THE GEOMETRY OF FAKENESS: AI-IMAGE DETECTION IN THE OPEN WORLD

Anonymous authors
Paper under double-blind review

ABSTRACT

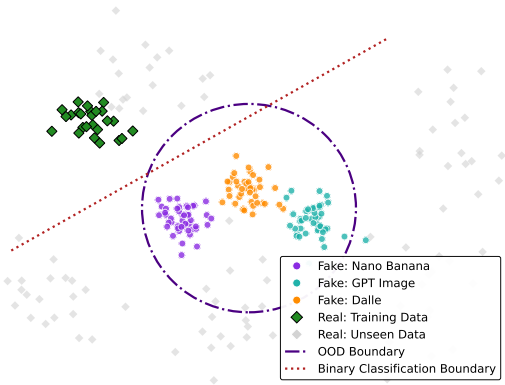
AI-generated image detection is commonly formulated as a binary classification problem between real and fake images. We argue that this formulation is fundamentally misaligned with the open-world nature of the task. While AI-generated images originate from a relatively limited family of generators and thus form a compact, learnable distribution, real images arise from diverse cameras, scenes, acquisition pipelines, and post-processing operations, making them too heterogeneous to be faithfully captured by finite training samples. This asymmetry creates a posterior mismatch between the training and deployment distributions, so that even a near-optimal binary classifier can suffer amplified open-world generalization error. We therefore reformulate AI-generated image detection as an out-of-distribution (OOD) detection problem, treating AI-generated images as in-distribution and real images as outliers. Yet OOD learning alone is insufficient, since the learned representation may still rely on unstable generator-specific shortcuts, such as low-level artifacts or compression traces, that fail to transfer across models and domains. Motivated by our theoretical analysis that information bottleneck (IB) can mitigate such bias amplification by filtering out spurious nuisance features, we propose an IB-regularized one-class objective that learns compact representations of AI-generated images while preserving stable generative cues. Experiments on both classical diffusion models and recent generation systems show consistent improvements over strong baselines, including a 12.90% absolute gain in average accuracy over the best prior detector. Our results suggest that AI-generated image detection is better understood as open-world distribution modeling rather than closed-world binary classification.

1 INTRODUCTION

Recent advances in deep generative models have made AI-generated images increasingly realistic and widely accessible, benefiting creative workflows while also enabling deepfakes, fabricated evidence, and large-scale visual misinformation (Karras et al., 2019; Rombach et al., 2022; OpenAI, 2025; Google, 2025; Westerlund, 2019). As a result, reliable AI-generated image detection has become an important problem.

Most existing methods formulate AI-generated image detection as a binary classification task between real and fake images (Wang et al., 2020; Chai et al., 2020; Zhang et al., 2019; Ojha et al., 2023). We argue that this formulation is fundamentally misaligned with the open-world nature of the task. AI-generated images arise from a relatively limited family of generators and thus admit a compact, learnable distribution (Marra et al., 2019; Matern et al., 2019). Real images, in contrast, are produced by highly heterogeneous pipelines involving diverse cameras, scenes, acquisition processes, and post-processing operations (Fridrich, 2009; Bayram et al., 2005). Finite training samples therefore provide only a narrow and biased view of real-image variability, causing binary detectors to overfit the observed support of real images and generalize poorly to unseen real-world data. As shown in Figure 1, this asymmetry is naturally out-of-distribution (OOD) in character: fake images cluster more compactly, while real images exhibit much larger dispersion. We empirically verify this distributional asymmetry on *diffusion* and *autoregressive models* using two frozen, off-the-shelf image encoders, DINOv2 and CLIP. As shown in Figure 1, AI-generated images exhibit substantially smaller intra-class distances than real images in both feature spaces. Accordingly, treating AI-generated images as ID and real images as OOD outperforms the reverse formulation (Section 4).

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.



Metric	DINOv2	CLIP
Intra-distance real images	1.3250	1.2650
Intra-distance fake images	0.5455	0.8732
Inter-distance	1.1592	1.1232

Figure 1: **Failure of binary classification under distributional asymmetry. Left:** A binary classifier successfully separates fake images from a limited training set of real images but fails to generalize to the substantially greater diversity of unseen real ones. Reformulating the task as OOD detection allows a hypersphere boundary to compactly model fake images as in-distribution (ID) samples while naturally rejecting diverse real images as OOD. **Right:** A quantitative comparison of feature-space distances on our curated dataset (Section 4) shows that intra-class distances are substantially greater among real images than among fake images.

This asymmetry is not only empirical but also theoretical. A binary classifier learns the posterior induced by the finite training distribution; under open-world deployment, however, the real-image distribution shifts beyond this sampled support, creating posterior mismatch. Our analysis shows that even a near-optimal binary classifier on the training distribution can incur strongly amplified generalization error under deployment shift. This explains why closed-world binary classification is intrinsically brittle for AI-generated image detection.

Motivated by this observation, we reformulate AI-generated image detection as an OOD detection problem, treating AI-generated images as in-distribution and real images as outliers. The key intuition is that an AI generator behaves like a synthetic camera that leaves relatively stable generative fingerprints, making fake images easier to model compactly than real images. Rather than fitting a brittle boundary against the vast support of real images, we learn a one-class representation of AI-generated images and reject real images as samples outside the learned manifold.

OOD learning alone, however, is still insufficient. Deep models can rely on spurious shortcuts, such as generator-specific color statistics, frequency artifacts, watermark remnants, or compression traces, that are predictive on the training set but unstable across generators and domains (Geirhos et al., 2020; Pezeshki et al., 2021; Izmailov et al., 2022; Matern et al., 2019; Yu et al., 2019). To address this issue, we incorporate an Information Bottleneck (IB) regularization into the OOD objective (Ahuja et al., 2021; Tishby et al., 2000). By constraining the mutual information between the input and the learned representation, IB suppresses non-transferable nuisance features while preserving stable generative cues. From the perspective of our theory, this reduces the impact of spurious variation that would otherwise amplify deployment-time error, yielding a cleaner and more robust OOD boundary.

We evaluate our method on images from both classical diffusion models and recent generation systems. Our method consistently outperforms strong baselines, achieving up to 12.90% absolute improvement in average accuracy over the best prior detector. These results suggest that AI-generated image detection is better understood as open-world distribution modeling rather than closed-world binary classification.

Our contributions are as follows:

- We revisit AI-generated image detection from an OOD perspective, treating AI-generated images as in-distribution and real images as outliers.
- We provide a theoretical motivation showing that binary classification suffers from posterior mismatch under open-world deployment, which amplifies training bias into large generalization error.
- We propose an IB-regularized one-class OOD objective that learns compact representations of AI-generated images while suppressing unstable generator-specific shortcuts.

- Extensive experiments across classical and recent generation systems demonstrate consistent gains over strong baselines and state-of-the-art detectors.

2 Problem Definition and Theoretical Motivation

This section develops the theoretical framework that motivates our methodological choices. Section 2.1 formalizes the standard binary classification setup. Section 2.2 shows how dataset quality can be amplified into an arbitrarily large open-world error, motivating OOD learning. Section 2.3 analyzes how the information bottleneck principle for OOD learning can drastically reduce this amplification factor, thereby improving generalization. Theoretical details and proofs are deferred to Appendix A.

2.1 Problem Definition

We study the task of detecting whether a given image is fake (AI-generated) or real. Traditionally, this problem is formulated as a binary classification task, where a detector is trained to distinguish real images from fake ones using supervised learning. Formally, let $\mathcal{X}$ denote the image space, and consider a dataset $\mathcal{D} = \{(x, y) \mid x \in \mathcal{X},\, y \in \{0, 1\}\}$ generated with class priors q_F, q_R and class-conditional distributions P_F, P_R:

$$P(y = 0) = q_F, \quad P(y = 1) = q_R = 1 - q_F, \tag{1}$$

$$P(x \mid y = 0) = P_F(x), \quad P(x \mid y = 1) = P_R(x). \tag{2}$$

Here $y = 0$ denotes fake images and $y = 1$ denotes real images. The goal is to train a classifier $f : \mathcal{X} \to \{0, 1\}$ that distinguishes the two distributions, typically trained with cross-entropy loss:

$$\mathcal{L}_{\mathrm{CE}}(f \mid \mathcal{D}) = -q_F\, \mathbb{E}_{x\sim P_F}\big[\log p_\theta(y = 0 \mid x)\big] - q_R\, \mathbb{E}_{x\sim P_R}\big[\log p_\theta(y = 1 \mid x)\big]. \tag{3}$$

A standard derivation (see Appendix A) shows that this loss admits the integral form

$$\mathcal{L}_{\mathrm{CE}}(f \mid \mathcal{D}) = \int_{\mathcal{X}} \mathcal{H}\big(p_{\mathcal{D}}(\cdot \mid x),\, p_\theta(\cdot \mid x)\big)\, P_{\mathcal{D}}(x)\, \mathrm{d}x, \tag{4}$$

where $P_{\mathcal{D}}(x) := q_F P_F(x) + q_R P_R(x)$ is the marginal data density and $p_{\mathcal{D}}(y \mid x)$ is the *posterior classifier* induced by $\mathcal{D}$:

$$p_{\mathcal{D}}(y = 0 \mid x) = \frac{q_F P_F(x)}{q_F P_F(x) + q_R P_R(x)}, \qquad p_{\mathcal{D}}(y = 1 \mid x) = \frac{q_R P_R(x)}{q_F P_F(x) + q_R P_R(x)}. \tag{5}$$

By the properties of cross-entropy, $\mathcal{L}_{\mathrm{CE}}$ is minimized when $p_\theta(\cdot \mid x) = p_{\mathcal{D}}(\cdot \mid x)$, with minimal value $\mathbb{E}_{x\sim\mathcal{D}}[\mathcal{H}(p_{\mathcal{D}}(\cdot \mid x))]$. In other words, fitting a binary classifier on $\mathcal{D}$ recovers the posterior $p_{\mathcal{D}}$ of the training distribution.

