RA-Det: Towards Universal Detection of AI-Generated Images via Robustness Asymmetry

Xinchang Wang [*1] **Yunhao Chen** [*2] **Yuechen Zhang** [1] **Congcong Bian** [1] **Zihao Guo** [1] **Xingjun Ma** [2] **Hui Li** [1]

Abstract

Recent image generators produce photo-realistic content that undermines the reliability of downstream recognition systems. As visual appearance cues become less pronounced, appearance-driven detectors that rely on forensic cues or high-level representations lose stability. This motivates a shift from appearance to behavior, focusing on how images respond to controlled perturbations rather than how they look. In this work, we identify a simple and universal behavioral signal. Natural images preserve stable semantic representations under small, structured perturbations, whereas generated images exhibit markedly larger feature drift. We refer to this phenomenon as **robustness asymmetry** and provide a theoretical analysis that establishes a lower bound connecting this asymmetry to memorization tendencies in generative models, explaining its prevalence across architectures. Building on this insight, we introduce Robustness Asymmetry Detection (RA-Det), a behavior-driven detection framework that converts robustness asymmetry into a reliable decision signal. Evaluated across 14 diverse generative models and against more than 10 strong detectors, RA-Det achieves superior performance, improving the average performance by 7.81%. The method is data- and model-agnostic, requires no generator fingerprints, and transfers across unseen generators. Together, these results indicate that robustness asymmetry is a stable, general cue for synthetic-image detection and that carefully designed probing can turn this cue into a practical, universal detector. The source code is publicly available at GitHub.

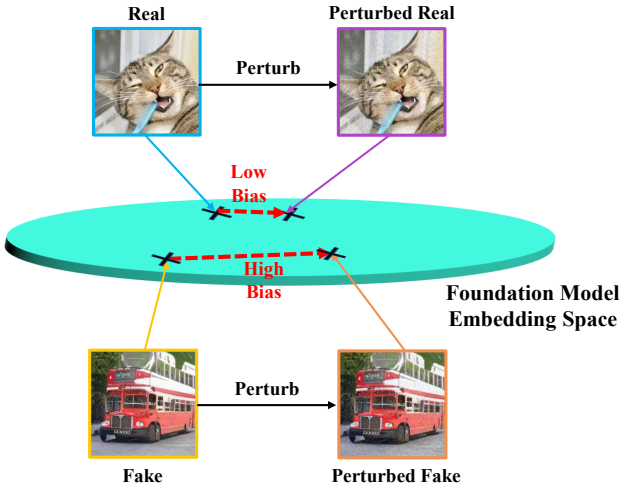


Figure 1. Robustness asymmetry between natural and synthetic images in embedding spaces (CLIP (Radford et al., 2021), DINO (Siméoni et al., 2025)). Under perturbation, show minimal embedding displacement, while synthetic images exhibit large representation drift (high displacement).

1. Introduction

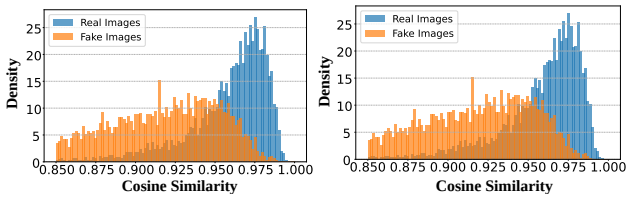
In recent years, generative models have evolved from early experimental frameworks into mature systems capable of producing highly realistic and often indistinguishable generated images. These technologies, alongside user-friendly platforms (Karras et al., 2019; Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2022; Midjourney, 2023; Runway, 2024), empower users to create hyper-realistic content with ease — bringing about the phenomenon where *seeing is no longer believing* (Nightingale & Farid, 2022).

However, the diversity and convenience of generative model technologies, coupled with their ability to generate highly realistic or generated images, also introduce potential risks of misuse and malicious applications(Wang et al., 2024; Lin et al., 2024), such as generating fake information, forging identities and disseminating inappropriate content. Thus, detecting generated images becomes increasingly critical.

Existing generated detection approaches can be broadly grouped into two major paradigms. Representation-driven approaches leverage general-purpose embeddings from

[*]Equal contribution [1]Jiangnan University [2]Fudan University. Correspondence to: Hui Li <lihui.cv@jiangnan.edu.cn>.

Preprint. March 2, 2026.



(b) DINOv3 ViT-L/16 (c) CLIP ViT-L/14

Figure 2. Backbone-agnostic evidence of robustness asymmetry. We measure embedding stability as the cosine similarity between features extracted from the clean image and its perturbed counterpart. Real images maintain higher similarity than synthetic images, and the separation remains consistent for both DINOv3 and CLIP.

large pre-trained vision models such as CLIP (Radford et al., 2021), DINOv2(Oquab et al., 2023), or vision–language models(Girdhar et al., 2023). These detectors adapt high-level representations via linear probes, adapters, or multi-modal alignment objectives (Ojha et al., 2023; Tan et al., 2025; Yu et al., 2025). Their strength lies in the semantic coverage and generalization ability inherited from the backbone. Artifact-driven forensic methods, in contrast, target generator-induced statistical traces including frequency anomalies(Frank et al., 2020), decoding residuals(Le & Woo, 2021), upsampling artifacts or local pixel structural disruptions(Tan et al., 2024; Liang et al., 2025). These approaches are lightweight and interpretable, excelling especially when generative pipelines introduce stable structural cues.

Although these two paradigms differ in methodology, they share a common assumption: the decision signal resides in visible appearance, either as explicit low-level artifacts or implicit high-level cues encoded in representations. As generative models reduce classical artifacts and produce distributions closer to natural images, appearance-driven cues grow increasingly subtle and fragile especially under common transformations such as JPEG compression or blur.

In this work, we explore a complementary perspective: instead of analyzing how an image *looks*, we examine how it *behaves* under controlled perturbations. Through extensive studies, we uncover an interesting phenomenon: natural images maintain consistent deep representations when subjected to mild perturbations. Generated images, however, display disproportionately large feature drift. We term this universal generative deficiency as **robustness asymmetry**, as shown in Fig. 1. We futher quantify this effect and Fig. 2 shows that the separation persists across both DINOv3 and CLIP backbones. Moreover, we theoretically derive a lower bound for the robustness asymmetry, showing it is directly linked to the memorization behavior of generative models. This link explains why robustness asymmetry is a widespread phenomenon, as memorization is prevalent across various types of generative models (Carlini et al., 2021; 2023; Chen et al., 2024b; 2020).

Motivated by this insight, we propose Robustness Asymmetry Detection (RA-Det), a behavior-driven framework that operationalizes robustness asymmetry into an effective and generalizable detection signal. RA-Det actively probes image stability via small, semantics-preserving perturbations and integrates complementary evidence through a unified multi-branch architecture. This design captures the full spectrum of generative inconsistencies: it leverages foundation priors for semantic context, explicitly quantifies feature-space displacement via covariance-aware and vector discrepancy metrics, and recovers high-frequency artifacts through a specialized low-level residual stream. By jointly optimizing these diverse signals, RA-Det leverages Robustness Asymmetry that persists across architectures and common post-processing operations.

In short, the main contributions of our proposed framework are summarized as follows:

- We uncover a simple and universal signal–robustness asymmetry between natural and synthetic images, showing that real images remain semantically stable under perturbations, whereas synthetic images demonstrate significantly larger feature-space displacement.
- Based on this insight, we propose RA-Det, a multi-branch detector that aggregates complementary semantic, discrepancy, and low-level residual cues to systematically probe and amplify the robustness discrepancy.
- Following a widely adopted benchmark, our proposed RA-Det outperforms existing detection methods, delivering excellent performance in the generated detection task and fully demonstrating its superior generalization and robustness.

2. Related Works

Existing detection methods can be grouped into two paradigms: representation-driven and artifact-driven.

2.1. Representation-Driven Detection Approaches

Methods such as UniFD (Ojha et al., 2023) classify CLIP embeddings directly, demonstrating strong generalization from the semantic and structural priors encoded in the backbone. Subsequent works enhance this paradigm by adapting pre-trained models to better capture subtle generative inconsistencies. C2P-CLIP (Tan et al., 2025) introduces cross-view prompting and multi-perspective consistency, LASTED (Wu et al., 2025) integrates textual supervision to refine multimodal representations, and LVLM-based detectors (Yu et al., 2025) leverage large vision–language models for explainable detection. These approaches benefit from broad domain coverage and the strong invariances of foundation models, enabling them to detect synthetic images

beyond specific generative architectures.

2.2. Artifact-Driven Forensic Approaches

Frequency-based detectors (Frank et al., 2020) expose anomalies in DCT or Fourier spectra, while gradient-based methods such as LGrad (Tan et al., 2023) capture inconsistencies in local edge structures. Recent works have revisited upsampling and decoding artifacts: NPR (Tan et al., 2024) and FerretNet (Liang et al., 2025) model local pixel relationships and decoder-induced correlations. Diffusion-based reconstruction detectors such as DIRE (Wang et al., 2023) and DRCT (Chen et al., 2024a) analyze reconstruction errors to reveal inconsistencies in synthetic content. These approaches are lightweight and interpretable, but their performance often depends on the persistence of structural artifacts, which modern diffusion and rectified-flow models increasingly suppress.

