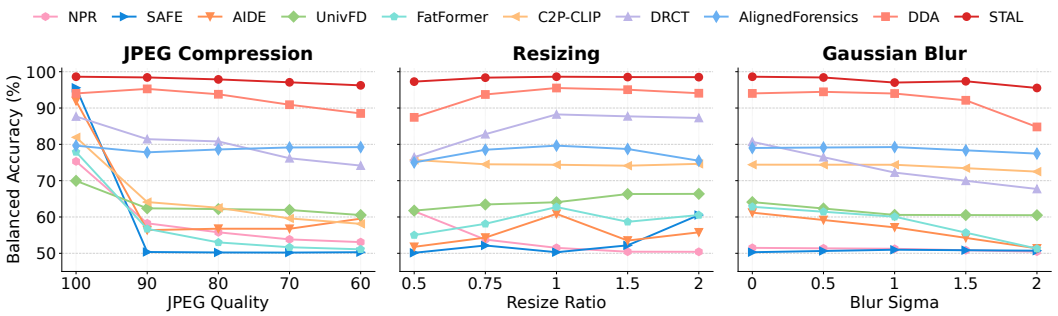


Figure 5: Robustness analysis on GenImage. We evaluate STAL and competing methods under JPEG compression, resizing, and Gaussian blur with increasing perturbation strengths.

leave-one-out experiment, the teacher without the tail band yields 94.4%, the tail-only teacher reaches 96.1%, and using both achieves the best performance of 97.0%. Under matched backbones, STAL outperforms the corresponding DDA variants. The larger gain with DINOv3-H+ indicates better compatibility with our frequency auxiliary supervision. These results validate both the auxiliary-supervision design and the contribution of spectral-tail modeling. **Additional results.** More detailed experiment results are provided in Appendix D.

5 Conclusion

We identify and systematically characterize spectral tail uplift, a phenomenon in which AI-generated images exhibit a pronounced uplift in the ultra-high-frequency tail of the spectrum, and explain it from the perspective of nonlinear harmonic accumulation in trained generative models. Based on this observation, we propose Spectral Tail Auxiliary Learning (STAL), a framework for generalizable AI-generated image detection that uses spectral-tail cues as auxiliary supervision during training. This design injects tail-related information into the spatial detector while keeping inference free from frequency-domain inputs or additional branches. Extensive experiments on multiple public benchmarks spanning diverse scenarios show that STAL achieves superior performance across generators and data distributions, highlighting its strong generalization ability and stability.

Table 2: Ablation studies on 9 benchmarks. (a) Frequency usage: how spectral information is used; (b) Frequency Branch Leave-One-Out: use frequency context and tail head or not to construct the auxiliary teacher; (c) Backbone: comparison under different spatial detector backbones. Results show the impact of each design choice on overall balanced accuracy.

Configuration	**BAL (%)**	**Δ**
(a) Frequency Usage		
Spatial-only baseline	94.2	–
Direct spatial-frequency inference	96.0	+1.8
Auxiliary frequency supervision (STAL)	**97.0**	**+2.8**
(b) Frequency Branch Leave-One-Out		
Spatial-only baseline	94.2	–
Without tail head	94.4	+0.2
Tail head only	96.1	+1.9
STAL	**97.0**	**+2.8**
Method	**BAL (%)**	**Δ**
(c) Backbone		
DDA (DINOv2-L) Chen et al. (2025)	90.0	–
STAL (DINOv2-L)	91.2	+1.2
DDA* (DINOv3-H+)	94.2	–
STAL (DINOv3-H+)	**97.0**	**+2.8**

Limitations. Although STAL shows strong performance, its frequency auxiliary branch primarily relies on radial spectrum statistics and local DCT statistics to capture frequency-domain information. While simple and effective, this design may not fully capture fine-grained spectral variations across different generative models. In future work, we would explore more adaptive frequency-band selection mechanisms and more lightweight forms of training-time frequency supervision.

AI USE STATEMENT

Generative AI tools are used solely as limited support for improving language and organizing the manuscript. The authors independently develop the framework and experimental design, assess the scientific interpretations, and make all final decisions concerning the claims presented in the study. Any AI-assisted code is reviewed and tested by the authors to confirm that it carries out the intended computations, and all numerical statements are checked against the corresponding recorded outputs. Likewise, all AI-assisted text is carefully reviewed, edited, and revised by the authors. Accordingly, the authors retain full responsibility for the final manuscript, including its text, mathematical content, results, software, and supplementary materials.

REFERENCES

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2024. doi: 10.1109/OJSP.2023.3337714.

Black Forest Labs. Flux.1 [dev]. Hugging Face model card, 2024. URL https://huggingface.co/black-forest-labs/FLUX.1-dev. Model card, accessed 2026-03-30.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=B1xsqj09Fm.

Bar Cavia, Eliahu Horwitz, Tal Reiss, and Yedid Hoshen. Real-time deepfake detection in the real world, 2024. URL https://openreview.net/forum?id=kkE7jlqKae.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man Cheung. A closer look at fourier spectrum discrepancies for cnn-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7200–7209, June 2021.

Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7621–7639. PMLR, 21–27 Jul 2024.

Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, and Shouhong Ding. Dual data alignment makes AI-generated image detector easier generalizable. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=C39ShJwtD5.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8789–8797, 2018.

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.

James W. Cooley and John W. Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965. doi: 10.1090/S0025-5718-1965-0178586-1.

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 (CVPR) Workshops*, pp. 4356–4366, June 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

David J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A*, 4(12):2379–2394, Dec 1987. doi: 10.1364/JOSAA.4.002379. URL https://opg.optica.org/josaa/abstract.cfm?URI=josaa-4-12-2379.

Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3247–3258. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/frank20a.html.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744, June 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

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.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations (ICLR)*, 2014. URL http://arxiv.org/abs/1312.6114.

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*, KDD '25, pp. 2405–2414. Association for Computing Machinery, 2025. ISBN 9798400712456.

Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone Von Gioi. Masksim: Detection of synthetic images by masked spectrum similarity analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3855–3865, June 2024.

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 David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 740–755, Cham, 2014. Springer International Publishing.

Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Jingdong Wang, and Yao Zhao. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10770–10780, June 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lare^2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17006–17015, 2024.

Midjourney, Inc. Midjourney. https://www.midjourney.com/home.