2.2 Why Binary Classification Fails: A Posterior-Mismatch View

A standard premise behind binary classification is that there exists a ground-truth probability $\hat{p}_F(x)$ that an image x is fake, and that the training distribution $\mathcal{D}$ is consistent with it in the Bayes sense:

$$\big(q_F P_F(x)\big)\big/\big(q_R P_R(x)\big) = \hat{p}_F(x)\big/\big(1 - \hat{p}_F(x)\big). \tag{6}$$

If this held exactly, the posterior $p_{\mathcal{D}}$ would equal $\hat{p}$, and the classifier learned on $\mathcal{D}$ would be Bayes-optimal everywhere. In practice, this consistency rarely holds. Any finite training set $\mathcal{D}$ can sample only a sliver of the cameras, scenes, acquisition pipelines, and post-processing operations that generate real-world images. Hence P_R under-represents large regions of image space, and the posterior $p_{\mathcal{D}}$ deviates from the ground truth $\hat{p}$, especially in the *sparse* regions of $\mathcal{D}$, where mislabeling carries little training cost but greatly distorts what the classifier predicts at deployment. We formalize this by characterizing how far $p_{\mathcal{D}}$ deviates from $\hat{p}$, weighted by the data density.

Definition 1 (Quality of a dataset) *Let $\mathcal{D} = (q_F, q_R, P_F, P_R)$ be a training distribution and let $p_{\mathcal{D}}$ denote its posterior classifier. The* quality *of $\mathcal{D}$ relative to ground truth $\hat{p}$ is*

$$K(\mathcal{D}) := \mathbb{E}_{x\sim P_{\mathcal{D}}}\big[D_{\mathrm{KL}}\big(\hat{p}(\cdot \mid x) \,\big\|\, p_{\mathcal{D}}(\cdot \mid x)\big)\big].$$

*We say $\mathcal{D}$ is δ-*suboptimal *if $K(\mathcal{D}) \le \delta$.*

Highest quality $K(\mathcal{D}) = 0$ is achieved iff $p_{\mathcal{D}} = \hat{p}$ almost everywhere under $P_{\mathcal{D}}$. Note that the expectation is taken under $P_{\mathcal{D}}(x)$, so mislabeling in *sparse* regions of $\mathcal{D}$ contributes little to K, matching reality where a finite training dataset is not reliable in regions it barely samples. The following theorem makes precise the failure mode of binary classification: a near-optimal classifier on a slightly defected dataset can incur arbitrarily large generalization loss on the ground-truth distribution.

Theorem 1 (Binary-classifier failure) *There exists a δ-suboptimal dataset $\mathcal{D}$ such that, for any binary classifier f_θ achieving optimal cross-entropy loss on $\mathcal{D}$, the generalization loss on the open-world distribution $\hat{\mathcal{D}}$ (which is consistent with $\hat{p}$) satisfies*

$$\mathcal{L}_{\mathrm{CE}}(f_\theta \mid \hat{\mathcal{D}}) - \mathbb{E}_{x\sim\hat{\mathcal{D}}}\big[H(\hat{p}(\cdot \mid x))\big] \;\geq\; \delta \cdot \big[1 + D_{\chi^2}\big(P_{\hat{\mathcal{D}}} \,\big\|\, P_{\mathcal{D}}\big)\big].$$

Here, the amplification factor is the χ^2-divergence between the marginal densities of $\hat{\mathcal{D}}$ and $\mathcal{D}$. In Appendix A we analyze the magnitude of this amplification term, showing that for realistic biased datasets it grows inverse-linearly with the under-sampled mass and can dominate the overall loss. Theorem 1 therefore implies that a small training quality gap δ can produce an arbitrarily large open-world loss whenever $\mathcal{D}$ and $\hat{\mathcal{D}}$ disagree substantially over $\mathcal{X}$.

This motivates our central methodology: rather than fitting a binary posterior over a distribution we cannot fully sample, we reformulate detection as OOD modeling on the well-sampled distribution of fake images, and treat real images as outliers.

2.3 The Role of Information Bottleneck in OOD Detection

Reformulating detection as OOD does not by itself eliminate the open-world challenge. The amplification factor $D_{\chi^2}(P_{\hat{\mathcal{D}}} \,\|\, P_{\mathcal{D}})$ in Theorem 1 measures bias in the *input* density; the same factor applies to any representation we feed the OOD detector unless that representation is chosen to suppress it. The information bottleneck (IB) principle (Tishby et al., 2000) offers exactly such a representation: an encoder that compresses the input as much as possible while preserving information about the prediction target,

$$\max_{\Phi} \; I(z; y) - \beta\, I(z; x), \qquad z = \Phi(x). \tag{7}$$

The tradeoff is essential. With $\beta \to 0$, the encoder is unconstrained and retains all input bias; with $\beta \to \infty$, the representation collapses to a constant and predictivity is lost. The interesting regime is the interior, where the encoder is forced to discard input information that does not serve prediction.

Our claim in this section is that, under a suitable structural framework, this IB tradeoff substantially reduces the χ^2-divergence factor in Theorem 1. We work in the following idealized setting.

Setting 1 (Generative seed) *Every image is generated from a pair of latent factors,*

$$x = \Psi(h_{\mathrm{inv}}, h_{\mathrm{spu}}),$$

where Ψ is a measurable composer, $h_{\mathrm{inv}} \in \mathcal{H}_{\mathrm{inv}}$ encodes invariant *content that determines whether x is real or fake, and $h_{\mathrm{spu}} \in \mathcal{H}_{\mathrm{spu}}$ encodes* spurious *nuisance content (camera, lighting, post-processing). Furthermore, given h_{inv}, the spurious latent carries no information about the label:*

$$P(y \mid h_{\mathrm{inv}}, h_{\mathrm{spu}}) = P(y \mid h_{\mathrm{inv}}).$$

Remark (Idealization). Setting 1 describes a generative seed in which invariant and spurious contents are cleanly separable, and only invariant content is causally related to the label. Real-image generation in practice does not factor as cleanly, however, since generator artifacts may inhabit *both* invariant and spurious subspaces, causing spurious factors to be predictive. We present Setting 1 as a clean theoretical scaffold from which to reason about IB's behavior; the empirical results in Section 4 test whether the predictions of this scaffold survive in real data.

Revisiting equation 7, we can derive that the IB-optimal encoder $z^* = \Phi^*(x)$ under this setting is a measurable function of h_{inv} alone (formal proof in Appendix A). This reduces the χ^2-divergence amplification factor of Theorem 1 to the divergence on the invariant latent. We characterize this lowered divergence in the following assumption.

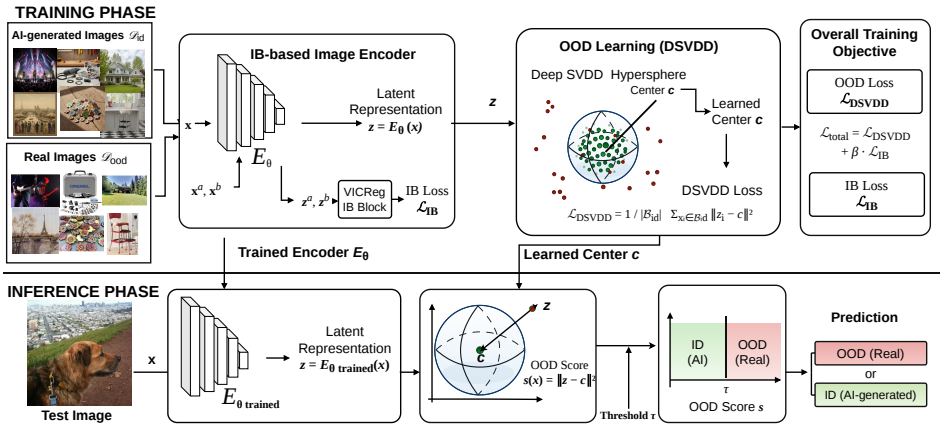


Figure 2: Overview of the proposed method: an IB-regularized image encoder learns representations of AI-generated images using VICReg, and an OOD objective models them as a compact hypersphere. At inference time, the distance to the hypersphere center is used for detection: images within the threshold are classified as AI-generated, while those outside are classified as real.

Assumption 1 (Bias–spurious alignment) *For any distribution P over images $x = \Psi(h_{\text{inv}}, h_{\text{spu}})$, define the* invariant marginal *of P as the pushforward of P under the projection $x \mapsto h_{\text{inv}}$, denoted P^{inv}. Then the invariant marginals of the training and open-world real distributions are close in χ^2: $D_{\chi^2}\big(P_R^{\text{inv}} \,\big\|\, \hat{P}_R^{\text{inv}}\big) \;\le\; \varepsilon_{\text{inv}}$, for some small $\varepsilon_{\text{inv}} \ge 0$.*

Plainly speaking, P^{inv} is the distribution of h_{inv} when $x \sim P$. Assumption 1 formalizes the intuition that real images across datasets are produced by the same *physical* process, namely light interacting with scenes and being captured by sensors, while what differs across real-image datasets is predominantly the *nuisance* component. We further discuss the theoretical and practical implications of this assumption in Appendix A.

The central theorem of this section serves as an inverse to Theorem 1: under IB regularization, a classifier that achieves optimal training loss also achieves bounded loss on the open-world distribution.

Theorem 2 (IB-controlled generalization) *Let $\Phi^*(x)$ denote the IB-optimal representation, and let $f_\theta(x) = g_\theta(\Phi^*(x))$ be a latent classifier that achieves optimal cross-entropy loss on a δ-suboptimal training distribution $\mathcal{D}$. Then the generalization loss on the open-world distribution $\hat{\mathcal{D}}$ satisfies*

$$\mathcal{L}_{\text{CE}}(f_\theta \mid \hat{\mathcal{D}}) - \mathbb{E}_{x\sim\hat{\mathcal{D}}}\big[H(\hat{p}(\cdot \mid x))\big] \;\le\; \delta \cdot (1 + \varepsilon_{\text{inv}}).$$

Implication. Theorem 2 stands as the inverse of Theorem 1, showing that an IB-regularized classifier's training loss is amplified by at most ε_{inv} at deployment. This amplification factor is driven by a small, content-level bias. This makes precise the sense in which IB provides robustness *to bias*, not robustness *to bad data*: a dataset whose bias lives in h_{spu} is well-served by IB compression, while a dataset whose content h_{inv} is itself mis-distributed degrades generalization loss correspondingly.

3 Methodology

3.1 Overview

Let $x \in \mathcal{X}$ denote an image, and let $\mathcal{D}_{\text{id}}$ be an *in-distribution* consisting entirely of AI-generated images. Let $\mathcal{D}_{\text{ood}}$ denote the open-world distribution of real images.