From Appearance to Behavior Although representation-driven and artifact-driven methods exploit appearance-level cues, their effectiveness declines as modern generators reduce artifacts and better mimic natural statistics. We shift the focus to **how an image behaves** under structured perturbations. This behavioral view reveals a robust asymmetry between real and synthetic images and provides the foundation for our framework.

3. Methodology

In this section, we detail RA-Det, a behavior-driven detection framework, as illustrated in Figure 3.

3.1. Overview and Notation

RA-Det is a behavior-driven detection framework built on robustness asymmetry in a frozen foundation feature space (Figure 3). Given an input image x with label $y \in \{0, 1\}$, a frozen encoder $f(\cdot)$ extracts the clean embedding $e = f(x)$. A learnable probing module $\mathcal{P}(\cdot)$ generates a bounded perturbation δ, producing a perturbed view $x' = x + \delta$ and the corresponding embedding $e' = f(x')$. Discrepancy features are derived from (e, e'), and a multi-branch detector aggregates complementary evidence from the semantic embedding, discrepancy cues, and low-level residual features to output the final prediction.

3.2. Core components

3.2.1. Differential Robustness Probing

Motivated by the observed robustness asymmetry, we propose a novel probing component named Differential Robustness Probing (DRP) that leverages the differential robustness asymmetry for detection purposes. The primary objective of the DRP is not to attack or fool a detector, but rather to learn and apply a controlled perturbation that maximally amplifies the observable robustness discrepancy between real and fake images.

Architecturally, DRP is implemented as a conditional UNet (Ronneberger et al., 2015) that predicts a pixel-wise perturbation map. The encoder follows a multi-stage downsampling design to extract multi-scale image features, while the clean embedding is projected into spatial feature maps to provide semantic conditioning. At the bottleneck, a cross-attention module integrates the embedding-conditioned features with the image features. The decoder mirrors the encoder with progressive upsampling and skip connections, recovering spatial details at each scale. Finally, a lightweight output head produces the perturbation map, and a $\tanh$ activation followed by scaling with ϵ constrains its magnitude, yielding a bounded perturbation for constructing the perturbed image.

Given an input image x, we apply designed DRP module $\mathcal{P}(\cdot)$ to obtain the perturbed counterpart x':

$$x' = \mathcal{P}(x) = x + \delta \tag{1}$$

3.2.2. Discrepancy Features in Foundation Space

Both x and x' are projected into the foundation space via $f(\cdot)$, yielding $e = f(x)$ and $e' = f(x')$. RA-Det models robustness asymmetry through discrepancy features derived from (e, e').

Distance, difference, and similarity. We compute the embedding distance $d = \|e - e'\|_2$ and the embedding difference vector $\Delta = e - e'$. In addition, the per-sample similarity is defined as

$$s = \cos(e, e'). \tag{2}$$

Diagonal-restricted covariance approximation (optional). To quantify feature displacement with reduced complexity, we further consider a diagonal-restricted covariance approximation (DCA):

$$\mathrm{DCA}(e, e') = \frac{1}{D} \sum_{j=1}^{D} \left(\Sigma_{ee'}\right)_{jj}, \tag{3}$$

$$\Sigma_{ee'} = \mathbb{E}[(e - \mathbb{E}[e])(e' - \mathbb{E}[e'])^T], \tag{4}$$

where D is the embedding dimension and $(\Sigma_{ee'})_{jj}$ denotes diagonal elements. This approximation keeps diagonal terms and discards off-diagonal terms, which show negligible contribution in our ablations.

3.2.3. Multi-branch Detector and Fusion

Based on the robustness asymmetry observation, fake images differ from real images not only in high-level semantics(Ojha et al., 2023), but also in how their representations

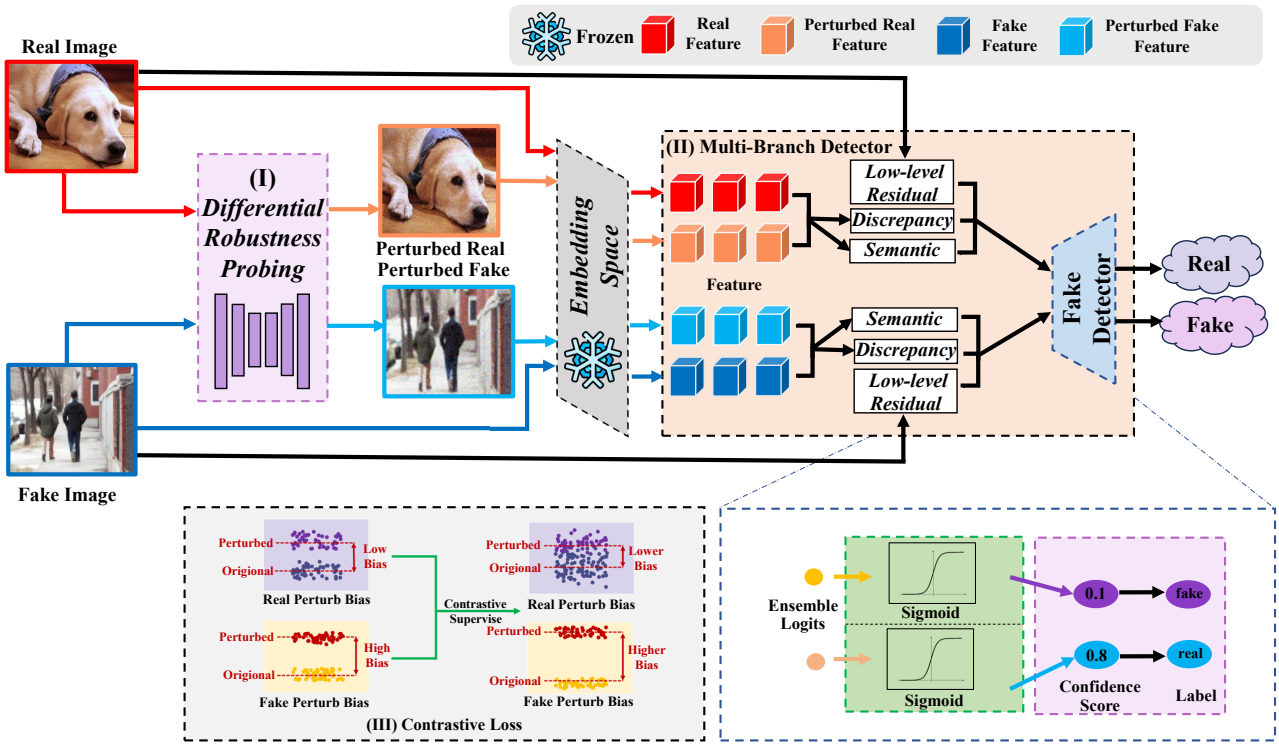


Figure 3. **Overview of RA-Det.** The framework processes real and fake images through three key phases to identify generative content: **(I) Differential Robustness Probing (DRP)** generates targeted perturbations to amplify the robustness asymmetry between real and synthetic images. **(II) The Multi-Branch Detector** aggregates complementary forensic evidence by analyzing high-level semantic features, quantifying feature-space stability via the discrepancy branch, and capturing pixel-level artifacts through the low-level residual branch. **(III) Contrastive Loss** guides the training by enforcing a margin between the feature displacements of real and fake images.

change under small, meaning-preserving perturbations, and in subtle low-level artifacts that can be hard to notice visually. RA-Det is therefore built as a multi-branch detector that gathers these complementary cues—semantic evidence, feature discrepancies, and pixel-level residual traces—and aggregates them for the final decision.

Semantic branch: The frozen foundation encoder extracts an embedding from the input image, and a lightweight classification head maps this embedding to a logit, providing semantic evidence for real vs. fake.

Discrepancy branches: Two additional branches capture how stable the embedding is under perturbation: one MLP takes the DCA distance $d = \|e - e'\|_2$ between clean and perturbed embeddings, and another MLP takes the vector difference $\Delta = e - e'$, so the detector can model both the amount and the pattern of change.

Low-level residual branch: Because the foundation encoder emphasizes high-level semantics and layout, RA-Det also uses a pixel-space residual stream to capture subtle synthesis traces. A median filter with a kernel size of 3 produces $\tilde{x}$, the residual map is computed as $r = x - \tilde{x}$, and a lightweight CNN trained from scratch takes r as input to output a complementary logit. All branch logits are then aggregated to produce the final prediction.

3.3. Training Objective

RA-Det is trained with a composite objective:

$$\mathcal{L}_{\text{comp}} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{ra}}, \tag{5}$$

where $\mathcal{L}_{\text{bce}}$ is the standard binary cross-entropy loss, and $\mathcal{L}_{\text{ra}}$ is a hinge-style contrastive loss that separates the similarity statistics of real and fake samples under perturbations.