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 (CVPR)*, pp. 24480–24489, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/081b08068e4733ae3e7ad019fe8d172f-Paper-Conference.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021.

Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=doBkiqESYq.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):5052–5060, Mar. 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 (CVPR)*, pp. 28130–28139, June 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7184–7192, Apr. 2025. doi: 10.1609/aaai.v39i7.32772.

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 (CVPR)*, 2020.

Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for AI-generated image detection. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–6, 2019.

Nan Zhong, Yiran Xu, Sheng Li, Zhenxing Qian, and Xinpeng Zhang. Patchcraft: Exploring texture patch for efficient ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

Mingjian Zhu, Hanting Chen, Qiangyu YAN, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 77771–77782. Curran Associates, Inc., 2023.

A Controlled Experiments of Spectral Tail Uplift

A.1 Spectral Tail Uplift under JPEG Compression

Due to the loss of high-frequency information caused by JPEG compression, we evaluate how JPEG compression affects spectral tail uplift. We first reconstruct real images with the SD-ema VAE decoder (Rombach et al., 2022) to obtain fake samples while reducing semantic bias. We then apply JPEG compression to the SD-VAE reconstructions with quality factors of 90, 80, 70, and 60, and compare them with real images and uncompressed SD-VAE reconstructions. The spectra are computed using the same pipeline as in the main spectral analysis. The yellow dashed curve shows the power-law decay fitted from real-image data as a reference.

As shown in Fig. 6, JPEG compression reduces the high-frequency energy of VAE reconstructions, and lower quality factors lead to lower overall energy in the tail region. However, even at JPEG quality 60, the tail uplift in VAE-reconstructed spectra is still not fully suppressed. A clear anomalous upward trend remains visible in the spectral tail. This suggests that spectral tail uplift is not limited to lossless reconstruction settings, but remains an identifiable structural spectral cue after common post-processing, providing useful evidence for generated-image detection.

A.2 Effects of Nonlinear Activations and Trained Weights

The spectra above show that the VAE reconstruction process introduces additional high-frequency energy and produces spectral tail uplift in the reconstructed samples. We therefore conduct controlled experiments to investigate how tail uplift arises. We hypothesize that tail uplift is related to harmonic accumulation induced by nonlinear activations in generative models. To verify our hypothesis, we study the effects of nonlinear activations and trained weights on spectral tail uplift. We use both pink-noise inputs and real-image inputs, and reconstruct them with SD-ema VAE, ensuring that each output is paired with its corresponding input to reduce semantic bias.

Effect of nonlinear activations. To isolate the effect of nonlinear activations on tail uplift, we keep the trained SD-VAE weights fixed and replace the original SiLU activations with different alternatives. We consider four settings: the original SiLU, Identity, ReLU, and LeakyReLU with negative slope 0.01. Here, Identity denotes the identity mapping, which is equivalent to completely removing nonlinear activations and reducing the network to a purely linear transformation. As shown in the Fig. 3, replacing SiLU with Identity reduces the tail uplift to 0.01 for pink-noise inputs and 0.00 for real-image inputs, almost eliminating the effect. It indicates that nonlinear activations are crucial for producing spectral tail uplift. In contrast, replacing SiLU with ReLU or LeakyReLU substantially strengthens the uplift, suggesting that different nonlinear activations induce harmonic generation with different strengths. These results suggest that tail uplift is closely related to harmonic accumulation caused by nonlinear activations in generative models.

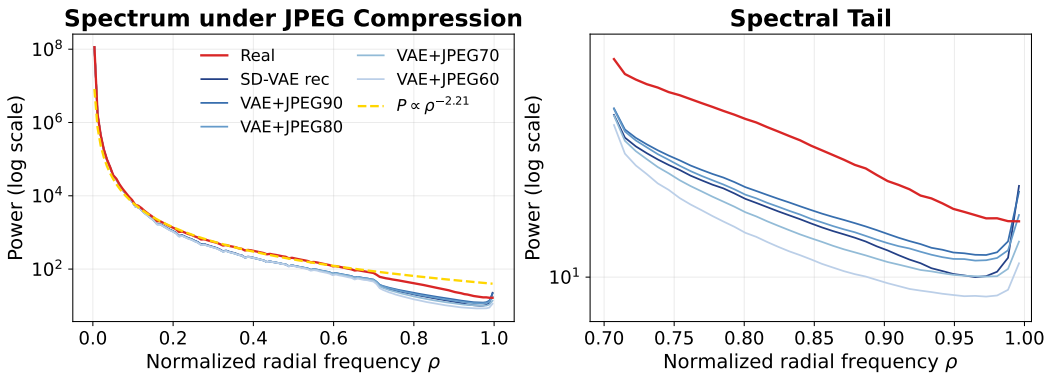


Figure 6: Effect of JPEG compression on the spectral tail. We apply JPEG compression with different quality factors to SD-VAE (Rombach et al., 2022) reconstructed images and compare their radial FFT power spectra with real images. The yellow curve denotes the power-law decay fitted from the data of real images and serves as a reference. **Left:** spectra over the full radial frequency range. **Right:** spectral tail over the local frequency range $\rho \in [0.7, 1]$.

Effect of trained weights. We further examine the role of generative-model weights by comparing trained weights with random weights. In the random-weight setting, we load the same SD-VAE architecture but reset all learnable weights using Kaiming random initialization (He et al., 2015). This keeps the architecture, activation functions, depth, channel dimensions, normalization layers, residual connections, and upsampling modules the same as those of the trained SD-VAE. The only difference is whether the weights have undergone generative-model training. As shown in the Fig. 7, the trained SD-VAE produces tail uplift values of $+0.12$ and $+0.14$ for pink-noise and real-image inputs, respectively. Under random weights, the same architecture produces much weaker uplift values of $+0.04$ and $+0.06$. The pink-noise input itself does not exhibit tail uplift, with an input baseline of $+0.00$. Therefore, the output uplift directly reflects how the network transforms the input spectrum. For pink-noise inputs, the $+0.04$ uplift under random weights can be attributed to weak harmonics generated enerated by nonlinear activations under random initialization. The trained VAE further amplifies this effect to $+0.12$, a three-fold increase. This suggests that training shapes the layer-wise frequency responses of the network, causing harmonic energy to accumulate across layers and become systematically concentrated in the spectral tail. The results on real-image inputs show the same trend.