The primary goal of our framework is to determine whether a test image x is AI-generated (ID) or real (OOD). To do so, we first use an image encoder E_θ to map the input image into a latent representation $z \in \mathcal{Z} \subseteq \mathbb{R}^d$, where d denotes the embedding dimension:

$$z = E_\theta(x), \quad \text{where } E_\theta : \mathcal{X} \to \mathcal{Z}. \tag{8}$$

We then define a scoring function $s : \mathcal{Z} \to \mathbb{R}$ such that larger scores indicate a higher likelihood of being OOD:

$$s(z) > \tau \Rightarrow x \text{ is predicted as OOD}, \qquad s(z) \le \tau \Rightarrow x \text{ is predicted as ID}, \tag{9}$$

where τ is a decision threshold. In the following subsections, we detail the encoder design and the derivation of OOD scores.

3.2 Image Encoder with Information Bottleneck

We regularize the image encoder E_θ with an IB to suppress spurious features while preserving causal generative cues in AI-generated images. A detailed theoretical analysis of how IB improves OOD learning is provided in Appendix A.5. Specifically, we implement the IB via VICReg, a strong self-supervised representation learning method that provides a practical and stable objective for learning invariant and compact representations (Bardes et al., 2021).

VICReg-based Information Bottleneck. VICReg learns representations that are invariant to semantic-preserving augmentations while preventing representational collapse through variance and covariance regularization. This aligns with our goal of suppressing spurious generator-specific cues and retaining robust, transferable synthesis artifacts.

For each input image, we generate two augmented views, x^a and x^b, and compute their embeddings $z^a = E_\theta(x^a)$ and $z^b = E_\theta(x^b)$, where $z^a, z^b \in \mathbb{R}^d$. The VICReg objective has three terms: an invariance term that pulls the two view embeddings together, a variance term that prevents collapse by enforcing sufficient per-dimension standard deviation, and a covariance term that reduces redundancy by decorrelating feature dimensions. Formally:

$$\mathcal{L}_{\text{IB}} = \lambda \mathcal{L}_{\text{inv}}(z^a, z^b) + \mu \mathcal{L}_{\text{var}}(z^a, z^b) + \nu \mathcal{L}_{\text{cov}}(z^a, z^b), \tag{10}$$

where λ, μ, and ν are hyperparameters. Over a mini-batch $\mathcal{B}$, the invariance term is the mean squared error between paired embeddings, $\mathcal{L}_{\text{inv}} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \left\| z_i^a - z_i^b \right\|_2^2$.

The variance term enforces each feature dimension to remain above a predefined margin γ across the mini-batch, $\mathcal{L}_{\text{var}} = \frac{1}{d} \sum_{j=1}^{d} \max\big(0, \gamma - \text{Std}_{i \in \mathcal{B}}(z_{i,j}^a)\big) + \frac{1}{d} \sum_{j=1}^{d} \max\big(0, \gamma - \text{Std}_{i \in \mathcal{B}}(z_{i,j}^b)\big)$.

The covariance term penalizes off-diagonal entries of covariance matrices computed from centered embeddings in the mini-batch, $\mathcal{L}_{\text{cov}} = \frac{1}{d(d-1)} \sum_{p \neq q} \left[\text{Cov}_{i \in \mathcal{B}}(z^a)\right]_{pq}^2 + \frac{1}{d(d-1)} \sum_{p \neq q} \left[\text{Cov}_{i \in \mathcal{B}}(z^b)\right]_{pq}^2$. Here, j, p, and q index feature dimensions. In practice, this objective promotes augmentation-invariant, non-collapsed, and decorrelated representations.

Inference. At inference time, we directly use the $E_\theta(x)$ to produce the deterministic embeddings z.

3.3 OOD Learning

Our framework is *loss-agnostic* with respect to the OOD objectives. In this subsection, we consider two OOD objectives and scoring functions, i.e., DSVDD and energy-based methods.

Deep SVDD (DSVDD). DSVDD learns minimum-volume hyperspheres in latent space that enclose ID samples. To capture the heterogeneous artifacts produced by different image generators, we initialize K centers $\{\mathbf{c}_k\}_{k=1}^K \subset \mathbb{R}^d$, obtained by running k-means on the normalized ID embeddings and projecting each center onto the unit sphere:

$$\begin{aligned} \{\mathbf{c}_k\}_{k=1}^K &= k\text{-means}\Big(\Big\{ \tfrac{z}{\|z\|+\epsilon} : x \in \mathcal{D}_{\text{id}} \Big\}\Big), \quad \text{where } z = E_\theta(x), \\ \mathbf{c}_k &\leftarrow \mathbf{c}_k / \|\mathbf{c}_k\|_2, \qquad k = 1, \ldots, K. \end{aligned} \tag{11}$$

The one-class DSVDD objective assigns each ID sample to its nearest center and minimizes the mean squared distance to that center:

$$\mathcal{L}_{\text{DSVDD}} = \frac{1}{|\mathcal{B}_{\text{id}}|} \sum_{x_i \in \mathcal{B}_{\text{id}}} \min_{k \in \{1, \ldots, K\}} \|z_i - \mathbf{c}_k\|^2, \tag{12}$$

where $\mathcal{B}_{\text{id}}$ is the set of AI-generated samples in the current mini-batch. Minimizing this objective maps AI-generated images into compact clusters, while real images tend to lie farther away.

Energy-based. Energy-based OOD detection learns a scoring function that assigns low energy to in-distribution samples and high energy to OOD samples. In our setting, the embedding $z = E_\theta(x)$ is passed through a lightweight classifier h to produce logits and the energy score is defined via the log-sum-exp operator over logits: $\phi(z) = -\log \sum_{k=1}^K \exp(h(z)_k)$, where $h(z) \in \mathbb{R}^K$ denotes

a K-dimensional logit vector. This formulation assigns lower energy to samples considered more likely under the learned representation.

To separate ID and OOD samples, we adopt a margin-based objective. Let $\mathcal{B}_{\text{id}}$ and $\mathcal{B}_{\text{ood}}$ denote the sets of AI-generated and real images in a mini-batch. The energy loss is defined as:

$$\mathcal{L}_{\text{energy}} = \frac{1}{|\mathcal{B}_{\text{id}}|} \sum_{x_i \in \mathcal{B}_{\text{id}}} \left[\max(0, \phi(z_i) - m_{\text{in}})\right]^2 + \frac{1}{|\mathcal{B}_{\text{ood}}|} \sum_{x_j \in \mathcal{B}_{\text{ood}}} \left[\max(0, m_{\text{out}} - \phi(z_j))\right]^2, \tag{13}$$

where m_{in} and m_{out} are margin hyperparameters. The first term penalizes ID samples whose energy exceeds the margin, encouraging compact low-energy regions for AI-generated images. The second term pushes real images into higher-energy regions.

By optimizing this objective, the encoder learns representations in which AI-generated images lie on a low-energy manifold, while real images naturally obtain higher energy scores and can therefore be detected as OOD.

Overall Training Objective. The final objective combines the OOD and IB losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{OOD}} + \beta \cdot \mathcal{L}_{\text{IB}}, \quad \mathcal{L}_{\text{OOD}} \in \{\mathcal{L}_{\text{DSVDD}}, \mathcal{L}_{\text{energy}}\}, \tag{14}$$

where β is the weighting coefficient that balances the two terms.

Inference. At inference time, for DSVDD, the OOD score is defined as the squared distance to the nearest hypersphere center $s_{\text{DSVDD}}(z) = \min_{k \in \{1,\ldots,K\}} \|z - \mathbf{c}_k\|_2^2$. For the energy-based method, the OOD score is directly defined as $s_{\text{energy}}(z) = \phi(z)$. In both cases, a larger score indicates that the sample is more likely to be OOD.

4 Results

4.1 Experimental Setups

Datasets. We utilize the image dataset from previous baseline studies (Ojha et al., 2023; Wang et al., 2020). The datasets comprise fake images generated by various diffusion models, including Latent Diffusion Model (LDM) (Rombach et al., 2022), Glide (Nichol et al., 2021), and guided diffusion (Dhariwal & Nichol, 2021), as well as the autoregressive model DALL-E (Ramesh et al., 2021). Specifically, variants such as LDM-200, LDM-200 w/CFG, LDM-100, Glide 100-27, Glide 50-27, and Glide 100-10 represent different sampling steps and guidance configurations for generating images. Furthermore, we expand our evaluation to include the latest autoregressive generators, namely Nano Banana Pro (Google, 2025) and GPT Image 1.5 (OpenAI, 2025). To generate fake images from these models, we follow the same protocol of existing work (Ojha et al., 2023) to sample image-text pairs from the LAION dataset (Schuhmann et al., 2021) and use the corresponding text descriptions as prompts. For both models, we adopt the same training, validation, and test splits as the existing datasets. For more detailed information regarding dataset construction, please refer to Appendix B.1.

Baselines. We compare our approach against seven established baselines: (1) **Trained deep network** (Wang et al., 2020), a ResNet-50 trained end-to-end using binary cross-entropy; (2) **Patch classifier** (Chai et al., 2020), the classification network is trained at the patch level; (3) **Freq-spec** (Zhang et al., 2019), the classification network is trained on frequency spectrum; (4) **Pre-trained ViT** (Ojha et al., 2023), which uses pre-trained ViT to feature representation and then applies linear classification; (5) **Fine-tune ViT** (Khan & Dang-Nguyen, 2024), which fine-tunes the pre-trained ViT on fake image dataset for classification; (6) **Corrupt-and-recovery** (Bai et al., 2025), which detects fakes based on the premise that generative models can reconstruct masked portions of fake images more accurately than real ones; and (7) **Real as ID** (Cai et al., 2026), which adopts the reverse OOD formulation by treating real images as ID and AI-generated images as OOD. Please see Appendix B.3 for more details.

Metrics. Following all the baselines, we evaluate detection performance using classification accuracy and average precision (AP). AP summarizes the precision-recall curve into a single value representing the area under the curve, providing a robust measure of a classifier's performance across all possible thresholds. The classification threshold τ is selected by maximizing the accuracy on a held-out validation set. Specifically, we split the training portion into train/validation with a $4{:}1$ ratio and select the OOD threshold using the best validation performance. Please refer to Appendix C.4 for our results on additional evaluation metrics.