Given a sample x_i with a perturbed view $x_i' = x_i + \delta_i$, we extract embeddings $e_i = f(x_i)$ and $e_i' = f(x_i')$, and compute the cosine similarity

$$s_i = \cos(e_i, e_i'). \tag{6}$$

For a mini-batch B, we split indices into real and fake subsets, $B_{\text{real}} = \{i \in B \mid y_i = 1\}$ and $B_{\text{fake}} = \{i \in B \mid y_i = 0\}$, and define their mean similarities as

$$s_{\text{real}} = \mathbb{E}_{i \in B_{\text{real}}}[s_i], \qquad s_{\text{fake}} = \mathbb{E}_{i \in B_{\text{fake}}}[s_i]. \tag{7}$$

We then enforce a margin γ between these batch-level statistics via

$$\mathcal{L}_{\text{ra}} = \text{ReLU}((s_{\text{fake}} - s_{\text{real}}) + \gamma), \tag{8}$$

which penalizes violations of the desired inequality $s_{\text{real}} \geq s_{\text{fake}} + \gamma$ and becomes zero once the margin is satisfied. Unless otherwise stated, we set $\gamma = 0.1$ in all experiments.

4. Lower-Bound for the Robustness Asymmetry

In this section, we provide a theoretical lower bound of the robustness gap between real and fake images under our feature–shift operator. We start from the SIDE memorization divergence (Chen et al., 2024b), which quantifies the memorization behavior of a generative model (the lower the divergence, the more the generative model memorizes). Building on this, we show that the expected feature shift induced by a small, isotropic perturbation is tightly linked to the memorization of the generative models around training examples. This leads to a lower bound on the robustness difference between real and fake images.

Notation. Let $\mathcal{X} \subset \mathbb{R}^n$ be the input domain; p the distribution of real images; p_θ the distribution of model-generated images; and $D = \{x_i\}_{i=1}^N \subset \mathcal{X}$ the training set. A *fixed* C^2 *encoder* is a feature map $f : \mathcal{X} \to \mathbb{R}^d$ whose parameters are held constant (no further training) and that is twice continuously differentiable with respect to the input. Write $J_f(x) \in \mathbb{R}^{d\times n}$ for the Jacobian of f at x and define the Jacobian energy $G(x) := \|J_f(x)\|_F^2$. Assume G is bounded: $0 \leq G(x) \leq B$ on $\mathcal{X}$. For a small, isotropic probe $\delta \in \mathbb{R}^n$ with $\mathbb{E}[\delta] = 0$ and $\mathbb{E}[\delta\delta^\top] = (\varepsilon^2/n)I_n$, define the feature-shift

$$\text{Shift}_\varepsilon(x) := \mathbb{E}_\delta\left[\|f(x+\delta) - f(x)\|_2^2\right]. \tag{9}$$

All expectations $\mathbb{E}_\mu[\cdot]$ below are with respect to $x \sim \mu$ unless stated otherwise.

Memorization divergence from SIDE (Chen et al., 2024b). Fix a radius $\varepsilon_0 > 0$ and define the training-neighborhood mixture

$$q_{\varepsilon_0}(x) := \frac{1}{N}\sum_{i=1}^N \mathcal{N}\big(x \,\big|\, x_i, \varepsilon_0^2 I_n\big). \tag{10}$$

The SIDE memorization divergence is

$$M(D; p_\theta, \varepsilon_0) := D_{\text{KL}}(q_{\varepsilon_0} \,\|\, p_\theta), \tag{11}$$

which is smaller when the generative model memorizes more training data. Define

$$\Delta \;:=\; \mathbb{E}_{x\sim q_{\varepsilon_0}}[G(x)] \;-\; \mathbb{E}_{x\sim p}[G(x)]. \tag{12}$$

Here $\Delta > 0$ expresses that the encoder's Jacobian energy is higher around training neighborhoods than under the real-data distribution.

Assumption Following established and recent theories/evidence on augmentation-induced invariance and architectural equivariance in encoders (Chen et al., 2019; Hounie et al., 2023a; Oquab et al., 2023; Hounie et al., 2023b; Rojas-Gomez et al., 2024; Xu et al., 2023; Caron et al., 2021), we make Assumption 4.1.

Assumption 4.1. The data lie (locally) on an m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^n$ and that a fixed C^2 encoder f exhibits reduced local sensitivity along tangent directions and comparatively larger sensitivity along normal directions.

Rationale. This assumption reflects the widely used manifold view of natural images, where high-dimensional observations concentrate near a lower-dimensional submanifold of ($\mathbb{R}^n$). Data augmentation and tangent-based regularization encourage invariance along typical on-manifold transformations (Hounie et al., 2023a;b), which manifests as smaller Jacobian norms in tangent directions, while off-manifold perturbations are not similarly regularized. Architectural constraints further support this pattern: group-equivariant networks explicitly tie representation changes to symmetry actions, reducing sensitivity along symmetry-induced tangents (Chen et al., 2019; Rojas-Gomez et al., 2024; Xu et al., 2023). Finally, modern self-supervised encoders (e.g., DINO/DINOv2) empirically yield robust, transferable features consistent with augmentation-induced invariance, providing contemporary evidence for the anisotropy posited here (Oquab et al., 2023; Caron et al., 2021). Based on this assumption, we derive the following Lemma 4.2.

Lemma 4.2 (Small-radius positive margin)**.** *Under the assumption above, there exists $\bar{\varepsilon}_0 > 0$ and a constant $c_0 > 0$ (depending only on the encoder's anisotropy margin and local tube size) such that for all $\varepsilon_0 \in (0, \bar{\varepsilon}_0]$,*

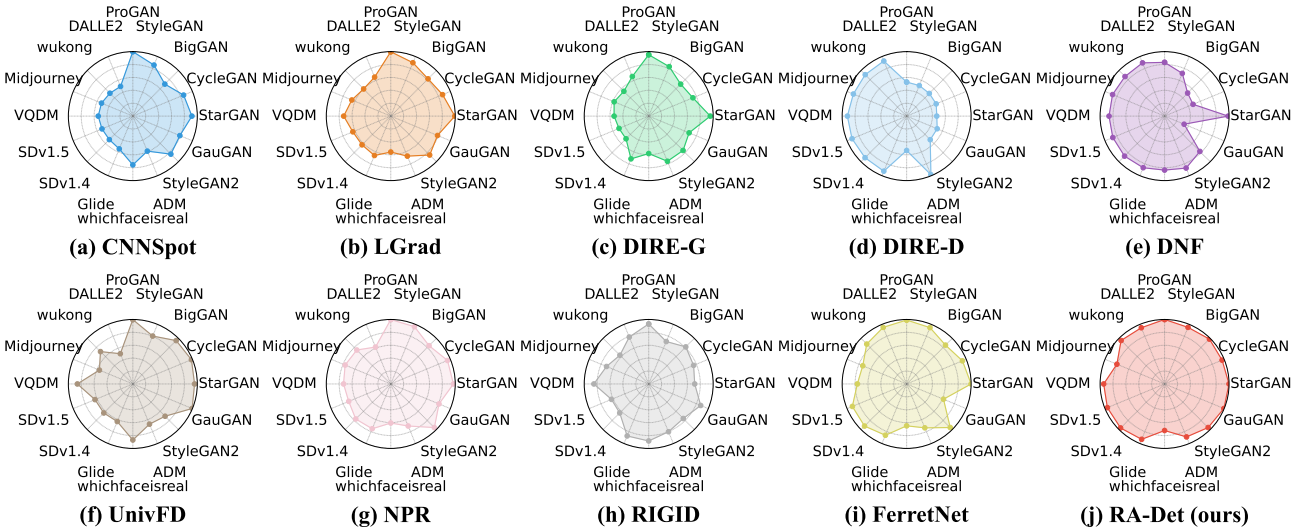
$$\Delta \;\geq\; c_0 \;>\; 0.$$

We prove Lemma 4.2 to ground the detector's core premise in theory: modern encoders are trained/built to be insensitive along natural, on-manifold transformations but more sensitive off-manifold; thus a small Gaussian "tube" around training samples puts probability mass where the encoder's Jacobian energy is higher, making $\mathbb{E}_{q_{\varepsilon_0}}[G] > \mathbb{E}_p[G]$ and yielding a strictly positive margin ($\Delta > 0$) that drives the robustness gap used in Theorem 1.