As shown in Fig. 8 and Fig. 9, we repeat the controlled experiments on activation functions and network weights across eight additional VAEs and tokenizers, and observe consistent results. Replacing the original nonlinear activation functions with the identity function eliminate the spectral tail uplift, while replacing the trained weights with randomly initialized weights substantially suppresses or completely eliminates the tail uplift. Taken as a whole, these controlled experiments indicate that tail uplift depends on two factors. Nonlinear activations provide the ability to generate new frequency components, while trained weights make these components accumulate across layers and align with the spectral tail. Removing either factor substantially weakens or even suppresses tail uplift. With both factors present, the uplift reaches its full strength, removing the nonlinearity eliminates it, and removing training reduces it to roughly one third. These results establish that both nonlinear activations and trained weights are required for tail uplift, consistent with harmonic accumulation induced by their joint effect.

B Formal Analysis of Spectral Tail Uplift

This appendix provides the formal statements and proofs used in the main text to analyze spectral tail uplift. The analysis is intentionally based on a minimal signal model: a one-dimensional periodic signal, a linear convolutional filter represented by its frequency response, and a pointwise nonlinear activation. This simplified model is not intended to describe every architectural detail of modern generators. Instead, it isolates the mechanism needed by the main text: pointwise nonlinear

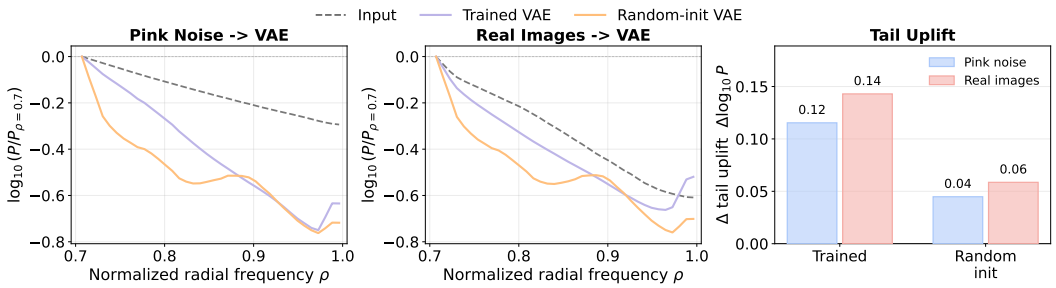


Figure 7: Comparison of spectral-tail under trained and random VAE weights. We keep the SD-VAE (Rombach et al., 2022) architecture fixed and compare trained weights with random-initialized weights using pink noise (**left**) and real images (**middle**) as inputs. Normalized curves show spectra on $\rho \in [0.7, 1]$. **Right:** tail uplift $\Delta \log_{10} P$, the rise from the tail's minimum to $\rho = 1$.

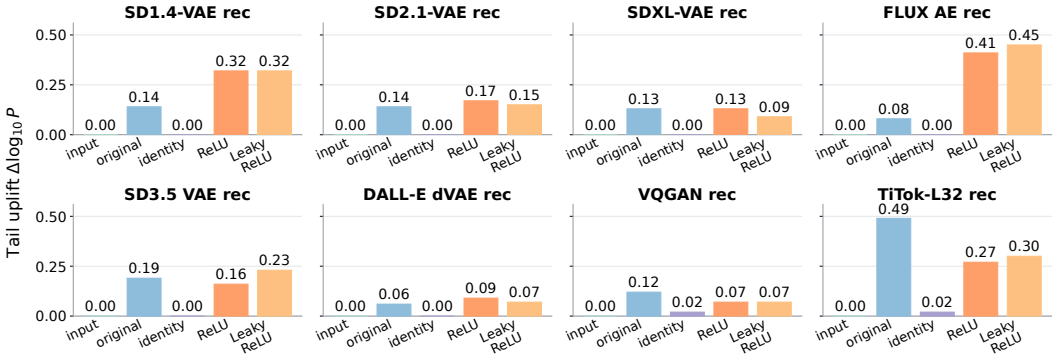


Figure 8: Controlled activation-function replacement experiments across eight different VAEs or tokenizers. We replace the original activation fuction with Identity, ReLU, or LeakyReLU, to expose the effect of nonlinear activations.

activations generate new harmonics, and cascaded filters modulate whether these harmonics survive along the decoder.

B.1 Harmonic Generation by Pointwise Nonlinearity

Theorem 1 (Harmonic generation). Let $x : \mathbb{R} \to \mathbb{R}$ be a real-valued 2π-periodic signal with finite bandwidth M,

$$x(t) = \sum_{m=-M}^{M} \hat{x}_m e^{imt}, \qquad \hat{x}_M \neq 0, \qquad \hat{x}_{-m} = \overline{\hat{x}_m}. \tag{19}$$

Let $\phi : \mathbb{R} \to \mathbb{R}$ be a degree-d polynomial activation,

$$\phi(z) = \sum_{q=0}^{d} a_q z^q, \qquad a_d \neq 0, \quad d \geq 2. \tag{20}$$

Then $\phi(x)$ has highest positive frequency dM, and the Fourier coefficient at this frequency is

$$\widehat{\phi(x)}_{dM} = a_d(\hat{x}_M)^d, \qquad \left|\widehat{\phi(x)}_{dM}\right|^2 = |a_d|^2 |\hat{x}_M|^{2d}. \tag{21}$$

Proof. Since ϕ is a polynomial, we can write

$$\phi(x(t)) = \sum_{q=0}^{d} a_q [x(t)]^q. \tag{22}$$

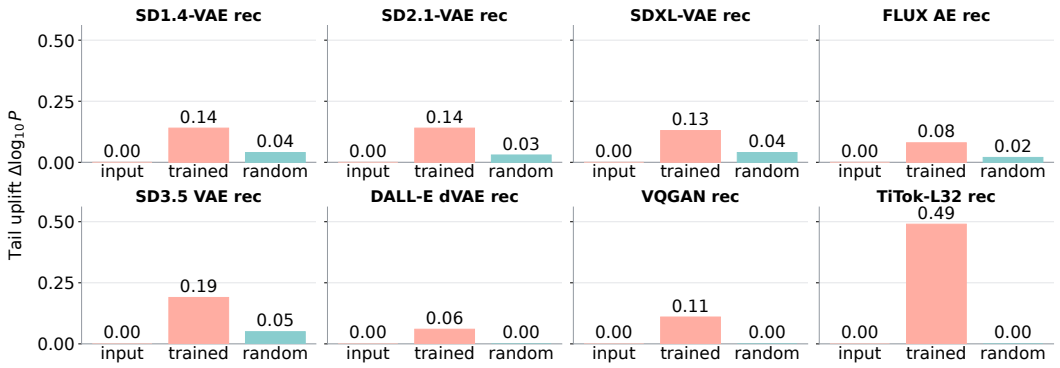


Figure 9: Controlled weights experiments across eight different VAEs or tokenizers. We replace the trained weights with random weights.