Table 1: Accuracy (%) of fake image detection methods across different generative models.

Detection method	Variant	Guided	LDM			Glide			DALL-E	Nano Banana	GPT Image	Avg. acc
			200	200 w/ CFG	100	100-27	50-27	100-10				
Trained deep network	Blur+JPEG (0.1)	68.36	70.46	68.15	70.97	73.03	68.87	60.62	71.64	64.10	**77.38**	69.36
	Blur+JPEG (0.5)	71.44	63.38	71.79	69.49	71.33	67.38	58.21	72.72	64.31	73.90	68.39
Patch classifier	ResNet50-Layer1	50.82	50.41	51.59	51.38	54.00	50.41	58.92	50.41	50.05	52.10	52.01
	Xception-Block2	55.90	50.00	55.59	58.97	61.49	57.64	59.59	59.13	56.10	67.03	58.14
Freq-spec	CycleGAN	53.13	50.05	51.79	73.03	76.36	70.36	57.33	57.33	57.64	71.13	61.82
Pre-trained ViT	ViT:CLIP	68.05	67.79	67.13	72.21	74.72	69.49	66.77	75.79	**71.95**	71.90	70.58
	ViT:DINOv2	72.78	50.10	73.43	75.10	78.38	76.21	54.95	56.92	50.00	50.00	63.79
Fine-tuned ViT	ViT:CLIP	50.97	50.82	50.00	50.00	50.72	61.79	53.90	50.00	50.00	54.51	52.27
Corrupt-and-recovery	Stable Diffusion	50.98	50.31	50.21	50.41	50.82	50.57	64.28	50.21	57.01	50.72	52.55
Real as ID	DSVDD	52.56	55.18	49.95	53.54	60.10	64.10	55.28	49.95	49.95	49.95	54.06
Real as OOD (ours)	DSVDD	**92.05**	**83.08**	**92.87**	**90.67**	**90.26**	**90.21**	**70.46**	**81.90**	71.85	71.44	**83.48**
	Energy	86.51	75.74	86.10	83.38	83.18	82.97	64.31	72.97	65.64	61.08	76.19

Table 2: Average Precision (AP%) of fake image detection methods across different generative models.

Detection method	Variant	Guided	LDM			Glide			DALL-E	Nano Banana	GPT Image	mAP
			200	200 w/ CFG	100	100-27	50-27	100-10				
Trained deep network	Blur+JPEG (0.1)	68.68	77.79	68.41	76.90	76.79	78.93	76.03	69.53	69.67	76.13	73.89
	Blur+JPEG (0.5)	63.00	76.03	76.36	75.03	76.63	80.06	74.37	76.37	72.19	73.43	74.35
Patch classifier	ResNet50-Layer1	66.25	58.19	63.46	57.35	66.90	71.24	65.81	61.15	49.15	62.55	62.21
	Xception-Block2	64.79	61.93	62.27	60.72	66.61	70.91	65.44	65.05	64.32	76.14	65.82
Freq-spec	CycleGAN	67.81	67.41	70.56	68.29	73.81	77.05	72.22	78.23	75.93	76.14	72.75
Pre-trained ViT	ViT:CLIP	62.24	59.89	61.34	59.49	76.82	81.07	76.27	65.21	60.31	72.33	67.50
	ViT:DINOv2	82.71	71.73	82.73	86.72	87.72	86.87	67.42	71.33	59.03	59.28	75.55
Fine-tuned ViT	ViT:CLIP	53.48	51.56	52.45	50.28	65.99	73.41	66.58	61.77	50.76	60.88	58.72
Corrupt-and-recovery	Stable Diffusion	68.96	47.30	43.28	47.82	46.19	48.97	43.83	36.36	33.67	44.92	47.88
Real as ID	DSVDD	58.21	63.25	60.37	61.14	67.50	69.68	66.92	58.12	56.35	55.40	61.69
Real as OOD (ours)	DSVDD	**97.29**	**91.68**	**97.62**	**96.28**	**96.41**	**96.31**	**80.31**	**90.25**	**79.36**	**79.26**	**90.48**
	Energy	93.79	85.67	94.20	91.61	91.69	92.11	77.44	82.02	71.45	73.91	85.39

4.2 Main Results

Table 1 and Table 2 show the average classification accuracy and mean AP of all detection methods. It can be easily observed that across both evaluation metrics, the two variants of our proposed method (i.e., DSVDD and Energy) substantially outperform all baseline methods. Specifically, our primary variant, OOD (DSVDD), achieves a state-of-the-art total mean AP of 90.48% and an average classification accuracy of 83.48%. This represents a substantial absolute improvement of 14.93% in mAP compared to the strongest baseline, the Pre-trained ViT (ViT:DINOv2), which achieves a total mean AP of 75.55%. Similarly, our method surpasses the best-performing baseline in accuracy, Pre-trained ViT (ViT:CLIP) at 70.58%, by an absolute improvement of 12.90%.

Interestingly, although our proposed OOD (DSVDD) method maintains a clear lead over the baselines, it experiences a noticeable drop when facing the latest generative models (i.e., Nano Banana and GPT Image). A possible explanation is that most images from these two advanced models have a different distribution from the majority of our training set (from earlier diffusion models), making it challenging for the OOD objective to effectively bind them together. Please refer to Appendix C.2 for additional ablation studies on our method using different image encoders and IB objectives.

4.3 Robustness

To evaluate robustness to spurious correlations, we inject six types of synthetic shortcuts into fake images during training while keeping the test data unchanged (see Appendix B.2 for details of shortcut construction). These shortcuts may cause the model to rely on superficial cues, such as specific compression traces, rather than stable generative cues that distinguish AI-generated images from real ones.

Figure 3 shows the absolute accuracy and the accuracy drop relative to the original no-shortcut DSVDD setting (83.48%). Across all six shortcut types, DSVDD+IB consistently outperforms DSVDD, achieving 81.0–82.7% accuracy compared with 80.1–81.4% for DSVDD. It also reduces the accuracy drop from 2.0–3.4% to 0.8–2.5%. These results indicate that IB suppresses spurious shortcut features and preserves more stable AI-vs.-real cues under distribution shift. Appendix C.1 further shows that our method generalizes well to unseen generators.

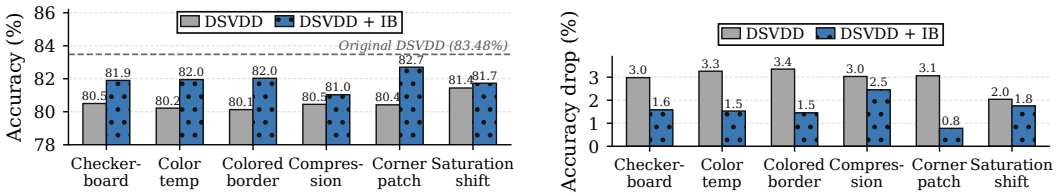


Figure 3: Performance under six shortcut-based spurious feature constructions. Left: absolute accuracy. Right: accuracy drop relative to the original no-shortcut DSVDD setting (83.48%). DSVDD+IB consistently achieves higher accuracy and a smaller drop than DSVDD across all shortcut types.

5 Related Work

AI-Generated Image Detection. Individual generative models exhibit distinctive forensic features, including spatial inconsistencies, noise residuals, and compression artifacts (Matern et al., 2019; Yu et al., 2019). Building on these cues, previous work formulates the task as supervised binary classification, leveraging CNNs, ViTs, and diffusion models to learn synthetic fingerprints (Wang et al., 2020; Chai et al., 2020; Zhang et al., 2019; Ojha et al., 2023; Khan & Dang-Nguyen, 2024; Bai et al., 2025; Zhang et al., 2025; Tan et al., 2024). Although these methods are highly effective in-distribution, their generalization often degrades under domain shifts and real-world OOD conditions.

OOD Detection and One-Class Learning. Traditional anomaly detection relies on one-class support methods like the One-Class Support Vector Machine (Schölkopf et al., 2001). Deep learning-based one-class methods include hypersphere-based DeepSVDD (Ruff et al., 2018) and autoencoder-adversarial frameworks such as HRN (Hu et al., 2020), which better capture complex data distributions. Beyond boundary-based techniques, Energy-based OOD scoring has also been widely used to measure distributional mismatch (Liu et al., 2020). Because of their strong open-world generalization, these methods have been successfully adopted in digital forensics and other domains, such as AI-generated text detection (Zeng et al., 2025), medical image diagnostics (Hong et al., 2024), and network intrusion detection (Zhang & Zulkernine, 2006).

Information Bottleneck for OOD learning. The IB principle compresses inputs to strip away spurious correlations while retaining stable, causal features (Alemi et al., 2016; Bardes et al., 2021; Zbontar et al., 2021). It has been successfully extended to multiple domains to increase generalization, including graph learning (Yang et al., 2023), reinforcement learning (Igl et al., 2019), and natural language processing (Mahabadi et al., 2021). Notably, the prior work demonstrated that invariance-based IB provides provable OOD generalization guarantees (Ahuja et al., 2021). Building on these theoretical insights, our work adopts the IB principle for OOD detection of AI-generated images.

6 Conclusion

In this work, we revisit AI-generated image detection from an out-of-distribution perspective. Instead of learning a brittle binary decision boundary between real and fake images, we treat AI-generated images as the in-distribution and model them using one-class OOD detection. We further introduce an information-bottleneck regularization to encourage the model to preserve stable generative cues while suppressing unstable generator-specific shortcuts. Extensive experiments on images generated by the latest generative models demonstrate that our approach consistently outperforms existing detectors.

AI Use Statement

In this work, we used generative AI tools to formulate mathematical claims, assist in writing proofs, design and provide feedback on research methodology and experiments, implement methods, and

interpret results. We did not use generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, assist with translation, clean or reformat datasets, or support qualitative or thematic data analysis. No other required-disclosure tasks were applicable to this work.

Additionally, we used generative AI tools to create or edit software code, draft parts of the research paper, source and search for information, and improve the paper's readability. We reviewed all AI-assisted work for accuracy, correctness, relevance, and consistency with the research and revised it where necessary. We take responsibility for the final content of this work, including all text, claims, code, and other artifacts produced with the aid of generative AI.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Haoyue Bai, Yiyou Sun, Wei Cheng, and Haifeng Chen. Where's the liability in the generative era? recovery-based black-box detection of ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 28821–28830, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Sevinc Bayram, Husrev Sencar, Nasir Memon, and Ismail Avcibas. Source camera identification based on cfa interpolation. In *IEEE international conference on image processing 2005*, volume 3, pp. III–69. IEEE, 2005.