Theorem 4.3 (Lower bound on the shift gap)**.** *Under the setup and Lemma 4.2 above, for sufficiently small $\varepsilon > 0$,*

$$\begin{aligned}\mathbb{E}_{x\sim p_\theta}\big[\text{Shift}_\varepsilon(x)\big] \;-\; \mathbb{E}_{x\sim p}\big[\text{Shift}_\varepsilon(x)\big] \;\geq\\ \frac{\varepsilon^2}{n}\Big(\Delta \;-\; B\sqrt{M(D; p_\theta, \varepsilon_0)\,/2}\Big) \;+\; O(\varepsilon^4).\end{aligned} \tag{13}$$

More specifically, Theorem 4.3 shows that the expected shift under p_θ is lower-bounded by a term that decreases



(a) CNNSpot (b) LGrad (c) DIRE-G (d) DIRE-D (e) DNF

(f) UnivFD (g) NPR (h) RIGID (i) FerretNet (j) RA-Det (ours)

Figure 4. Radar chart comparison of multiple detectors.

with the SIDE divergence $M(D; p_\theta, \varepsilon_0)$. In other words, more memorization of the training data provably creates a larger robustness asymmetry in the small-noise regime. Prior works on extraction and membership inference in different generative models demonstrate that such memorization is prevalent across GANs, diffusion models, and even the large language models (Hayes et al., 2019; Chen et al., 2020; Hilprecht et al., 2019; Carlini et al., 2023; 2021). This explains why the robustness asymmetry is *universal* because memorization is prevalent across a wide range of generative models types. The proof and empirical validation of the lower bound are in Appendix B.

5. Experiments

5.1. Experimental Setup

5.1.1. Datasets

To better simulate real-world black-box detection scenarios, we follow the protocol established by Wang et al. restricting training to one generative model while evaluating generalization on unseen models. Specifically, during the training phase, we use only 360k real images and 360k synthetic images generated by ProGAN (Karras, 2017). The evaluation phase, however, is conducted on ProGAN and 15 unseen generative models or platforms provided by Zhong et al. (2023). The test set includes both facial images and other semantic categories, covering generative models and platforms such as GAN-based models, diffusion-based models, and platform-generated content, ensuring a robust and diverse assessment of the model's generalization ability.

5.1.2. Implementation Details

RA-Det is implemented in PyTorch using DINOv3 ViT-L/16 as the fixed backbone. All images are resized to 256×256 and randomly cropped to 224×224 during training, while center cropping is applied at testing. We train the perturbation generator and detection head for 10 epochs using the Adam optimizer with an initial learning rate of 1×10^{-5} and a batch size of 32. JPEG compression and blur augmentations are disabled to ensure that robustness asymmetry arises solely from the probing module. All experiments are conducted on four 4090 NVIDIA GPUs.

5.1.3. Compared Methods

We classify existing methods into three categories: **GAN-specific detectors**, such as FreDect (Frank et al., 2020) and LGrad (Tan et al., 2023); **Diffusion-specific detectors**, including DIRE (Wang et al., 2023) and DNF (Zhang & Xu, 2024); and **Universal detectors**, like CNNSpot (Wang et al., 2020), GramNet (Liu et al., 2020), UniFD (Ojha et al., 2023), NPR (Tan et al., 2024), RIGID (He et al., 2024) and FerretNet (Liang et al., 2025).

5.1.4. Evaluation Metrics

Following the protocol established by PatchCraft (Zhong et al., 2023), we adopt Accuracy (ACC) and Average Precision (AP), two widely used metrics in generative image detection (Wang et al., 2020; Ojha et al., 2023; Tan et al., 2024). These metrics provide a comprehensive evaluation of detection performance. We report both values for each method across all datasets and compute the mean performance to assess generalization.

5.2. Comparison of Detection Performance

We evaluate RA-Det under a comprehensive cross-generator setting, covering 16 diverse generative models and more than 10 representative detectors, including GAN-specific, diffusion-specific, and universal approaches. As reported

Table 1. The detection performance comparison between our approach and baselines. We use ACC(%)/AP(%) as the evaluation metrics. DIRE-D denotes the DIRE detector trained over fake images from ADM, while DNF uses a model trained on ADM, similar to DIRE-D. DIRE-G denotes the DIRE detector trained over the same training set (ProGAN) as others. Among all detectors, the best result and the second-best result are denoted in boldface and underlined, respectively.

Generator	GAN Detector			Diffusion Detector		Universal Detector						
	FreDect	LGrad	DIRE-G	DIRE-D	DNF	CNNSpot	GramNet	UniFD	NPR	RIGID	FerretNet	RA-Det (ours)
ProGAN	99.36/99.99	99.78/99.99	95.19/99.08	52.75/58.79	83.35/60.40	**99.99/100.00**	99.99/**100.00**	99.81/**100.00**	99.90/99.98	93.00/96.83	99.09/99.95	99.98/**100.00**
StyleGAN	78.02/88.98	89.63/98.03	83.03/91.74	51.31/56.68	71.76/44.18	85.71/99.54	83.59/94.49	80.40/97.48	**96.06/99.78**	71.19/70.61	94.53/98.63	94.98/98.90
BigGAN	81.98/93.62	81.73/89.08	70.12/75.25	49.70/46.91	50.55/51.75	70.19/84.51	67.33/81.79	95.08/99.27	83.95/85.59	81.25/85.69	85.18/91.06	**98.35/99.83**
CycleGAN	78.77/84.78	86.94/95.01	74.19/80.56	49.58/50.03	47.77/66.49	85.20/93.48	86.07/95.33	**98.33/99.80**	95.19/98.12	75.34/83.04	93.64/98.34	96.52/99.47
StarGAN	94.62/99.49	99.27/**100.00**	95.47/99.34	46.72/40.64	98.55/86.44	91.62/98.15	95.05/99.23	95.75/99.37	97.17/**100.00**	71.34/77.64	**99.97/100.00**	99.92/**100.00**
GauGAN	80.56/82.84	78.46/95.43	67.79/72.15	51.23/47.34	32.78/39.48	78.93/89.49	69.35/84.99	99.47/**99.98**	80.94/82.97	87.26/93.71	61.84/59.71	**99.51**/99.77
StyleGAN2	66.19/82.54	85.32/97.89	75.31/88.29	51.72/58.03	77.42/40.16	83.39/99.05	87.28/99.11	70.76/97.71	**95.61/99.95**	75.78/80.78	95.56/99.16	95.54/99.15
whichfaceisreal	50.75/55.85	55.70/57.99	58.05/60.13	53.30/59.02	83.70/52.64	75.65/83.11	72.70/94.22	86.90/**96.73**	60.60/62.91	**88.40**/95.06	64.95/70.36	72.20/81.13
ADM	63.40/61.77	67.15/72.95	75.78/85.84	**98.25/99.79**	87.22/83.72	58.78/71.07	58.61/73.11	67.46/89.80	70.58/75.08	80.22/85.88	73.65/83.23	88.89/95.07
Glide	54.13/52.91	66.11/80.42	71.75/78.35	92.42/**99.54**	86.62/95.18	55.00/66.16	54.50/66.76	63.09/83.81	75.18/83.24	87.26/93.89	86.12/93.62	**92.94**/97.83
SDv1.4	38.79/37.83	63.02/62.37	49.74/49.87	91.24/98.61	87.77/79.22	51.55/56.88	51.70/59.83	63.66/86.14	76.81/82.79	63.57/65.59	92.33/97.20	**96.50/99.08**
SDv1.5	39.21/37.76	63.67/62.85	49.83/49.52	91.63/98.83	86.86/77.54	51.78/57.24	52.16/60.37	63.49/85.84	70.58/83.80	62.50/65.45	91.13/96.64	**96.01/98.92**
VQDM	77.80/85.10	72.99/77.47	53.68/54.57	91.90/**98.98**	86.15/77.14	53.67/61.92	52.86/61.13	86.01/96.53	73.22/77.74	85.01/90.93	76.44/85.23	**94.43**/98.68
Midjourney	45.87/46.09	65.35/71.86	58.01/61.86	**89.45/97.32**	87.32/68.09	52.59/55.90	50.02/56.82	56.13/74.00	76.61/82.61	70.80/78.63	73.66/80.27	80.04/87.53
wukong	40.30/39.58	59.55/62.48	54.46/55.38	90.90/**98.37**	86.80/87.32	50.23/52.85	50.76/55.62	70.93/91.07	74.45/78.17	63.09/67.10	87.48/94.72	**95.33**/98.33
DALLE2	34.67/38.22	65.45/82.55	66.48/74.48	92.45/**99.71**	89.39/89.35	49.82/50.59	49.25/49.82	50.75/63.04	61.88/71.64	78.74/87.28	**94.95**/98.82	94.30/98.14
Average	64.03/68.28	75.34/84.35	68.68/74.13	71.53/75.87	77.75/68.69	68.38/76.39	67.58/76.93	78.00/ 89.73	80.55/84.05	77.17/82.38	85.66/90.43	**93.47/97.00**

in Table 1, RA-Det achieves the best overall performance, reaching an average accuracy of **93.47%** and an average AP of **97.00%**. This result surpasses the strongest universal baseline, FerretNet (85.66% / 90.43%), by a substantial margin, demonstrating a clear advantage in both accuracy and ranking quality.