For each q, the Fourier coefficient of $[x(t)]^q$ at frequency k is the q-fold convolution of the Fourier coefficients of x:

$$\widehat{[x]^q}_k = \sum_{\substack{m_1+\cdots+m_q=k \\ |m_j|\le M}} \hat{x}_{m_1}\hat{x}_{m_2}\cdots\hat{x}_{m_q}. \tag{23}$$

Because each m_j lies in $[-M, M]$, the support of $[x]^q$ is contained in $[-qM, qM]$. Therefore, any term with $q < d$ cannot contribute to frequency dM.

For $q = d$ and $k = dM$, the constraint $m_1 + \cdots + m_d = dM$ with $m_j \le M$ has only one feasible solution: $m_1 = \cdots = m_d = M$. Hence

$$\widehat{[x]^d}_{dM} = (\hat{x}_M)^d. \tag{24}$$

Combining this with Eq. equation 22 gives

$$\widehat{\phi(x)}_{dM} = \sum_{q=0}^{d} a_q \widehat{[x]^q}_{dM} = a_d(\hat{x}_M)^d. \tag{25}$$

Since $a_d \neq 0$ and $\hat{x}_M \neq 0$, this coefficient is nonzero, so the highest positive frequency is exactly dM. Taking the squared magnitude gives Eq. equation 21.

B.2 Harmonic-Chain Propagation in a Decoder Cascade

Theorem 2 (Harmonic-chain propagation). Consider the L-layer cascade

$$y_\ell = H_\ell * x_{\ell-1}, \qquad x_\ell = \phi(y_\ell), \qquad \ell = 1, \ldots, L, \tag{26}$$

where H_ℓ is a linear convolutional filter with frequency response $H_\ell(k)$, and ϕ is the degree-d polynomial activation in Eq. equation 20. Let the input be a single-tone signal

$$x_0(t) = A\cos(k_0 t), \qquad \hat{x}_{0,k_0} = \hat{x}_{0,-k_0} = A/2. \tag{27}$$

Assume the filters are non-degenerate along the highest-order harmonic chain, i.e.,

$$H_\ell(d^{\ell-1}k_0) \neq 0, \qquad \ell = 1, \ldots, L. \tag{28}$$

Then the output power at the top harmonic $d^L k_0$ is

$$\left|\hat{x}_L(d^L k_0)\right|^2 = |a_d|^{\frac{2(d^L-1)}{d-1}} \left(\frac{A}{2}\right)^{2d^L} \prod_{\ell=1}^{L} \left|H_\ell(d^{\ell-1}k_0)\right|^{2d^{L-\ell+1}}. \tag{29}$$

Proof. We prove the result by induction along the highest-order harmonic chain. A linear convolution does not change the frequency support; in the Fourier domain it only rescales each coefficient:

$$\hat{y}_\ell(k) = H_\ell(k)\hat{x}_{\ell-1}(k). \tag{30}$$

For $L = 1$, the pre-activation coefficient at frequency k_0 is

$$\hat{y}_1(k_0) = H_1(k_0)\frac{A}{2}. \tag{31}$$

Applying Theorem 1 to y_1 gives

$$\hat{x}_1(dk_0) = a_d \left(H_1(k_0)\frac{A}{2} \right)^d. \tag{32}$$

Thus

$$|\hat{x}_1(dk_0)|^2 = |a_d|^2 |H_1(k_0)|^{2d} \left(\frac{A}{2} \right)^{2d}, \tag{33}$$

which is Eq. equation 29 when $L = 1$.

Now assume that after layer $\ell - 1$ the contribution of the highest-order harmonic chain satisfies

$$\left|\hat{x}_{\ell-1}(d^{\ell-1}k_0)\right|^2 = |a_d|^{\frac{2(d^{\ell-1}-1)}{d-1}} \left(\frac{A}{2} \right)^{2d^{\ell-1}} \prod_{j=1}^{\ell-1} \left|H_j(d^{j-1}k_0)\right|^{2d^{\ell-j}}. \tag{34}$$

The next linear filter gives

$$\left|\hat{y}_\ell(d^{\ell-1}k_0)\right|^2 = \left|H_\ell(d^{\ell-1}k_0)\right|^2 \left|\hat{x}_{\ell-1}(d^{\ell-1}k_0)\right|^2. \tag{35}$$

Along the highest-order chain, the degree-d term contributes to frequency $d^\ell k_0$ by combining the component at $d^{\ell-1}k_0$ exactly d times. Other lower-frequency components cannot reach this top frequency through the degree-d term. Therefore, the chain contribution obeys

$$\left|\hat{x}_\ell(d^\ell k_0)\right|^2 = |a_d|^2 \left|\hat{y}_\ell(d^{\ell-1}k_0)\right|^{2d}. \tag{36}$$

Substituting Eqs. equation 34 and equation 35 into Eq. equation 36 yields

$$\left|\hat{x}_\ell(d^\ell k_0)\right|^2 = |a_d|^{\frac{2(d^\ell-1)}{d-1}} \left(\frac{A}{2} \right)^{2d^\ell} \prod_{j=1}^{\ell} \left|H_j(d^{j-1}k_0)\right|^{2d^{\ell-j+1}}. \tag{37}$$

This is the desired expression for the highest-order chain at layer ℓ, so the induction is complete.

B.3 Scope of the Formal Model

Theorems 1 and 2 are used as a mechanism-level explanation rather than a complete generative model. Several points are important for avoiding over-interpretation.

Theorem 1 is stated for polynomial activations because this setting gives an exact closed-form coefficient. For smooth activations used in modern decoders, a finite Taylor approximation on the bounded activation range leads to the same type of local harmonic-generation mechanism. For piecewise linear activations, harmonics can also be introduced when the input crosses nonlinear kink regions, although the exact coefficients differ. We therefore use the polynomial model as an analytically tractable abstraction of the pointwise activation layer.