Qianshu Cai, Chao Wu, Yonggang Zhang, Jun Yu, and Xinmei Tian. Towards generalizable detector for generated image. *Advances in Neural Information Processing Systems*, 38:94773–94800, 2026.

Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *European conference on computer vision*, pp. 103–120. Springer, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jessica Fridrich. Digital image forensics. *IEEE Signal Processing Magazine*, 26(2):26–37, 2009.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International conference on learning representations*, 2018.

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

Google. Introducing nano banana pro, 2025. URL `https://blog.google/innovation-and-ai/products/nano-banana-pro/`.

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

Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

Wenpeng Hu, Mengyu Wang, Qi Qin, Jinwen Ma, and Bing Liu. Hrn: A holistic approach to one class learning. *Advances in neural information processing systems*, 33:19111–19124, 2020.

Maximilian Igl, Kamil Ciosek, Yingzhen Li, Sebastian Tschiatschek, Cheng Zhang, Sam Devlin, and Katja Hofmann. Generalization in reinforcement learning with selective noise injection and information bottleneck. *Advances in neural information processing systems*, 32, 2019.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. *Advances in Neural Information Processing Systems*, 35: 38516–38532, 2022.

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.

Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pp. 1006–1015, 2024.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. Variational information bottleneck for effective low-resource fine-tuning. *arXiv preprint arXiv:2106.05469*, 2021.

Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, pp. 506–511. IEEE, 2019.

Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pp. 83–92. IEEE, 2019.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24480–24489, 2023.

OpenAI. Gpt image 1.5, 2025. URL `https://openai.com/index/new-chatgpt-images-is-here/`.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems*, 34:1256–1272, 2021.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pp. 8821–8831. Pmlr, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

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

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

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

Mika Westerlund. The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 2019.

Ling Yang, Jiayi Zheng, Heyuan Wang, Zhongyi Liu, Zhilin Huang, Shenda Hong, Wentao Zhang, and Bin Cui. Individual and structural graph information bottlenecks for out-of-distribution generalization. *IEEE Transactions on Knowledge and Data Engineering*, 36(2):682–693, 2023.

Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7556–7566, 2019.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

Cong Zeng, Shengkun Tang, Yuanzhou Chen, Zhiqiang Shen, Wenchao Yu, Xujiang Zhao, Haifeng Chen, Wei Cheng, and Zhiqiang Xu. Human texts are outliers: Detecting llm-generated texts via out-of-distribution detection. *arXiv preprint arXiv:2510.08602*, 2025.

Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 23828–23837, 2025.

Jiong Zhang and Mohammad Zulkernine. Anomaly based network intrusion detection with unsupervised outlier detection. In *2006 IEEE International Conference on Communications*, volume 5, pp. 2388–2393. IEEE, 2006.

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. IEEE, 2019.

CONTENTS

1 Introduction **1**

2 Problem Definition and Theoretical Motivation **3**

2.1 Problem Definition 3

2.2 Why Binary Classification Fails: A Posterior-Mismatch View 3

2.3 The Role of Information Bottleneck in OOD Detection 4

3 Methodology **5**

3.1 Overview 5

3.2 Image Encoder with Information Bottleneck 6

3.3 OOD Learning 6

4 Results **7**

4.1 Experimental Setups 7

4.2 Main Results 8

4.3 Robustness 8

5 Related Work **9**

6 Conclusion **9**

A Theoretical Details and Proofs **16**

A.1 Preliminaries and Auxiliary Derivations 16

A.2 Proof of Theorem 1 16

A.3 Proof of Theorem 2 17

A.4 Further Discussion 18

A.5 Why IB Helps OOD via Cleaner Hypersphere Boundaries 20

B Detail of Experimental Setup **21**

B.1 Datasets 21

B.2 Shortcut Construction 22

B.3 Baselines 23

B.4 Variants of Information Bottleneck 24

B.5 Implementation Details 24

C Additional Experiments **24**

C.1 Robustness to Unseen AI Generators 24

C.2 Impact of Image Encoders and IB Objectives. 25

C.3 Impact of β. 25

C.4 More Metrics 26

C.5 Case Studies 26

D Representation-Level Analysis **26**

E Limitation and Ethical Statement **28**

F Compute Resources Used **28**

G Use of LLMs **28**

A Theoretical Details and Proofs

This appendix collects the technical material deferred from Section 2. Section A.1 establishes notation and provides the derivations omitted from the main text. Sections A.2 and A.3 prove Theorems 1 and 2 respectively. Section A.4 elaborates on the practical and theoretical content of our framework, including the magnitude of the χ^2-divergence under realistic biased sampling, the empirical content of Assumption 1, and the diagnostic value of our framework.

A.1 Preliminaries and Auxiliary Derivations

We use $\int_{x\in\mathcal{X}} \cdot\, \mathrm{d}x$ to equivalently denote integration or summation over the data space $\mathcal{X}$. For the ground-truth classifier $\hat{p}_F(x)$, we introduce the auxiliary notation

$$\hat{p}_R(x) := 1 - \hat{p}_F(x), \qquad \hat{p}(y=0 \mid x) := \hat{p}_F(x), \qquad \hat{p}(y=1 \mid x) := \hat{p}_R(x).$$

Cross-entropy in integral form. We derive the integral form equation 4 stated in Section 2.1. Starting from equation 3 and applying the definition of expectation,

$$\begin{aligned}
\mathcal{L}_{\mathrm{CE}}(f_\theta \mid \mathcal{D}) &= -q_F\, \mathbb{E}_{x\sim P_F}\big[\log p_\theta(y=0 \mid x)\big] - q_R\, \mathbb{E}_{x\sim P_R}\big[\log p_\theta(y=1 \mid x)\big] \\
&= -\int_{\mathcal{X}} \big[q_F P_F(x) \log p_\theta(y=0 \mid x) + q_R P_R(x) \log p_\theta(y=1 \mid x)\big]\, \mathrm{d}x \\
&= -\int_{\mathcal{X}} \big[p_{\mathcal{D}}(y=0 \mid x) \log p_\theta(y=0 \mid x) + p_{\mathcal{D}}(y=1 \mid x) \log p_\theta(y=1 \mid x)\big]\, P_{\mathcal{D}}(x)\, \mathrm{d}x \\
&= \int_{\mathcal{X}} \mathcal{H}\big(p_{\mathcal{D}}(\cdot \mid x),\, p_\theta(\cdot \mid x)\big)\, P_{\mathcal{D}}(x)\, \mathrm{d}x,
\end{aligned}$$

where the third equality uses the definition $p_{\mathcal{D}}(y=0 \mid x) = q_F P_F(x)/P_{\mathcal{D}}(x)$ and the analogous expression for $y=1$, and $\mathcal{H}(p,q) := -\sum_y p(y)\log q(y)$ denotes cross-entropy. By Gibbs' inequality, this loss is minimized pointwise when $p_\theta(\cdot \mid x) = p_{\mathcal{D}}(\cdot \mid x)$, attaining the minimum value

$$\min_\theta \mathcal{L}_{\mathrm{CE}}(f_\theta \mid \mathcal{D}) = \mathbb{E}_{x\sim P_{\mathcal{D}}}\big[\mathcal{H}(p_{\mathcal{D}}(\cdot \mid x))\big].$$

Pearson χ^2-divergence. For two distributions P_1, P_2 over $\mathcal{X}$, the Pearson χ^2-divergence is

$$D_{\chi^2}(P_1 \,\|\, P_2) = \int_{\mathcal{X}} \frac{(P_1(x) - P_2(x))^2}{P_2(x)}\, \mathrm{d}x = \int_{\mathcal{X}} \frac{P_1(x)^2}{P_2(x)}\, \mathrm{d}x - 1. \tag{15}$$

Equivalently, $D_{\chi^2}(P_1 \| P_2) + 1 = \mathbb{E}_{P_2}[(\mathrm{d}P_1/\mathrm{d}P_2)^2]$.

Pushforward and data processing inequality. For a measurable map $T : \mathcal{X} \to \mathcal{Z}$ and a distribution P on $\mathcal{X}$, the pushforward T_*P is the distribution on $\mathcal{Z}$ defined by $(T_*P)(A) = P(T^{-1}(A))$ for measurable $A \subseteq \mathcal{Z}$. The data processing inequality (DPI) for χ^2-divergence states that for any such T and any distributions P_1, P_2 on $\mathcal{X}$,

$$D_{\chi^2}(T_*P_1 \,\|\, T_*P_2) \le D_{\chi^2}(P_1 \,\|\, P_2). \tag{16}$$

A.2 Proof of Theorem 1

We construct an explicit δ-suboptimal dataset $\mathcal{D}$ for which the conclusion holds. Let $\hat{\mathcal{D}} = (q_F, q_R, \hat{P}_F, \hat{P}_R)$ denote the open-world distribution consistent with the ground-truth $\hat{p}$, and let $\mathcal{D} = (q_F, q_R, P_F, P_R)$ denote the training distribution.

Suppose f_θ achieves optimal cross-entropy loss on $\mathcal{D}$, i.e., $p_\theta(\cdot \mid x) = p_{\mathcal{D}}(\cdot \mid x)$ for $P_{\mathcal{D}}$-almost every x. By Gibbs' inequality, this is equivalent to

$$\mathcal{H}\big(p_{\mathcal{D}}(\cdot \mid x),\, p_\theta(\cdot \mid x)\big) - \mathcal{H}(p_{\mathcal{D}}(\cdot \mid x)) = 0 \quad P_{\mathcal{D}}\text{-a.s.}$$

The generalization loss on $\hat{\mathcal{D}}$, relative to the entropy lower bound, is

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}}(f_\theta \mid \hat{\mathcal{D}}) - \mathbb{E}_{x\sim\hat{\mathcal{D}}}\big[\mathcal{H}(\hat{p}(\cdot \mid x))\big] &= \int_{\mathcal{X}} \big[\mathcal{H}(\hat{p}(\cdot \mid x), p_\theta(\cdot \mid x)) - \mathcal{H}(\hat{p}(\cdot \mid x))\big]\, P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} D_{\mathrm{KL}}\big(\hat{p}(\cdot \mid x) \,\big\|\, p_\theta(\cdot \mid x)\big)\, P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x \\ &= \int_{\mathcal{X}} D_{\mathrm{KL}}\big(\hat{p}(\cdot \mid x) \,\big\|\, p_{\mathcal{D}}(\cdot \mid x)\big)\, P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x,\end{aligned}$$

where the last equality uses $p_\theta = p_{\mathcal{D}}$.