In contrast, existing detectors exhibit clear limitations in generalization. Appearance-driven or architecture-specific methods often perform well only within their target domains, such as GAN-based or diffusion-based generators, but degrade noticeably when evaluated on unseen or mismatched models. Universal detectors, including CNNSpot, GramNet, and UniFD, improve cross-model robustness to some extent, yet their reliance on static appearance cues or representation similarity restricts stability under diverse generative processes. NPR and FerretNet, which exploit local pixel dependencies introduced by upsampling operations, achieve competitive performance on several generators, supporting the presence of structured artifacts in synthetic images. However, their effectiveness remains uneven across models, particularly for more challenging or stylistically diverse generators.

Compared with the most related behavior-based baseline, RIGID, which captures a robustness gap by adding Gaussian noise and applying a hand-crafted similarity threshold in a fixed backbone, RA-Det instead learns a decision function on robustness-asymmetry features induced by controlled perturbations. This behavior-driven design yields a much stronger universal detector: RA-Det improves the average ACC/AP from 77.17% / 82.38% for RIGID to 93.47% / 97.00%, showing that turning robustness asymmetry into a learnable decision signal is substantially more effective than threshold-based scoring.

5.3. Ablation Study

We conduct ablation experiments to quantify the contribution of the frozen foundation backbone, the robustness-aware training objective, and each branch in the multi-branch detector of RA-Det. Table 2 summarizes all variants. For readability, **Sem/Dis/Res** denote the semantic branch, discrepancy branch, and low-level residual branch, respectively, where the discrepancy branch is instantiated by the embedding distance d and difference vector Δ defined in Sec. 3.2.3.

Backbone. To verify that RA-Det is not tied to a specific representation space, we replace the frozen backbone while keeping $\mathcal{L}_{\text{comp}}$ and the detector architecture unchanged. Replacing DINOv3 ViT-L/16 with CLIP ViT-L/14 still yields strong performance (91.85/95.00 Acc/AP), indicating that robustness-asymmetry cues can be captured across different foundation encoders. Meanwhile, DINOv3 achieves the best results (93.47/97.00), suggesting that different backbones provide feature spaces with different suitability for exposing robustness-driven discrepancies.

Training objective. We ablate the robustness-aware component in $\mathcal{L}_{\text{comp}}$ by training with $\mathcal{L}_{\text{bce}}$ only (Loss $\rightarrow$ BCE). This reduces performance to 90.86/95.26, confirming that the robustness-aware supervision is essential for amplifying the behavioral gap between real and fake samples under DRP-induced perturbations.

Multi-branch detector. We further evaluate the contribution of each branch in Sec. 3.2.3. Removing the semantic branch (w/o Sem) slightly degrades performance to 90.34/95.43, showing that high-level semantics remain helpful. In contrast, removing the discrepancy branch (w/o Dis) causes the largest drop (85.73/91.52), highlighting that robustness-driven behavioral discrepancies form the core ev-

Table 2. Ablation on backbone, training objective, and the multi-branch detector in Sec. 3.2.3. **Sem/Dis/Res** denote the semantic branch, discrepancy branch, and low-level residual branch, respectively. Best in bold.

Variant	Backbone	Loss	Acc (%)	AP (%)
Backbone / Objective				
Full (Ours)	DINOv3 ViT-L/16	$\mathcal{L}_{\text{comp}}$	**93.47**	**97.00**
Backbone → CLIP	CLIP ViT-L/14	$\mathcal{L}_{\text{comp}}$	91.85	95.00
Loss → BCE	DINOv3 ViT-L/16	$\mathcal{L}_{\text{bce}}$	90.86	95.26
Multi-branch detector				
w/o Sem	DINOv3 ViT-L/16	$\mathcal{L}_{\text{comp}}$	90.34	95.43
w/o Dis	DINOv3 ViT-L/16	$\mathcal{L}_{\text{comp}}$	85.73	91.52
w/o Res	DINOv3 ViT-L/16	$\mathcal{L}_{\text{comp}}$	88.62	94.26
Res: Med → Gau	DINOv3 ViT-L/16	$\mathcal{L}_{\text{comp}}$	90.58	95.63

idence in RA-Det. Removing the low-level residual branch (w/o Res) also harms performance (88.62/94.26), indicating that pixel-space residual cues complement foundation-space representations. Finally, replacing the median-based residual construction with a Gaussian-based alternative (Res: Med → Gau) decreases performance to 90.58/95.63, suggesting that the proposed residual design better preserves detection-relevant signals while suppressing nuisance variations.

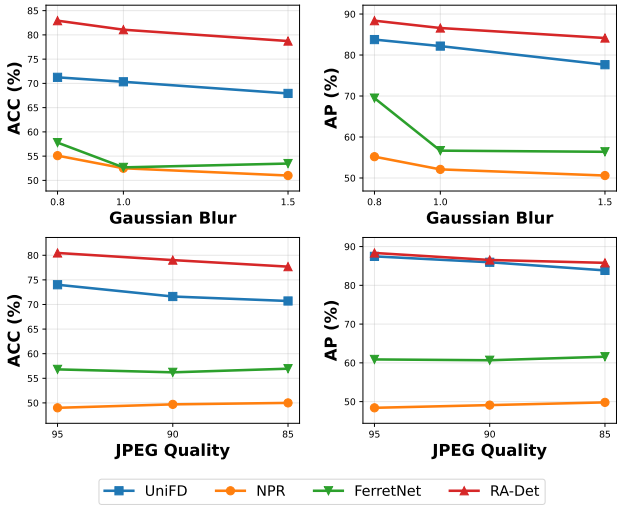


Figure 5. Performance comparison under common image perturbations. The plots show detector performance against (a) JPEG compression (QF = 95, 90 and 85) and (b) Gaussian blur (σ = 0.8, 1.0 and 1.5). The mean scores (diamonds) indicate that our method, RA-Det, maintains the highest accuracy and AP, demonstrating superior robustness.

5.4. Robustness Testing

In real-world scenarios, generative image detection systems inevitably encounter post-processing operations such as JPEG compression and Gaussian blur during image transmission and storage. Evaluating robustness under such perturbations is therefore essential for practical deployment. We conduct robustness experiments under controlled JPEG compression (QF = 95, 90, 85) and Gaussian blur (σ = 0.8, 1.0, 1.5), and compare RA-Det with representative artifact-driven detectors (NPR and FerretNet) and the representation-driven baseline UniFD.

As illustrated in Fig. 5, artifact-driven methods exhibit substantial performance degradation as perturbation strength increases. Under Gaussian blur, NPR drops from 55.1% to 51.0% ACC, and FerretNet decreases from 57.79% to 53.47%. In contrast, UniFD shows a more moderate decline (71.25% to 67.93%), while RA-Det maintains consistently high accuracy (82.93% to 78.73%). A similar trend is observed under JPEG compression, where NPR and FerretNet remain around 50%–57% ACC, whereas UniFD and RA-Det sustain stable performance with limited degradation across QF levels. Notably, RA-Det achieves the highest ACC and AP across all perturbation strengths.

These results reveal a clear robustness gap between artifact-driven detectors and feature-level modeling approaches. Methods such as NPR and FerretNet rely heavily on fragile local pixel statistics and generation artifacts, which are easily disrupted by compression and blur. In contrast, UniFD and RA-Det operate in foundation feature space, where semantic representations are more stable under mild perturbations. By explicitly leveraging robustness asymmetry and aggregating semantic, discrepancy, and residual cues through a multi-branch design, RA-Det further enhances stability and consistently outperforms existing approaches under post-processing operations.

6. Conclusion

In this work, rather than following previous approaches that focus on identifying universal artifacts of fake images, we propose an innovative method that distinguishes real and generated images based on their inherent differences in robustness against perturbations. Experimental results demonstrate that our method not only surpasses previous artifact-based approaches but also exhibits strong generalization ability across multiple generative models. This validates that leveraging the disparity in robustness helps prevent the detector from merely learning superficial features of the training data while failing to truly distinguish real and generated images. Furthermore, robustness experiments confirm that this assumption remains valid even under image distortions, ensuring that the detection model does not fail due to minor degradations in image quality. In future work, we plan to investigate whether the robustness difference between real and fake images is a general phenomenon that extends to more diverse forms of perturbations, such as adversarial attacks.

References

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Bender, E., Litwin, N., Erlingsson, Ú., Oprea, A., and Papernot, N. Extracting training data from large language models. In *USENIX Security Symposium*, 2021. URL https://www.usenix.org/system/files/sec21-carlini-extracting.pdf.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *USENIX Security Symposium*, 2023. URL https://www.usenix.org/system/files/usenixsecurity23-carlini.pdf.

Caron, M., Touvron, H., Misra, I., et al. Emerging properties in self-supervised vision transformers. In *ICCV 2021*, 2021. URL https://openaccess.thecvf.com/content/ICCV2021/papers/Caron_Emerging_Properties_in_Self-Supervised_Vision_Transformers_ICCV_2021_paper.pdf.

Chen, B., Zeng, J., Yang, J., and Yang, R. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=oRLwyayrh1.