Theorem 2 does not claim that arbitrary filters necessarily amplify the harmonic chain. It shows where the filter gains enter the top-harmonic power. If a gain $H_\ell(d^{\ell-1}k_0)$ is close to zero, the corresponding harmonic path is suppressed; if the learned filters preserve or amplify these frequencies, the path can survive to the spectral tail. Thus, convolutional weights should be interpreted as a modulation and propagation factor rather than as an unconditional source of high-frequency energy.

Connection to controlled observations. The formal results above identify a mechanism by which decoder pointwise nonlinear activations can create harmonics and by which convolutional filters can modulate their propagation. This mechanism is supported by our controlled SD-VAE experiments. Under the same measurement protocol, real images follow the expected decaying spectral profile and do not exhibit a tail-uplift pattern. After SD-VAE reconstruction, the reconstructed fake images show a clear uplift shape at the spectral tail. To isolate the role of pointwise activations, we fix the trained VAE weights and replace all SiLU activation functions with identity mappings. The tail uplift is then almost completely suppressed, while replacing SiLU with other nonlinear activations restores or strengthens the effect. These observations support our interpretation that pointwise activations inject additional harmonics, while learned convolutional filters modulate how these harmonics propagate through the decoder.

C Implementation Details

C.1 Training Data and View Construction

During training, we use real images from MSCOCO (Lin et al., 2014) and their reconstructed counterparts to construct training data. For the spatial branch, we follow DDA (Chen et al., 2025) to build aligned training pairs, mitigating reconstruction-induced low-level pixel biases and encouraging the detector to learn more generalizable cues. Different from the spatial branch, the auxiliary frequency branch uses a frequency-preserving view. This view shares the necessary geometric transformations with the spatial view, but avoids augmentations that substantially alter the spectral shape, such as strong compression, resampling, or other operations that may distort spectral-tail cues. This design allows the training-time frequency teacher to capture spectral tail uplift consistently, rather than learning spectral perturbations caused by data augmentation.

C.2 More Training Details

We train STAL for 1 epoch on 8 NVIDIA V100 32GB GPUs with a batch size of 16 per GPU. The spatial detector uses DINOv3 ViT-H+/16 (Siméoni et al., 2025) with LoRA (Hu et al., 2022) fine-tuning, where the input resolution is 336×336, the LoRA rank is 8, and $\alpha = 1.0$. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1×10^{-4} and a weight decay of 0.005. The frequency-related losses follow a curriculum schedule and are discarded together with the frequency branch at inference time, so STAL introduces no additional inference cost. The frequency-related losses follow a lightweight curriculum schedule: the weight is linearly warmed up in the first 800 steps, kept at full strength in early training, and then cosine-annealed to zero between 15% and 45% of the estimated training steps. This encourages the spatial detector to absorb spectral-tail cues during training while avoiding dependence on the auxiliary frequency branch.

C.3 Evaluation Protocol

All experiments report balanced accuracy, defined as the average of real-image accuracy and generated-image accuracy to mitigate the effect of class imbalance. For datasets with multiple subsets, we report the arithmetic mean and standard deviation of balanced accuracy across subsets. All methods (Tan et al., 2024b; Ojha et al., 2023; Liu et al., 2024; Li et al., 2025; Tan et al., 2025; Yan et al., 2025; Chen et al., 2024; Rajan et al., 2025; Chen et al., 2025)are evaluated using their official model weights. DDA* denotes a DDA variant using the same DINOv3-H+ (Siméoni et al., 2025) backbone as STAL, which provides a controlled comparison under matched backbone. DDA* and our method are trained under the same setting with fixed random seeds, and the best checkpoint is selected for evaluation to ensure fairness.

Table 3: Comparison on EvalGEN. Balanced accuracy (%) is reported.

Method	Flux	GoT	Infinity	NOVA	OmniGen	Avg
NPR (CVPR'24) Tan et al. (2024b)	0.7	0.2	6.5	4.7	2.2	2.9 ± 2.7
UnivFD (CVPR'23) Ojha et al. (2023)	4.0	9.2	15.7	8.3	39.6	15.4 ± 14.2
FatFormer (CVPR'24) Liu et al. (2024)	9.9	47.9	44.7	98.3	27.3	45.6 ± 33.1
SAFE (KDD'25) Li et al. (2025)	1.0	0.5	1.9	0.6	1.6	1.1 ± 0.6
C2P-CLIP (AAAI'25) Tan et al. (2025)	8.7	49.6	35.3	86.4	14.5	38.9 ± 31.2
AIDE (ICLR'25) Yan et al. (2025)	17.9	24.7	3.4	16.3	33.4	19.1 ± 11.1
DRCT (ICML'24) Chen et al. (2024)	72.5	81.4	77.9	84.6	72.5	77.8 ± 5.4
AlignedForensics (ICLR'25) Rajan et al. (2025)	32.0	72.3	74.0	84.8	77.0	68.0 ± 20.7
DDA (NeurIPS'25) Chen et al. (2025)	89.9	99.5	97.8	99.5	99.5	97.2 ± 4.2
DDA*	**97.9**	**99.6**	99.8	99.8	**99.7**	$\mathbf{99.3 \pm 0.8}$
STAL (ours)	95.7	99.5	**99.9**	**99.9**	99.5	98.9 ± 1.8

D More Experiment Results

D.1 More Comparison Results

In this section, we provide complete detailed results on the benchmarks (Zhu et al., 2023; Chen et al., 2024; Bammey, 2024; Chen et al., 2025; Wang et al., 2020; Zhong et al., 2023; Cozzolino et al., 2024; Cavia et al., 2024) that are not fully reported in the Section 4.2, as a supplement. Table 4 reports results on AIGCDetectBenchmark, covering various GAN-based and diffusion-based generators. It provides a comprehensive evaluation of detector generalization across different generator architectures. STAL achieves the best overall performance with an accuracy of 96.9%, outperforming the second-best method by 4.5%. In addition, Table 5 reports the results on two in-the-wild

Table 4: Comparison on AIGCDetectionBenchmark. Balanced accuracy (%) is reported.