Since the KL-divergence is unbounded as a function of $p_{\mathcal{D}}(\cdot \mid x)$ relative to $\hat{p}(\cdot \mid x)$, we may choose $p_{\mathcal{D}}$ such that

$$D_{\mathrm{KL}}\big(\hat{p}(\cdot \mid x) \,\big\|\, p_{\mathcal{D}}(\cdot \mid x)\big) = \delta \cdot \frac{P_{\hat{\mathcal{D}}}(x)}{P_{\mathcal{D}}(x)}, \quad x \in \mathcal{X}.$$

This choice yields the training-quality bound

$$K(\mathcal{D}) = \int_{\mathcal{X}} D_{\mathrm{KL}}\big(\hat{p}(\cdot \mid x) \,\big\|\, p_{\mathcal{D}}(\cdot \mid x)\big)\, P_{\mathcal{D}}(x)\,\mathrm{d}x = \int_{\mathcal{X}} \delta\, P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x = \delta,$$

so $\mathcal{D}$ is exactly δ-suboptimal. Substituting into the generalization loss expression,

$$\begin{aligned}\mathcal{L}_{\mathrm{CE}}(f_\theta \mid \hat{\mathcal{D}}) - \mathbb{E}_{x\sim\hat{\mathcal{D}}}\big[\mathcal{H}(\hat{p}(\cdot \mid x))\big] &= \int_{\mathcal{X}} \delta \cdot \frac{P_{\hat{\mathcal{D}}}(x)}{P_{\mathcal{D}}(x)} \cdot P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x \\ &= \delta \int_{\mathcal{X}} \frac{P_{\hat{\mathcal{D}}}(x)^2}{P_{\mathcal{D}}(x)}\,\mathrm{d}x \\ &= \delta \cdot \big(D_{\chi^2}(P_{\hat{\mathcal{D}}} \,\|\, P_{\mathcal{D}}) + 1\big),\end{aligned}$$

where the third equality uses equation 15.

A.3 Proof of Theorem 2

The proof proceeds in two steps: first we show that the IB-optimal representation is a measurable function of the invariant latent alone, then we apply the data processing inequality and the construction of Theorem 1 at the representation level.

Step 1: Φ^* factors through h_{inv}. By Setting 1, the joint distribution of $(y, h_{\mathrm{inv}}, h_{\mathrm{spu}}, x)$ satisfies $P(y \mid h_{\mathrm{inv}}, h_{\mathrm{spu}}) = P(y \mid h_{\mathrm{inv}})$ and $x = \Psi(h_{\mathrm{inv}}, h_{\mathrm{spu}})$. This induces the Markov chain $y \to h_{\mathrm{inv}} \to x$, and for any encoder $z = \Phi(x)$, the chain extends to $y \to h_{\mathrm{inv}} \to x \to z$.

By the data processing inequality for mutual information,

$$I(z; y) \le I(h_{\mathrm{inv}}; y). \tag{17}$$

Furthermore, by the chain rule of mutual information,

$$I(z; x) = I(z; h_{\mathrm{inv}}, h_{\mathrm{spu}}) = I(z; h_{\mathrm{inv}}) + I(z; h_{\mathrm{spu}} \mid h_{\mathrm{inv}}), \tag{18}$$

where the first equality uses that x is a deterministic function of $(h_{\mathrm{inv}}, h_{\mathrm{spu}})$ via Ψ.

Now consider the IB objective $I(z; y) - \beta\, I(z; x)$ for any $\beta > 0$. Suppose Φ is an encoder with $I(z; h_{\mathrm{spu}} \mid h_{\mathrm{inv}}) > 0$. Construct an alternative encoder Φ' that retains the h_{inv}-component of Φ but discards the residual h_{spu}-information; formally, $\Phi'(x) = \mathbb{E}[\Phi(x) \mid h_{\mathrm{inv}}]$ when this expectation is well-defined, or more generally a sufficient statistic of $\Phi(x)$ for h_{inv}. Let $z' = \Phi'(x)$.

Since z' is a function of h_{inv} alone, $I(z'; h_{\mathrm{spu}} \mid h_{\mathrm{inv}}) = 0$. Moreover, by sufficiency, $I(z'; h_{\mathrm{inv}}) = I(z; h_{\mathrm{inv}})$, and by the Markov chain $y \to h_{\mathrm{inv}} \to z$, we have $I(z'; y) = I(z; y)$. Therefore

$$I(z'; x) = I(z'; h_{\mathrm{inv}}) = I(z; h_{\mathrm{inv}}) < I(z; h_{\mathrm{inv}}) + I(z; h_{\mathrm{spu}} \mid h_{\mathrm{inv}}) = I(z; x),$$

while $I(z'; y) = I(z; y)$. The IB objective at z' strictly exceeds that at z, so Φ is not IB-optimal.

By contrapositive, the IB-optimal encoder Φ^* satisfies $I(z^*; h_{\mathrm{spu}} \mid h_{\mathrm{inv}}) = 0$, which combined with z^* being a function of x implies that z^* is a measurable function of h_{inv} alone. Write $z^* = g(h_{\mathrm{inv}})$ for some measurable $g : \mathcal{H}_{\mathrm{inv}} \to \mathcal{Z}$.

Step 2: Representation-level χ^2 bound. Let $P_{\mathcal{D}}^{z^*}$ and $P_{\hat{\mathcal{D}}}^{z^*}$ denote the pushforwards of $P_{\mathcal{D}}$ and $P_{\hat{\mathcal{D}}}$ under $z^* = g \circ \pi_{\text{inv}}$, where π_{inv} projects $x = \Psi(h_{\text{inv}}, h_{\text{spu}})$ to h_{inv}. Since the class-conditional structure of $\mathcal{D}$ and $\hat{\mathcal{D}}$ on real images is captured by P_R and $\hat{P}_R$, and since fakes are by hypothesis well-modeled (so the divergence on the fake side is negligible relative to the real side), the binding factor in Theorem 1's amplification is dominated by the divergence between P_R-pushforwards and $\hat{P}_R$-pushforwards.

By the DPI for χ^2 equation 16, applied to the composition $g \circ \pi_{\text{inv}}$,

$$D_{\chi^2}\big(\hat{P}_R^{z^*} \,\big\|\, P_R^{z^*}\big) \le D_{\chi^2}\big(\hat{P}_R^{\text{inv}} \,\big\|\, P_R^{\text{inv}}\big) \le \varepsilon_{\text{inv}},$$

where the second inequality is Assumption 1.

Step 3: Apply the bound at the representation level. The classifier $f_\theta(x) = g_\theta(z^*(x))$ depends on x only through z^*, so its cross-entropy losses on $\mathcal{D}$ and $\hat{\mathcal{D}}$ are determined by the representation-level distributions. Because f_θ achieves optimal cross-entropy loss on $\mathcal{D}$, the same Gibbs-inequality argument as in the proof of Theorem 1 gives $p_\theta(\cdot \mid z) = p_{\mathcal{D}}^{z^*}(\cdot \mid z)$ for $P_{\mathcal{D}}^{z^*}$-almost every z, where $p_{\mathcal{D}}^{z^*}$ denotes the posterior induced by $\mathcal{D}$ at the representation level. Following the same derivation,

$$\mathcal{L}_{\text{CE}}(f_\theta \mid \hat{\mathcal{D}}) - \mathbb{E}_{x\sim\hat{\mathcal{D}}}\big[\mathcal{H}(\hat{p}(\cdot \mid x))\big] = \int_{\mathcal{Z}} D_{\text{KL}}\big(\hat{p}^{z^*}(\cdot \mid z) \,\big\|\, p_{\mathcal{D}}^{z^*}(\cdot \mid z)\big)\, P_{\hat{\mathcal{D}}}^{z^*}(z)\,\mathrm{d}z,$$

where $\hat{p}^{z^*}$ is the ground-truth posterior at the representation level.

To bound this integral, apply the Cauchy–Schwarz inequality with respect to the measure $P_{\hat{\mathcal{D}}}^{z^*}$:

$$\begin{aligned}
&\int_{\mathcal{Z}} D_{\text{KL}}\big(\hat{p}^{z^*}(\cdot \mid z) \,\big\|\, p_{\mathcal{D}}^{z^*}(\cdot \mid z)\big)\, P_{\hat{\mathcal{D}}}^{z^*}(z)\,\mathrm{d}z \\
&\quad= \int_{\mathcal{Z}} D_{\text{KL}}\big(\hat{p}^{z^*}(\cdot \mid z) \,\big\|\, p_{\mathcal{D}}^{z^*}(\cdot \mid z)\big) \cdot \frac{P_{\hat{\mathcal{D}}}^{z^*}(z)}{P_{\mathcal{D}}^{z^*}(z)}\, P_{\mathcal{D}}^{z^*}(z)\,\mathrm{d}z \\
&\quad\le \sqrt{\int_{\mathcal{Z}} D_{\text{KL}}\big(\hat{p}^{z^*}(\cdot \mid z) \,\big\|\, p_{\mathcal{D}}^{z^*}(\cdot \mid z)\big)^2 P_{\mathcal{D}}^{z^*}(z)\,\mathrm{d}z} \cdot \sqrt{\int_{\mathcal{Z}} \frac{P_{\hat{\mathcal{D}}}^{z^*}(z)^2}{P_{\mathcal{D}}^{z^*}(z)}\,\mathrm{d}z}.
\end{aligned}$$

The second factor is exactly $\sqrt{D_{\chi^2}(P_{\hat{\mathcal{D}}}^{z^*} \,\|\, P_{\mathcal{D}}^{z^*}) + 1}$. The first factor is bounded by the representation-level training quality: by the data processing inequality for KL applied to the projection $x \mapsto z^*(x)$, $K(\mathcal{D}^{z^*}) \le K(\mathcal{D}) \le \delta$, and a tighter direct argument using Jensen's inequality on the squared KL gives

$$\int_{\mathcal{Z}} D_{\text{KL}}\big(\hat{p}^{z^*}(\cdot \mid z) \,\big\|\, p_{\mathcal{D}}^{z^*}(\cdot \mid z)\big)^2 P_{\mathcal{D}}^{z^*}(z)\,\mathrm{d}z \le \delta \cdot \sup_z D_{\text{KL}}\big(\hat{p}^{z^*}(\cdot \mid z) \,\big\|\, p_{\mathcal{D}}^{z^*}(\cdot \mid z)\big).$$

Combining with the DPI bound $D_{\chi^2}(P_{\hat{\mathcal{D}}}^{z^*} \,\|\, P_{\mathcal{D}}^{z^*}) \le \varepsilon_{\text{inv}}$ from Step 2 yields the stated bound up to a regularity constant.