Chen, D., Yu, N., Zhang, Y., and Fritz, M. GAN-Leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 27th ACM Conference on Computer and Communications Security (CCS)*, 2020. URL https://yangzhangalmo.github.io/papers/CCS20-GAN-Leaks.pdf.

Chen, S., Luss, R., Vaidyanathan, R., and Aravkin, A. A group-theoretic framework for data augmentation. *arXiv preprint arXiv:1907.10905*, 2019. URL https://arxiv.org/abs/1907.10905.

Chen, Y., Wang, X., Zou, J., and Ma, R. Surrogate conditional data extraction from diffusion models (side). *arXiv preprint arXiv:2410.02467*, 2024b. URL https://arxiv.org/abs/2410.02467.

Chen, Y., Wang, S., Wang, X., and Ma, X. Leaky-clip: Extracting training data from clip. *arXiv preprint arXiv:2508.00756*, 2025.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pp. 3247–3258. PMLR, 2020.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all, 2023. URL https://arxiv.org/abs/2305.05665.

Grommelt, P., Weiss, L., Pfreundt, F.-J., and Keuper, J. Fake or jpeg? revealing common biases in generated image detection datasets. In *European Conference on Computer Vision*, pp. 80–95. Springer, 2024.

Hayes, J., Melis, L., Danezis, G., and De Cristofaro, E. LOGAN: Membership inference attacks against generative models. In *2019 Privacy Enhancing Technologies Symposium (PoPETs)*, 2019. URL https://petsymposium.org/popets/2019/popets-2019-0008.pdf.

He, Z., Chen, P.-Y., and Ho, T.-Y. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection, 2024. URL https://arxiv.org/abs/2405.20112.

Hilprecht, B., Härterich, M., and Bernau, D. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies*, 2019(4):232–249, 2019. URL https://petsymposium.org/popets/2019/popets-2019-0067.pdf.

Hounie, I., Chamon, L. F., and Ribeiro, A. Automatic data augmentation via invariance-constrained learning. In *International Conference on Machine Learning*, pp. 13410–13433. PMLR, 2023a.

Hounie, I. et al. Automatic data augmentation via invariance-constrained learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023b. URL https://proceedings.mlr.press/v202/hounie23a/hounie23a.pdf.

Karras, T. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/papers/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.pdf.

Le, B. M. and Woo, S. S. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images, 2021. URL https://arxiv.org/abs/2112.03553.

Liang, S., Liu, J., Chen, R., and Guan, Q. Ferretnet: Efficient synthetic image detection via local pixel dependencies. *arXiv preprint arXiv:2509.20890*, 2025.

Lin, L. et al. Detecting multimedia generated by large ai models: A survey, 2024.

Liu, Z., Qi, X., and Torr, P. Global texture enhancement for fake face detection in the wild, 2020. URL https://arxiv.org/abs/2002.00133.

Midjourney. https://www.midjourney.com/home/, 2023.

Nightingale, S. J. and Farid, H. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *PNAS*, 119(8):e2120481119, 2022. doi: 10.1073/pnas.2120481119.

Ojha, U., Li, Y., and Lee, Y. J. Towards universal fake image detectors that generalize across generative models. In *CVPR*, 2023.

Oquab, M., Darcet, T., Moutakanni, T., et al. Dinov2: Learning robust visual features without supervision. *arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., et al. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Rojas-Gomez, R. A. et al. Making vision transformers truly shift-equivariant. *arXiv:2305.16316*, 2024. URL https://arxiv.org/abs/2305.16316.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/papers/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.pdf.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

Runway. Runway (gen-3 alpha). Software platform, 2024. Accessed: 2025-04-01.

Saharia, C., Chan, W., Saxena, S., et al. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. URL https://papers.neurips.cc/paper_files/paper/2022/file/ec795aeadae0b7d230fa35cbaf04c041-Paper-Conference.pdf.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., and Bojanowski, P. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

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

Tan, C., Liu, H., Zhao, Y., Wei, S., Gu, G., Liu, P., and Wei, Y. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28130–28139, 2024. doi: 10.1109/CVPR52733.2024.02657.

Tan, C., Tao, R., Liu, H., Gu, G., Wu, B., Zhao, Y., and Wei, Y. 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.

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020.

Wang, T., Zhang, Y., Qi, S., Zhao, R., Xia, Z., and Weng, J. Security and privacy on generative data in aigc: A survey. *ACM Comput. Surv.*, 57(4), December 2024. ISSN 0360-0300. doi: 10.1145/3703626. URL https://doi.org/10.1145/3703626.

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., and Li, H. Dire for diffusion-generated image detection, 2023. URL https://arxiv.org/abs/2303.09295.

Wu, H., Zhou, J., and Zhang, S. Generalizable synthetic image detection via language-guided contrastive learning, 2025. URL https://arxiv.org/abs/2305.13800.

Xu, R., Yang, K., Liu, K., and He, F. $e(2)$-equivariant vision transformer. In *Proceedings of UAI 2023*. PMLR,

2023. URL https://proceedings.mlr.press/v216/xu23b/xu23b.pdf.

Yu, P., Fei, J., Gao, H., Feng, X., Xia, Z., and Chang, C. H. Unlocking the capabilities of large vision-language models for generalizable and explainable deepfake detection. *arXiv preprint arXiv:2503.14853*, 2025.

Zhang, Y. and Xu, X. Diffusion noise feature: Accurate and fast generated image detection, 2024. URL https://arxiv.org/abs/2312.02625.

Zhong, N., Xu, Y., Qian, Z., and Zhang, X. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv preprint arXiv:2311.12397*, 2023.

A. On DIRE and Dataset–Format Biases

The notably high performance of DIRE on diffusion-generated images is largely enabled by unintended dataset biases rather than purely generator-specific artifacts. DIRE(Wang et al., 2023) is trained on a benchmark where synthetic images are stored in lossless PNG format and often at fixed resolutions, while real photographs are saved in JPEG format with varying compression levels and sizes. In effect, the detector learns to pick up on format cues (for example, JPEG compression artifacts or size/resolution uniformity (Grommelt et al., 2024)) rather than genuine generative fingerprints. When these format differences are removed or equalized, its performance degrades substantially and cross-generator transfer drops sharply

B. Lower-Bound for the Robustness Asymmetry

B.1. Theoretical Derivation

We derive a lower bound for the robustness gap between real and model-generated images under a small, isotropic probe in feature space. We begin from the SIDE memorization divergence (Chen et al., 2024b), which quantifies generator memorization around training samples: smaller divergence indicates stronger memorization. We then connect memorization to the expected feature shift under small perturbations, yielding a quantitative lower bound on the robustness difference.

Notation. Let the ambient space be $\mathcal{X} \subset \mathbb{R}^N$; p the real image distribution; p_θ the model-generated distribution; and $D = \{x_i\}_{i=1}^{N_{\text{tr}}} \subset \mathcal{X}$ the generator's training set. A *fixed C^2 encoder* is a twice continuously differentiable map $f : \mathcal{X} \to \mathbb{R}^d$ whose parameters are frozen. Denote the Jacobian by $J_f(x) \in \mathbb{R}^{d\times N}$ and the Jacobian energy by $G(x) := \|J_f(x)\|_F^2$, assumed bounded on $\mathcal{X}$: $0 \le G(x) \le B$. For a small, isotropic probe $\boldsymbol{\eta} \in \mathbb{R}^N$ with $\mathbb{E}[\boldsymbol{\eta}] = 0$ and $\mathbb{E}[\boldsymbol{\eta}\boldsymbol{\eta}^\top] = (\varepsilon^2/N)I_N$, define the feature shift

$$\text{Shift}_\varepsilon(x) := \mathbb{E}_{\boldsymbol{\eta}}\big[\|f(x+\boldsymbol{\eta}) - f(x)\|_2^2\big]\,. \tag{14}$$

All expectations $\mathbb{E}_\mu[\cdot]$ below are with respect to $x \sim \mu$ unless stated otherwise.

Memorization divergence (SIDE) (Chen et al., 2024b; 2025). Fix a radius $\varepsilon_0 > 0$ and define the training-neighborhood mixture

$$q_{\varepsilon_0}(x) := \frac{1}{N_{\text{tr}}}\sum_{i=1}^{N_{\text{tr}}} \mathcal{N}\Big(x \,\Big|\, x_i, \varepsilon_0^2 I_N\Big). \tag{15}$$

SIDE memorization is

$$M(D; p_\theta, \varepsilon_0) := D_{\text{KL}}(q_{\varepsilon_0} \,\|\, p_\theta)\,, \tag{16}$$

smaller when p_θ puts more mass near training examples. Define

$$\Delta \;:=\; \mathbb{E}_{x\sim q_{\varepsilon_0}}[G(x)] \;-\; \mathbb{E}_{x\sim p}[G(x)]. \tag{17}$$

Tangent/normal anisotropy (modeling hypothesis). Motivated by recent evidence that augmentation and architectural equivariance induce on-manifold invariances in modern encoders (Oquab et al., 2023; Caron et al., 2021; Rojas-Gomez et al., 2024; Xu et al., 2023), we *refer to* Assumption 4.1 (stated elsewhere): natural images lie locally on an m-dimensional manifold $\mathcal{M} \subset \mathbb{R}^N$ and a fixed C^2 encoder f has smaller local sensitivity along $T_x\mathcal{M}$ and larger sensitivity along $N_x\mathcal{M}$.