Method	ADM	DALL·E 2	GLIDE	Midjourney	VQDM	BigGAN	CycleGAN	GauGAN	ProGAN	SDXL	SD14	SD15	StarGAN	StyleGAN	StyleGAN2	WFR	Wukong	Avg
NPR (CVPR'24) Tan et al. (2024b)	43.8	20.0	41.2	53.4	48.4	53.1	76.6	42.2	58.7	59.6	55.1	55.0	67.4	57.9	54.6	58.8	57.4	53.1 ± 12.2
UnivFD (CVPR'23) Ojha et al. (2023)	62.5	50.0	61.3	55.1	76.9	87.5	96.9	98.8	**99.4**	58.2	55.6	55.7	95.1	80.0	69.4	69.2	61.1	72.5 ± 17.3
FatFormer (CVPR'24) Liu et al. (2024)	80.2	68.5	91.1	54.4	88.0	**99.2**	**99.5**	**99.1**	98.5	71.7	67.5	67.2	99.4	**98.0**	**98.8**	88.3	75.6	85.0 ± 14.9
SAFE (KDD'25) Li et al. (2025)	49.5	49.5	53.0	49.0	50.2	52.2	51.9	50.0	50.0	49.8	49.7	49.8	50.1	50.0	50.0	49.8	50.3	50.3 ± 1.1
C2P-CLIP (AAAI'25) Tan et al. (2025)	71.6	52.3	73.5	56.6	73.7	98.4	96.8	98.8	99.3	62.3	77.5	76.9	**99.6**	93.1	79.4	94.8	79.4	81.4 ± 15.6
AIDE (ICLR'25) Yan et al. (2025)	52.9	51.1	60.2	49.8	69.3	70.1	93.6	60.6	89.0	49.6	51.6	51.0	72.1	66.5	59.0	80.6	54.5	63.6 ± 13.9
DRCT (ICML'24) Chen et al. (2024)	79.9	89.2	89.2	85.5	88.6	81.4	91.0	93.8	71.1	88.3	91.4	91.0	53.0	62.7	63.8	73.9	90.8	81.4 ± 12.2
AlignedForensics (ICLR'25) Rajan et al. (2025)	51.6	52.0	55.6	96.2	72.1	51.2	49.5	50.8	50.7	95.1	**99.7**	**99.6**	53.8	52.7	51.6	50.0	**99.6**	66.6 ± 21.6
DDA (NeurIPS'25) Chen et al. (2025)	89.5	94.6	89.6	95.6	76.6	91.0	72.5	92.7	92.8	99.4	98.7	98.6	72.7	87.8	90.2	52.1	98.8	87.8 ± 12.6
DDA*	97.0	**98.8**	97.5	96.8	97.9	91.3	90.8	90.9	81.8	98.8	97.8	97.7	75.2	83.5	86.2	91.8	97.8	92.4 ± 7.0
STAL (ours)	**97.3**	98.4	**98.0**	**98.2**	**99.4**	97.7	95.9	97.8	95.3	**99.8**	99.5	99.3	81.7	95.9	96.7	**96.4**	99.4	**96.9 ± 4.2**

Table 5: Comparison on SynthWildx and WildRF. Balanced accuracy (%) is reported.

Method	SynthWildx				WildRF			
	DALL·E 3	Firefly	Midjourney	Avg.	Facebook	Reddit	Twitter	Avg.
NPR (CVPR'24) Tan et al. (2024b)	43.6	61.3	44.5	49.8 ± 10.0	78.1	61.0	51.3	63.5 ± 13.6
UnivFD (CVPR'23) Ojha et al. (2023)	45.4	65.3	46.2	52.3 ± 11.3	49.1	60.2	56.5	55.3 ± 5.7
FatFormer (CVPR'24) Liu et al. (2024)	46.5	61.6	48.3	52.1 ± 8.2	54.1	68.1	54.4	58.9 ± 8.0
SAFE (KDD'25) Li et al. (2025)	49.4	48.2	49.6	49.1 ± 0.7	50.9	74.1	37.5	57.2 ± 18.5
C2P-CLIP (AAAI'25) Tan et al. (2025)	56.9	61.4	53.0	57.1 ± 4.2	54.4	68.4	55.9	59.6 ± 7.7
AIDE (ICLR'25) Yan et al. (2025)	63.4	48.8	51.9	54.7 ± 7.7	57.8	71.5	45.8	58.4 ± 12.9
DRCT (ICML'24) Chen et al. (2024)	58.3	56.4	50.5	55.1 ± 1.8	46.6	53.1	55.2	50.6 ± 3.5
AlignedForensics (ICLR'25) Rajan et al. (2025)	85.5	58.5	92.2	78.8 ± 17.8	89.4	69.1	81.8	80.1 ± 10.3
DDA (NeurIPS'25) Chen et al. (2025)	92.3	87.3	93.1	90.9 ± 3.1	93.1	86.4	91.5	90.3 ± 3.5
DDA*	92.3	91.7	92.0	92.0 ± 0.3	96.6	95.5	96.9	96.3 ± 0.7
STAL (ours)	**95.6**	**94.8**	**95.4**	**95.2 ± 0.4**	**98.4**	**96.7**	**98.6**	**97.9 ± 1.0**

Table 6: Comparison on GenImage. Balanced accuracy (%) is reported.

Method	Midjourney	SDv1.4	SDv1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg
NPR (CVPR'24) Tan et al. (2024b)	53.4	55.1	55.0	43.8	41.2	57.4	48.4	57.7	51.5 ± 6.3
UnivFD (CVPR'23) Ojha et al. (2023)	55.1	55.6	55.7	62.5	61.3	61.1	76.9	84.4	64.1 ± 10.8
FatFormer (CVPR'24) Liu et al. (2024)	52.1	53.6	53.8	61.4	65.5	60.9	72.5	82.2	62.8 ± 10.4
SAFE (KDD'25) Li et al. (2025)	49.0	49.7	49.8	49.5	53.0	50.3	50.2	50.9	50.3 ± 1.2
C2P-CLIP (AAAI'25) Tan et al. (2025)	56.6	77.5	76.9	71.6	73.5	79.4	73.7	85.9	74.4 ± 8.4
AIDE (ICLR'25) Yan et al. (2025)	58.2	77.2	77.4	50.4	54.6	70.5	50.8	50.6	61.2 ± 11.9
DRCT (ICML'24) Chen et al. (2024)	82.4	88.3	88.2	76.9	86.1	87.9	85.4	87.0	84.7 ± 2.7
AlignedForensics (ICLR'25) Rajan et al. (2025)	97.5	**99.7**	**99.6**	52.4	57.6	**99.6**	75.0	50.6	79.0 ± 22.7
DDA (NeurIPS'25) Chen et al. (2025)	95.6	98.7	98.6	89.5	89.6	98.7	76.5	86.5	91.7 ± 7.8
DDA*	96.8	97.8	97.7	97.0	97.5	97.8	97.9	97.0	97.4 ± 0.4
STAL (ours)	**98.2**	99.5	99.3	**97.3**	**98.0**	99.4	**99.4**	**97.9**	**98.6 ± 0.9**

Table 7: Comparison on DRCT-2M. Balanced accuracy (%) is reported..