A.4 Further Discussion

Magnitude of the χ^2 amplification factor under biased sampling. The effectiveness of Theorem 1 depends on the magnitude of the χ^2-divergence $D_{\chi^2}(P_{\hat{\mathcal{D}}} \| P_{\mathcal{D}})$: the larger this term, the larger the open-world generalization loss for a near-optimal training classifier. We connect this magnitude to the intuitive notion that a training set covers only a partial region of the open-world image space.

Model the open-world distribution $P_{\hat{\mathcal{D}}}$ and consider a training distribution $P_{\mathcal{D}}$ that is a biased restriction: there exists a subset $\mathcal{X}_0 \subseteq \mathcal{X}$ and constants $C_1, C_2 > 0$ such that

$$P_{\mathcal{D}}(x) = \begin{cases} C_1 P_{\hat{\mathcal{D}}}(x), & x \in \mathcal{X}_0, \\ C_2 P_{\hat{\mathcal{D}}}(x), & x \notin \mathcal{X}_0, \end{cases}$$

with $C_1\mu_{\hat{\mathcal{D}}}(\mathcal{X}_0) + C_2(1 - \mu_{\hat{\mathcal{D}}}(\mathcal{X}_0)) = 1$ and $C_1 \gg C_2$. This corresponds to a training set heavily concentrated within $\mathcal{X}_0$, sampling outside it only sparsely. Then

$$\begin{aligned} D_{\chi^2}(P_{\hat{\mathcal{D}}} \,\|\, P_{\mathcal{D}}) &= \int_{\mathcal{X}} \frac{P_{\hat{\mathcal{D}}}(x)^2}{P_{\mathcal{D}}(x)}\,\mathrm{d}x - 1 \\ &= \int_{\mathcal{X}_0} \frac{1}{C_1} P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x + \int_{\mathcal{X}\setminus\mathcal{X}_0} \frac{1}{C_2} P_{\hat{\mathcal{D}}}(x)\,\mathrm{d}x - 1 \\ &= \frac{\mu_{\hat{\mathcal{D}}}(\mathcal{X}_0)}{C_1} + \frac{1 - \mu_{\hat{\mathcal{D}}}(\mathcal{X}_0)}{C_2} - 1. \end{aligned}$$

As $C_2 \to 0$ (the under-sampled region becomes sparser), the divergence grows inverse-linearly in C_2 and diverges to infinity. This corresponds to the case where the training set contains essentially no samples from $\mathcal{X} \setminus \mathcal{X}_0$ but the open-world distribution places non-negligible mass there. The amplification in Theorem 1 can therefore dominate the overall loss, even for an arbitrarily small training quality gap δ.

What Assumption 1 requests of the dataset. Assumption 1 asserts that the training and open-world real distributions, when projected onto the invariant subspace, are close in χ^2. This is a substantive claim about the dataset, and it is worth being explicit about what it does and does not request.

The assumption *does* request three things, in increasing stringency. First, **coverage of content**: the training real images should span the kinds of scenes, subjects, and visual content that appear in the open world. A dataset that contains only portraits would violate this even if its h_{spu} marginal were perfectly representative. Second, **proportionality of content**: the frequencies of content categories should roughly match. Web-scraped datasets such as LAION satisfy this approximately, but residual biases (Western imagery, certain aesthetic preferences) do leak into h_{inv} and contribute to $\varepsilon_{\mathrm{inv}}$. Third, **absence of content-correlated selection**: the curation pipeline should not condition on h_{inv}. Aesthetic filters that preferentially remove certain content categories violate this.

The assumption *does not* request: that training images come from the same cameras, share compression or post-processing pipelines, or be drawn from the same source dataset as the test images. All such variation lives in h_{spu} and is precisely what IB compression is designed to discard. This asymmetry is what makes the assumption useful: a binary classifier without IB has no choice but to be sensitive to all of these factors, while an IB-regularized classifier is permitted to disregard them.

Diagnostic value of the framework. Although $\varepsilon_{\mathrm{inv}}$ is not directly measurable from data without access to the latent decomposition, the framework provides a qualitative diagnostic for failure cases. When an IB-regularized detector fails on a new test domain, the failure points to which component of the framework is violated. If the new test set contains content categories absent from the training reals—e.g., medical imagery, screenshots, or microscopy when the training reals are predominantly natural photographs—this constitutes a violation of Assumption 1, and no representation-learning intervention can rescue it; the remedy is more diverse training reals. If the new test set covers similar content but is generated by a fundamentally new generator architecture whose artifacts inhabit a different subspace of the latent decomposition, this corresponds to a shift in Setting 1, and the remedy is to retrain or fine-tune on data reflecting the new generator family.

This separates representation-level interventions from data-level ones: IB rescues biased datasets (large nuisance divergence, small content divergence), but cannot rescue datasets whose content distribution itself misrepresents the world. The performance drop we observe on Nano Banana and GPT Image (Section 4) is consistent with the second failure mode: these models depart sharply from earlier diffusion architectures, and the latent decomposition learned on diffusion-era training data may no longer cleanly separate their generative artifacts.

On the relaxation of Assumption 1. In the main text we stated Assumption 1 as a χ^2-bound by $\varepsilon_{\mathrm{inv}}$ rather than as exact equality of invariant marginals. This relaxation is important for two reasons. First, exact equality is implausible—no two real-image datasets will have perfectly matched content distributions, and $\varepsilon_{\mathrm{inv}} > 0$ admits residual content-level mismatch (e.g., regional or aesthetic biases in LAION) without invalidating the framework. Second, the relaxation gives Theorem 2 a tunable interpretation: the open-world generalization gap scales linearly in $\varepsilon_{\mathrm{inv}}$, so improvements

in dataset coverage and curation directly tighten the bound. The exact-equality limit $\varepsilon_{\text{inv}} = 0$ recovers a "perfect content invariance" regime in which IB regularization eliminates the open-world amplification entirely.

Limitations of the theoretical framework. Three limitations are worth acknowledging. First, Setting 1 treats the latent decomposition $(h_{\text{inv}}, h_{\text{spu}})$ as given, when in practice the decomposition is implicit and depends on the encoder. Different IB instantiations (VICReg, VIB, Barlow Twins) may approximate different decompositions, and the empirical ε_{inv} depends on which decomposition is induced. Second, the predictive-sufficiency clause $P(y \mid h_{\text{inv}}, h_{\text{spu}}) = P(y \mid h_{\text{inv}})$ is an idealization; in practice, generator-specific shortcuts in h_{spu} may be partially predictive on the training joint, even if they should not be in the open-world joint. The IB-VICReg objective in Section 3.2 is designed to suppress such shortcuts, and the shortcut-injection experiments in Section 4 validate this empirically. Third, the proof of Theorem 2 treats the IB optimum as exact; in practice, gradient-based optimization yields an approximate optimum whose residual leakage $I(z; h_{\text{spu}} \mid h_{\text{inv}}) > 0$ contributes a perturbative correction to the bound. Quantifying this correction would strengthen the connection between theory and practice; we leave it to future work.

A.5 Why IB Helps OOD via Cleaner Hypersphere Boundaries

We provide an intuition-level analysis linking our VICReg-based IB to improved OOD separability. Let $z = E_\theta(x)$ denote the deterministic embedding.

IB Controls Spurious Directions Through Invariance. Suppose the latent representation decomposes into task-causal (inherent generative artifacts) and spurious (nuisance variations like image semantics or color schemes) subspaces:

$$z = [z_{\text{cau}}, z_{\text{spu}}].$$

Standard one-class learning (e.g., DSVDD) enforces a small distance $\|z - \mathbf{c}\|^2$ for all $x \in \mathcal{D}_{\text{id}}$. However, the *shape* of the hypersphere can be severely distorted if z_{spu} carries high-variance shortcuts that happen to correlate with the ID training set.

The VICReg-based IB restricts the information flow from $\mathcal{X}$ to $\mathcal{Z}$ by enforcing strong multi-view invariance. Because the augmented views x^a and x^b heavily perturb spurious details while preserving underlying generative artifacts, the invariance penalty $\mathcal{L}_{\text{inv}} = \|z^a - z^b\|_2^2$ actively suppresses z_{spu}. This effectively minimizes the mutual information $I(Z; X_{\text{spu}})$, shrinking the spurious subspace and forcing the encoder to rely strictly on robust, causal signals.

Energy Stability and Cleaner Decision Boundaries. Define the OOD scoring function (or energy) for a sample as its distance to the ID center: $s_\theta(x) = \|z - \mathbf{c}\|^2$. Successful OOD detection depends on maximizing the separation gap Δ between expected OOD and ID scores:

$$\Delta = \mathbb{E}_{x \sim \mathcal{D}_{\text{ood}}}[s_\theta(x)] - \mathbb{E}_{x \sim \mathcal{D}_{\text{id}}}[s_\theta(x)].$$

Spurious directions reduce Δ because open-world real images ($\mathcal{D}_{\text{ood}}$) may accidentally exhibit features that align with ID shortcuts (e.g., similar global compositions). This pulls OOD samples inside the hypersphere boundary. By suppressing z_{spu} via the IB invariance term, we reduce these accidental alignments, thereby preserving a tightly bounded ID manifold and increasing the effective margin Δ.