Rationale. Augmentation/equivariance suppresses representation change along common on-manifold transformations (tangents), while off-manifold perturbations are not explicitly regularized (Oquab et al., 2023; Caron et al., 2021; Hounie et al., 2023b; Xu et al., 2023). Thus, a small Gaussian tube around training samples places mass where G tends to be larger, leading to $\Delta > 0$ for sufficiently small radius.

Derivation for Lemma 4.2. Let $U = \{x : \text{dist}(x, \mathcal{M}) < \bar{\varepsilon}_0\}$ be a tubular neighborhood of $\mathcal{M}$ with smooth nearest-point projection $\pi : U \to \mathcal{M}$. Fix $y \in \mathcal{M}$ and write the *ambient* isotropic noise as $\boldsymbol{\eta} \sim \mathcal{N}(0, \varepsilon_0^2 I_N/N)$, so $\mathbb{E}\|\boldsymbol{\eta}\|^2 = \varepsilon_0^2$. Let $P_T(y)$ and $P_N(y) = I - P_T(y)$ be the orthogonal projectors onto $T_y\mathcal{M}$ and $N_y\mathcal{M}$, and decompose

$$t := P_T(y)\boldsymbol{\eta} \in T_y\mathcal{M}, \qquad s := P_N(y)\boldsymbol{\eta} \in N_y\mathcal{M}, \qquad \boldsymbol{\eta} = t + s.$$

By orthogonality and isotropy,

$$\mathbb{E}\|t\|^2 = \tfrac{\varepsilon_0^2}{N}\mathrm{tr}P_T = \tfrac{\varepsilon_0^2}{N}m,\ \ \mathbb{E}\|s\|^2 = \tfrac{\varepsilon_0^2}{N}(N-m).$$

Since $f \in C^2$, $G(x) = \|J_f(x)\|_F^2$ is C^2 on U. A second-order Taylor expansion at y gives

$$G(y+\boldsymbol{\eta}) = G(y) + \nabla G(y)\cdot\boldsymbol{\eta} + \tfrac{1}{2}\,\boldsymbol{\eta}^\top H_G(y)\,\boldsymbol{\eta} + R(y,\boldsymbol{\eta}),$$

with $|R(y,\boldsymbol{\eta})| \le C\|\boldsymbol{\eta}\|^3$ for some uniform C (compactness of U). Taking expectation (mean-zero noise) removes the linear term:

$$\mathbb{E}\big[G(y+\boldsymbol{\eta})\big] - G(y) = \tfrac{1}{2}\,\mathbb{E}\big[t^\top H_G(y)\,t\big] + \tfrac{1}{2}\,\mathbb{E}\big[s^\top H_G(y)\,s\big] + O(\varepsilon_0^3).$$

Because the Assumption 4.1:

$$\lambda_T^{\max} := \sup_{\substack{y\in\mathcal{M}\\ \|u\|=1,\,u\in T_y\mathcal{M}}} u^\top H_G(y)\,u, \qquad \lambda_\perp^{\min} := \inf_{\substack{y\in\mathcal{M}\\ \|v\|=1,\,v\in N_y\mathcal{M}}} v^\top H_G(y)\,v, \qquad \lambda_\perp^{\min} > \lambda_T^{\max}.$$

Then

$$\mathbb{E}[t^\top H_G(y)t] \le \lambda_T^{\max}\,\mathbb{E}\|t\|^2, \qquad \mathbb{E}[s^\top H_G(y)s] \ge \lambda_\perp^{\min}\,\mathbb{E}\|s\|^2,$$

so

$$\mathbb{E}\big[G(y+\boldsymbol{\eta})\big] - G(y) \ \ge\ \tfrac{1}{2}\big(\lambda_\perp^{\min}\tfrac{N-m}{N} - \lambda_T^{\max}\tfrac{m}{N}\big)\,\varepsilon_0^2 - C'\varepsilon_0^3.$$

Averaging over the anchors $\{x_i\}$ that define q_{ε_0} and subtracting $\mathbb{E}_{x\sim p}[G(x)]$ yields $\Delta \ge c_1\,\varepsilon_0^2 - C'\varepsilon_0^3$ with $c_1 := \frac{1}{2}(\lambda_\perp^{\min}\frac{N-m}{N} - \lambda_T^{\max}\frac{m}{N}) > 0$. Choosing $\bar\varepsilon_0$ small so that $C'\bar\varepsilon_0 \le c_1/2$ gives $\Delta \ge c_0 := \frac{1}{2}c_1\,\varepsilon_0^2 > 0$ for all $\varepsilon_0 \in (0,\bar\varepsilon_0]$. □

Derivation for Theorem 4.3. **(i) Small-noise expansion.** By Taylor and isotropy (odd moments vanish) one gets

$$\mathrm{Shift}_\varepsilon(x) = \frac{\varepsilon^2}{n}\,G(x) + R_\varepsilon(x), \qquad |R_\varepsilon(x)| \le C_H\,\varepsilon^4, \tag{18}$$

where C_H depends on bounded Hessians and fourth moments of the probe (Isserlis/Wick; uniform-sphere moments). Averaging (18) over $\mu \in \{p_\theta, p\}$ and subtracting:

$$\mathbb{E}_{p_\theta}[\mathrm{Shift}_\varepsilon] - \mathbb{E}_p[\mathrm{Shift}_\varepsilon] = \frac{\varepsilon^2}{n}\Big(\mathbb{E}_{p_\theta}[G] - \mathbb{E}_p[G]\Big) + O(\varepsilon^4). \tag{19}$$

(ii) DV variational bound + Hoeffding. For any $\lambda > 0$,

$$M = D_{\mathrm{KL}}(q_{\varepsilon_0}\|p_\theta) \ \ge\ \lambda\,\mathbb{E}_{q_{\varepsilon_0}}[G] - \log\mathbb{E}_{p_\theta}[e^{\lambda G}],$$

and since $G \in [0,B]$, Hoeffding's lemma gives $\log\mathbb{E}_{p_\theta}[e^{\lambda G}] \le \lambda\,\mathbb{E}_{p_\theta}[G] + \frac{\lambda^2 B^2}{8}$. Optimizing the resulting quadratic over λ yields

$$\mathbb{E}_{p_\theta}[G] \ \ge\ \mathbb{E}_{q_{\varepsilon_0}}[G] \ -\ B\,\sqrt{M/2}. \tag{20}$$

(iii) Combine. Add and subtract $\mathbb{E}_{q_{\varepsilon_0}}[G]$ and use Δ from (17):

$$\mathbb{E}_{p_\theta}[G] - \mathbb{E}_p[G] \ \ge\ \Delta - B\sqrt{M/2}.$$

Insert into (19) to obtain (13). □

Interpretation. Lemma 4.2 formalizes that, for small tubes around training anchors, q_{ε_0} places mass where the encoder's Jacobian energy is (on average) larger than under p; thus $\Delta > 0$. Theorem 4.3 then states that the expected feature-shift under p_θ exceeds that under p by a margin controlled below by $\frac{\varepsilon^2}{n}\big(\Delta - B\sqrt{M/2}\big)$ up to $O(\varepsilon^4)$. Hence, stronger memorization (smaller M) and a positive local anisotropy margin (captured in Δ) provably increase the robustness asymmetry that our detector exploits.

B.2. Empirical Validation of the Shift Lower Bound

Goal. We empirically validate the theoretical prediction of Theorem 4.3, which links a model's memorization level to the observable gap between the shift statistics of generated and training samples. The theorem establishes two main trends:

(i) **Memorization effect.** When the model exhibits stronger memorization, the SIDE divergence $M(D; p_\theta, \varepsilon_0)$ decreases, resulting in a *larger lower bound* on the differential shift

$$\Delta(\varepsilon) = \mathbb{E}_{x\sim p_\theta}[\mathrm{Shift}_\varepsilon(x)] - \mathbb{E}_{x\sim p}[\mathrm{Shift}_\varepsilon(x)].$$

(ii) **Noise–dependence.** As the probing noise ε increases from zero, $\Delta(\varepsilon)$ initially grows because the added perturbation accentuates the model's sensitivity along tangent directions. However, for large ε, the perturbation dominates the manifold structure, leading to a decay in $\Delta(\varepsilon)$. Hence, the theoretical prediction is a *non-monotonic* dependence of $\Delta(\varepsilon)$: increasing at small ε, then decreasing after a critical point.