Method	LDM	SDv1.4	SDv1.5	SDv2	SDXL	SDXL-Refiner	SD-Turbo	SDXL-Turbo	LCM-SDv1.5	LCM-SDXL	SDv1-Ctrl	SDv2-Ctrl	SDXL-Ctrl	Avg
NPR (CVPR'24) Tan et al. (2024b)	33.0	29.1	29.0	35.1	33.2	28.4	27.9	27.9	29.4	30.2	28.4	28.3	34.7	30.4 ± 2.7
UnivFD (CVPR'23) Ojha et al. (2023)	85.4	56.8	56.4	58.2	63.2	55.0	56.5	53.0	54.5	65.9	68.0	65.4	75.9	62.6 ± 9.5
FatFormer (CVPR'24) Liu et al. (2024)	55.9	48.2	48.2	48.2	48.2	48.3	48.2	48.2	48.3	50.6	49.7	49.9	59.8	50.1 ± 3.6
SAFE (KDD'25) Li et al. (2025)	50.3	50.1	50.0	50.0	49.9	50.1	50.0	50.0	50.1	50.0	49.9	50.0	54.7	50.4 ± 1.3
C2P-CLIP (AAAI'25) Tan et al. (2025)	83.0	51.7	51.7	52.9	51.9	64.6	51.7	50.6	52.0	66.1	56.9	54.7	77.8	58.9 ± 10.8
AIDE (ICLR'25) Yan et al. (2025)	64.4	74.9	75.1	58.5	53.5	66.3	52.8	52.8	70.0	54.3	65.9	53.6	53.9	61.2 ± 8.6
DRCT (ICML'24) Chen et al. (2024)	96.7	96.3	96.3	94.9	96.2	93.5	93.4	92.9	91.2	95.0	95.6	92.7	92.0	94.4 ± 1.8
AlignedForensics (ICLR'25) Rajan et al. (2025)	**99.9**	**99.9**	**99.9**	99.6	90.2	81.3	99.7	89.4	99.7	90.0	**99.9**	99.2	87.6	95.1 ± 6.5
DDA (NeurIPS'25) Chen et al. (2025)	99.2	98.9	99.0	98.3	98.0	96.8	97.9	94.8	95.9	98.2	98.7	99.0	99.4	98.0 ± 1.4
DDA*	99.8	99.5	99.5	99.7	99.7	**99.7**	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.7 ± 0.1
STAL (ours)	**99.9**	99.7	99.6	**99.8**	**99.8**	**99.7**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.8 ± 0.1**

datasets, SynthWildx and WildRF. STAL achieves a balanced accuracy of 95.2% on SynthWildx and 97.9% on WildRF. These results indicate that STAL also achieves strong performance in real-world scenarios. Tables 6 and 7 present detailed results on GenImage and DRCT-2M, respectively, covering a wide range of generative models and variants, including GAN- and diffusion-based generators. Table 8 further evaluates STAL on Synthbuster, which includes images from diverse commercial and open-source generative models. Table 9 reports results on ForenSynths, focusing on generalization to early GAN-based generators and image synthesis methods. Finally, Table 3 reports detection performance on EvalGEN, which contains recent generative models. Across these detailed benchmarks, STAL achieves the best or near-best average performance in most cases. These results suggest that the proposed auxiliary frequency supervision improves cross-generator generalization across both classical and recent generative architectures.

Table 8: Comparison on Synthbuster. Balanced accuracy (%) is reported.

Method	DALL·E 2	DALL·E 3	Firefly	GLIDE	Midjourney	SD 1.3	SD 1.4	SD 2	SDXL	Avg
NPR (CVPR'24) Tan et al. (2024b)	51.1	49.3	46.5	48.5	52.8	51.4	51.8	46.0	52.8	50.0 ± 2.6
UnivFD (CVPR'23) Ojha et al. (2023)	83.5	47.4	89.9	53.3	52.5	70.4	69.9	75.7	68.0	67.8 ± 14.4
FatFormer (CVPR'24) Liu et al. (2024)	59.4	39.5	60.3	72.7	44.4	53.7	54.0	52.3	69.1	56.1 ± 10.7
SAFE (KDD'25) Li et al. (2025)	58.0	9.9	10.3	52.2	56.7	59.4	59.1	53.0	59.5	46.5 ± 20.8
C2P-CLIP (AAAI'25) Tan et al. (2025)	55.6	63.2	59.5	86.7	52.9	75.2	76.7	69.2	77.7	68.5 ± 11.4
AIDE (ICLR'25) Yan et al. (2025)	34.9	33.7	24.8	65.0	57.5	74.1	73.7	53.2	68.4	53.9 ± 18.6
DRCT (ICML'24) Chen et al. (2024)	77.2	86.6	84.1	82.6	73.7	86.6	86.6	83.2	71.3	84.8 ± 3.6
AlignedForensics (ICLR'25) Rajan et al. (2025)	50.2	48.9	51.7	53.5	98.7	98.8	98.8	98.6	97.3	77.4 ± 25.0
DDA (NeurIPS'25) Chen et al. (2025)	86.3	90.0	91.9	76.5	93.5	92.9	92.7	93.3	93.5	90.1 ± 5.6
DDA*	**99.8**	**99.9**	**99.9**	**99.5**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	**99.8 ± 0.1**
STAL (ours)	99.6	**99.9**	99.8	98.6	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**	99.7 ± 0.4

Table 9: Comparison on ForenSynths. Balanced accuracy (%) is reported.