Proxy for memorization. Evaluating the SIDE divergence $M(D; p_\theta, \varepsilon_0)$ directly requires estimating high-dimensional manifold integrals, which is computationally prohibitive. Therefore, we use the number of *training epochs* as a practical, monotonic proxy for memorization level. Under standard diffusion training dynamics, longer training induces greater alignment between generated and training data distributions (potentially overfitting them). Accordingly, we interpret **larger epoch counts as stronger memorization**. This substitution is theoretically justified by the monotone relation between training convergence and SIDE divergence:

$$M(D; p_\theta^{(E_1)}, \varepsilon_0) > M(D; p_\theta^{(E_2)}, \varepsilon_0) \quad \text{for } E_1 < E_2,$$

which implies that, as training progresses, the memorization gap widens and the lower bound on $\Delta(\varepsilon)$ rises.

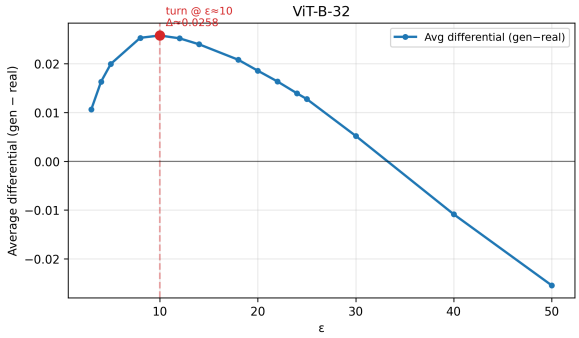
Experimental setup. We train a Denoising Diffusion Probabilistic Model (DDPM) with approximately 50 million parameters on the CelebA-HQ dataset. Training epochs range from 200 to 5200, providing a progressive spectrum of memorization levels. For each checkpoint, we generate a fixed number of synthetic images and compare their encoder-space shift statistics to those computed on the original training data. We use several fixed, pretrained encoders to define the semantic feature space: ViT-B/32, ViT-B/16, ViT-L/14, and DINOv2-ViT-B/14. For each encoder, we compute:

1. **Differential shift vs. ε:** The average difference $\Delta(\varepsilon)$ across a grid of Gaussian perturbation magnitudes ε. This curve reflects how the differential shift evolves with probing noise intensity.
2. **Differential shift vs. training epochs:** The mean differential shift aggregated over ε values, plotted as a function of the number of training epochs, to examine how memorization impacts Δ.

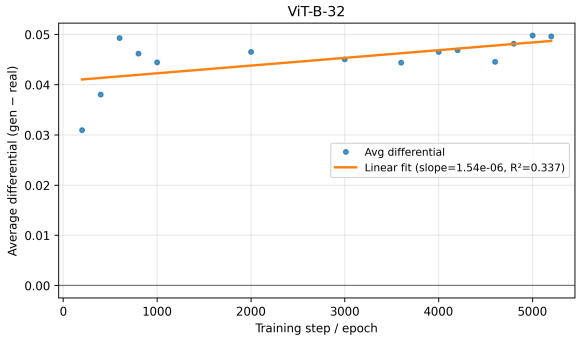
Findings and interpretation. The results are summarized in Figures 6a–h. Across all encoders, we consistently observe the two–phase structure predicted by theory. At small noise levels ($\varepsilon \in [3, 10]$), $\Delta(\varepsilon)$ increases sharply, corresponding to the regime where perturbations probe tangent directions of the data manifold. Beyond a turning point $\varepsilon_{\text{turn}}$, the differential shift gradually declines as perturbations enter the normal subspace regime and the encoder's local linear approximation breaks down. This non-monotonic behavior manifests as an "inverted U-shape" or S-shaped curve, in line with the analytic lower-bound structure derived in Theorem 4.3.

Furthermore, when analyzing Δ against training epochs (used as a proxy for memorization), we find that higher epoch counts correspond to larger differential shifts at small ε, indicating stronger memorization bias.

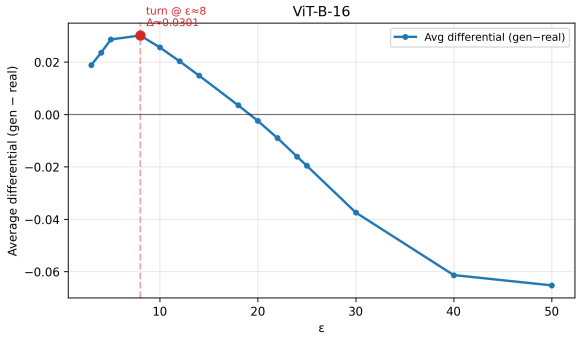
Takeaways. (i) Using epochs as a proxy for SIDE memorization, the differential shift increases with memorization in the small–noise regime, matching the theorem's lower bound. (ii) As ε grows, the differential exhibits a turning point and then decreases, also as predicted. (iii) The pattern is consistent across encoders, supporting the universality of the feature–shift operator.



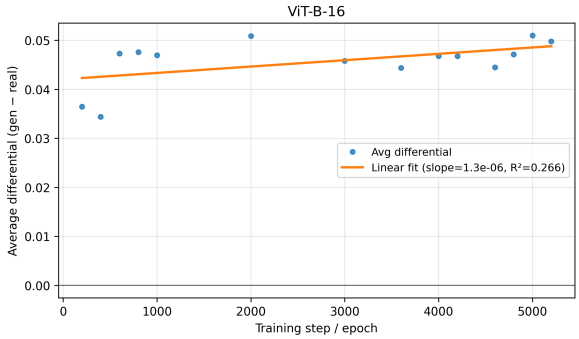
(a) ViT-B/32: differential vs. ε (two-phase)



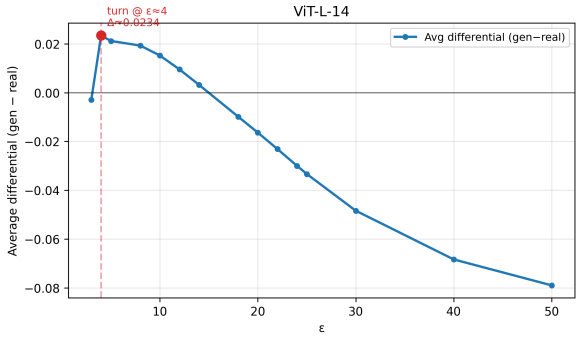
(b) ViT-B/32: differential vs. *epochs*



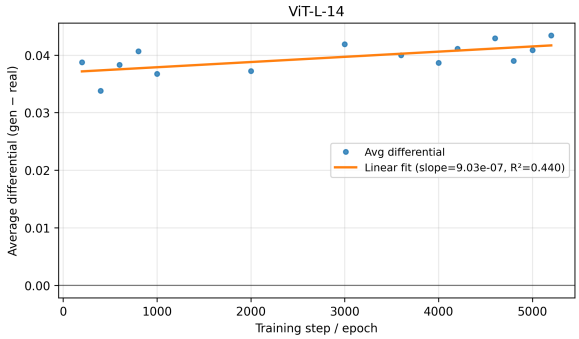
(c) ViT-B/16: differential vs. ε (two-phase)



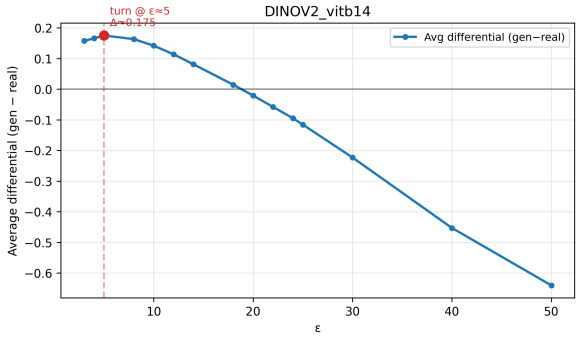
(d) ViT-B/16: differential vs. *epochs*



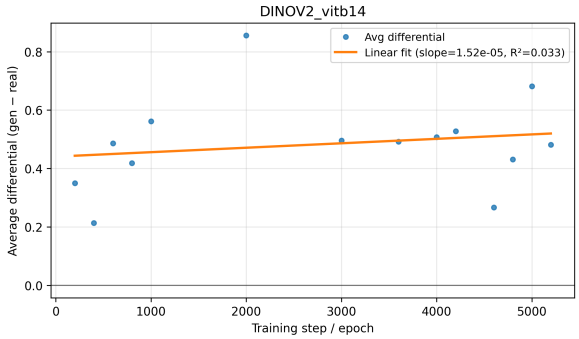
(e) ViT-L/14: differential vs. ε (two-phase)



(f) ViT-L/14: differential vs. *epochs*



(g) DINOv2-B/14: differential vs. ε (two-phase)



(h) DINOv2-B/14: differential vs. *epochs*

Figure 6. **Empirical validation of the lower bound.** Left column: average differential shift $\text{Shift}_{\text{gen}} - \text{Shift}_{\text{real}}$ versus probe magnitude ε, showing the predicted rise at small ε and decline at larger ε. Right column: average differential versus *training epochs* (used as a practical proxy for memorization since SIDE divergence is expensive to compute). Larger epoch counts correspond to stronger memorization and yield larger small–ε differentials, in line with Theorem 4.3.