Method	BigGAN	CRN	CycleGAN	DeepFake	GauGAN	IMLE	ProGAN	SAN	SeeingDark	StarGAN	StyleGAN	StyleGAN2	WFR	Avg
NPR (CVPR'24) Tan et al. (2024b)	53.1	0.4	76.6	35.7	42.2	5.3	58.7	48.4	63.6	67.4	57.9	54.6	58.8	47.9 ± 22.6
UnivFD (CVPR'23) Ojha et al. (2023)	87.5	55.7	96.9	69.4	98.8	68.1	**99.4**	58.2	62.2	95.1	80.0	69.4	69.2	77.7 ± 16.1
FatFormer (CVPR'24) Liu et al. (2024)	**99.3**	72.1	**99.5**	**93.0**	**99.3**	72.1	98.4	70.8	81.9	99.4	**98.1**	**98.9**	88.3	90.0 ± 11.8
SAFE (KDD'25) Li et al. (2025)	52.2	50.0	51.9	50.1	50.0	50.0	50.0	50.9	41.1	50.1	50.0	50.0	49.8	49.7 ± 2.7
C2P-CLIP (AAAI'25) Tan et al. (2025)	98.4	93.3	96.8	92.6	98.8	93.2	99.3	63.2	**94.7**	**99.6**	93.1	79.4	94.8	92.0 ± 10.1
AIDE (ICLR'25) Yan et al. (2025)	70.1	12.2	93.6	53.2	60.6	15.9	89.0	55.3	44.2	72.1	66.5	59.0	80.6	59.4 ± 24.6
DRCT (ICML'24) Chen et al. (2024)	81.4	78.4	91.0	51.5	93.8	82.6	71.1	84.9	72.2	53.0	62.7	63.8	73.9	73.9 ± 13.4
AlignedForensics (ICLR'25) Rajan et al. (2025)	51.2	50.4	49.5	71.7	50.8	49.7	50.7	67.6	51.4	53.8	52.7	51.6	50.0	53.9 ± 7.1
DDA (NeurIPS'25) Chen et al. (2025)	91.0	87.0	72.5	76.5	92.7	89.7	92.8	94.7	58.6	72.7	87.8	90.2	52.1	81.4 ± 13.9
DDA*	91.4	70.5	90.8	74.2	90.9	70.5	81.9	94.3	83.9	75.2	83.6	86.2	91.8	83.5 ± 8.5
STAL (ours)	97.7	**96.9**	95.9	70.8	97.8	**96.8**	95.3	**97.0**	84.4	81.7	95.9	96.7	**96.4**	**92.6 ± 8.3**

D.2 More Robustness Analysis

In addition to cross-dataset generalization, we further evaluate the stability of STAL under common image post-processing operations. All experiments are conducted on GenImage (Zhu et al., 2023), where we apply three types of perturbations, including JPEG compression, resizing, and Gaussian blur, and report balanced accuracy. Figure 5 shows the robustness results under these three post-processing operations.

JPEG compression. Under JPEG compression, when the quality factor decreases from $Q = 100$ to $Q = 60$, the balanced accuracy of STAL drops from 98.62% to 96.24%, corresponding to only a 2.38 percentage point decrease. More importantly, as the compression becomes stronger, the advantage of STAL over the second-best method does not shrink: STAL leads by 3.04 percentage points at $Q = 100$ and by 7.73 percentage points at $Q = 60$. This shows that STAL performs better not only on uncompressed or mildly compressed images, but also under stronger JPEG recompression.

Resizing. For resizing, using $\alpha = 1.0$ as the reference point, STAL still reaches 97.27% at $\alpha = 0.5$, with a decrease of 1.35 percentage points. At $\alpha = 0.75$, 1.5, and 2.0, it obtains 98.36%, 98.52%, and 98.50%, respectively, which are nearly unchanged from the result at $\alpha = 1.0$. Across all resize ratios, STAL outperforms the second-best method by at least 3.11 percentage points. Under the most challenging strong downsampling setting, i.e., $\alpha = 0.5$, the margin further increases to 9.84 percentage points.

Gaussian blur. Under Gaussian blur, blur is the most challenging perturbation for STAL among the three types in terms of performance drop, but the maximum decrease remains within 3.10 percentage points. In terms of comparison with other methods, STAL ranks first at all blur strengths and exceeds the second-best method by 10.73 percentage points at $\sigma = 2.0$.

Overall, STAL exhibits the strongest robustness across all perturbation settings, and its advantage becomes larger under more challenging conditions such as strong JPEG compression, strong downsampling, and strong blur. These results indicate that using frequency-domain information as auxiliary supervision for the spatial detector improves cross-generator generalization while maintaining strong adaptability to common image post-processing operations.

D.3 Heatmap Visualization

As shown in Fig. 10, we visualize the attention maps of the spatial-only baseline and STAL. The spatial-only model, trained without frequency-domain auxiliary supervision, exhibits relatively dif-

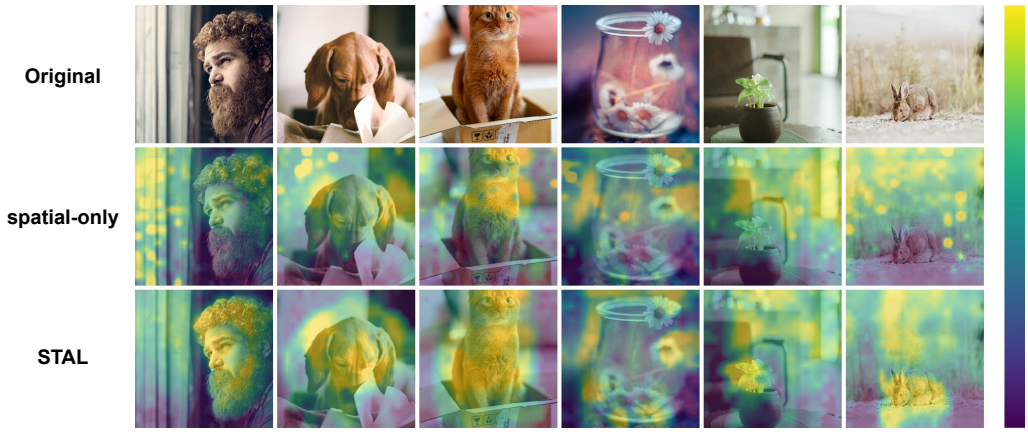


Figure 10: Visualization of model attention. From top to bottom, the three rows show the original images, heatmaps generated by the spatial-only model, and heatmaps generated by STAL, respectively. Colors from yellow to purple indicate attention values from high to low.

fuse attention responses. In contrast, STAL yields more localized responses over regions associated with high-frequency cues, suggesting that the auxiliary frequency supervision encourages the spatial encoder to attend more selectively to frequency-relevant structures.