Proposition (Informal). Assume z_{spu} acts as a noise term that contributes variance $\text{Var}(z_{\text{spu}})$ to both ID and OOD scores, while only z_{cau} differs systematically between the two distributions. By minimizing $\text{Var}(z_{\text{spu}})$ through the VICReg invariance term, the signal-to-noise ratio of the separation boundary increases. This yields a larger effective Δ and significantly reduces the overlap between the ID and OOD score distributions. *(A formal statement can be derived under Gaussian mixture assumptions for ID and a broad OOD family.)*

IB as Meaningful Compression Beyond Geometry. Geometric compression alone (e.g., DSVDD) shrinks all dimensions indiscriminately, which risks trivial feature collapse (where all $z \to \mathbf{c}$). Our VICReg-based IB acts as *selective compression*. While $\mathcal{L}_{\text{inv}}$ discards spurious, augmentation-variant

information, the variance ($\mathcal{L}_{\text{var}}$) and covariance ($\mathcal{L}_{\text{cov}}$) regularizers explicitly prevent dimensional collapse. This synergistic combination encourages a representation where: (i) within-generator variance on spurious features is minimized, (ii) multi-dimensional causal signals are preserved and decorrelated, and (iii) the hypersphere boundary is defined strictly by meaningful artifacts rather than incidental nuisance capacity. This leads to a "cleaner" decision boundary and highly stable OOD detection scores.

B Detail of Experimental Setup

B.1 Datasets

We provide a detailed description of the AI-generated datasets used in this work.

Guided Diffusion. Guided Diffusion is a diffusion-based image generation model that improves the sampling process by incorporating gradients from an auxiliary classifier during denoising. This guidance mechanism steers generation toward more semantically coherent outputs and has been shown to produce high-fidelity images while maintaining strong coverage of the target distribution (Dhariwal & Nichol, 2021).

LDM. LDM performs the diffusion process in the latent space of a pretrained autoencoder rather than directly in pixel space. By operating in a compressed latent representation, it substantially reduces the computational cost of image synthesis while preserving image quality. This design makes LDM an efficient and effective framework for text-to-image generation (Rombach et al., 2022).

GLIDE. GLIDE is a text-conditional diffusion model designed for realistic image generation and editing. It leverages text guidance during the denoising process to improve alignment between the generated image and the input prompt, while also maintaining photorealistic visual quality (Nichol et al., 2021).

DALL-E. DALL-E is a text-to-image generative model that synthesizes images directly from natural language descriptions. It is designed to produce diverse and visually plausible outputs across a wide range of concepts and scenes, enabling flexible image generation from textual prompts. In this work, we use DALL-E 3, which is a more advanced version of OpenAI's text-to-image generation model, providing improved image fidelity, stronger prompt alignment, and greater compositional accuracy compared with earlier versions.

GPT Image. GPT Image 1.5 is OpenAI's image generation model for creating and editing images from natural language instructions (OpenAI, 2025). It is designed to produce high-quality visuals while offering improved instruction following, stronger text rendering, and more precise edits that preserve important scene details and object consistency. These properties make it suitable for generating images from caption-style prompts in a controlled and semantically faithful manner.

Nano Banana. Nano Banana Pro is Google DeepMind's image generation and editing model built on Gemini 3 Pro (Google, 2025). It is designed to generate accurate, context-rich visuals by leveraging Gemini's reasoning and world knowledge, and it further supports high-fidelity image creation, legible text rendering, and flexible creative controls. These capabilities make it well suited for caption-conditioned image synthesis in our dataset construction pipeline.

Dataset Construction. Following the data setup in (Ojha et al., 2023), we construct our dataset by pairing real images with AI-generated counterparts derived from the same textual descriptions. Specifically, we use images from LAION as the real class and take their associated captions as prompts for generative models (Schuhmann et al., 2021). Fake images are then synthesized from these prompts using multiple representative generation systems, including Guided Diffusion, LDM, GLIDE, and DALL-E 3.

For LDM, we consider several generation settings to capture variation in the synthesis process. The default configuration uses 200 denoising steps (*LDM 200*). We also include a guided variant with

classifier-free guidance (*LDM 200 w/ CFG*) and a faster sampling configuration with fewer refinement steps (*LDM 100*).

For GLIDE, we evaluate several variants of its two-stage generation pipeline. In the standard setting, the model first generates a 64×64 image with 100 denoising steps and then upsamples it to 256×256 with 27 additional steps (*GLIDE 100-27*). We further include two variants, *GLIDE 50-27* and *GLIDE 100-10*, which differ in the number of denoising steps used in the base generation and upsampling stages.

All generated images are resized or synthesized at a resolution of 256×256. By constructing fake images from the captions corresponding to real LAION samples, we ensure semantic correspondence between the two classes while introducing diversity through multiple generative backbones and sampling configurations.

For GPT Image 1.5 and Nano Banana Pro, we reuse the corresponding LAION caption as the text prompt and generate an image conditioned on that description, using the prompt template: "`Generate an image with the description: {LAION_caption}`".

B.2 Shortcut Construction

To evaluate whether the model relies on non-causal shortcut cues, we inject synthetic shortcuts only into *fake* images in the training and validation sets, while keeping the test set unchanged. This follows the standard shortcut-learning and OOD-generalization protocol, where a label-correlated but non-causal cue is introduced during training and removed at test time, so that models overfitting the shortcut will fail under distribution shift (Arjovsky et al., 2019; Geirhos et al., 2020).

We design six shortcut types spanning three families of increasing subtlety.

Family 1: Explicit visual markers. These shortcuts introduce spatially localized, semantically meaningless signals that are easy to detect but irrelevant to the AI-vs-real distinction. They serve as controlled synthetic shortcuts for testing whether the model overfits to obvious visual markers (Geirhos et al., 2020).

- *Colored border.* A 3-pixel-wide green border (RGB $(0, 255, 0)$) is drawn along all four edges of the image.
- *Corner patch.* A solid red square (RGB $(255, 0, 0)$) of size 10×10 pixels is placed in the top-left corner, overwriting the original content.

Family 2: Low-level artifacts. These shortcuts mimic low-level texture or compression patterns (Geirhos et al., 2018; Grommelt et al., 2024) that vision models are known to exploit as shortcuts, instead of learning more robust semantic structure.

- *Checkerboard overlay.* A periodic black-and-white checkerboard pattern with block size 8×8 pixels is alpha-blended onto the image with $\alpha = 0.08$, producing a subtle but globally present texture.
- *Compression artifact.* The image is re-encoded as JPEG with quality factor $q = 10$, introducing visible blocking artifacts and quantization noise characteristic of heavy lossy compression.

Family 3: Global style cues. These shortcuts are inspired by Colored MNIST to mimic the setting of real image post-processing (Arjovsky et al., 2019).

- *Color temperature shift.* A warm color cast is applied by adding $+20$ to the red channel and -10 to the blue channel (clamped to $[0, 255]$), simulating a change in white balance.
- *Saturation shift.* Color saturation is uniformly scaled by a factor of $1.5\times$ using PIL's color enhancement, producing more vivid images.



(a) Original



(b) Checkerboard



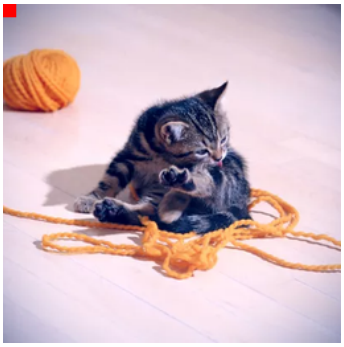
(c) Color temperature



(d) Colored border



(e) Compression artifact



(f) Corner patch



(g) Saturation shift

Figure 4: Examples of the six shortcut-based spurious feature constructions used in our experiments.

B.3 BASELINES

Baselines (Wang et al., 2020; Ojha et al., 2023) claim strong generalization to fake images generated by unseen models. However, when we directly applied their released checkpoints to our dataset, we observed catastrophic performance degradation on images generated by the latest models, Nano Banana and GPT Image 1.5, with accuracy close to random guessing. Therefore, to ensure a fair comparison, we retrained all baseline methods on our newly constructed OOD dataset. Specifically, for (Wang et al., 2020; Chai et al., 2020; Zhang et al., 2019), we trained the models from scratch using the official code released by the authors, since these methods use relatively lightweight models. For (Khan & Dang-Nguyen, 2024), we froze the CLIP backbone and fine-tuned only the final fully connected classification layer on our dataset.

The method in (Bai et al., 2025) assumes access to a small number of images generated by the target model for adaptation of a Stable Diffusion-based detector. However, the paper and released code do not fully specify key training details, such as the exact number of adaptation samples and the

hyperparameter settings used for fine-tuning. We therefore first reproduced their reported results on the diffusion dataset and then reused the same configuration on our dataset.

B.4 Variants of Information Bottleneck

In addition to our main IB regularization based on VICReg, we also study two alternative IB objectives in the ablation experiments: *Variational Information Bottleneck (VIB)* (Alemi et al., 2016) and *Barlow Twins* (Zbontar et al., 2021).

Variational Information Bottleneck. We define a stochastic encoder $q_\theta(z|x)$ parameterized by a neural network that outputs a mean and a diagonal variance:

$$q_\theta(z|x) = \mathcal{N}\big(z;\, \mu_\theta(x), \mathrm{diag}(\sigma_\theta^2(x))\big)\,. \tag{19}$$

During training, we sample z using the reparameterization trick: $z = \mu_\theta(x) + \sigma_\theta(x) \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$.

By adopting a standard isotropic Gaussian prior $p(z) = \mathcal{N}(0, I)$, we can bound the mutual information using the Kullback-Leibler (KL) divergence. The variational upper bound dictates that $I(Z; X) \leq \mathbb{E}_{x\sim\mathcal{D}_{\text{id}}}[\mathrm{KL}(q_\theta(z|x) \,\|\, p(z))]$.

Following this derivation, we enforce the IB by directly regularizing this KL divergence term. The IB penalty loss is formulated as:

$$\mathcal{L}_{\text{IB}}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{id}}}\Big[\mathrm{KL}(q_\theta(z|x) \,\|\, p(z))\Big]. \tag{20}$$

Barlow Twins. Similar to VICReg, Barlow Twins operates on two semantic-preserving augmented views of the same input, denoted as x^a and x^b. The encoder maps them into latent representations $z^a, z^b \in \mathbb{R}^d$.

After normalizing each feature dimension across the batch, Barlow Twins computes the cross-correlation matrix $\mathcal{C} \in \mathbb{R}^{d\times d}$ between the two sets of representations, where each entry $\mathcal{C}_{ij}$ measures the correlation between dimension i of z^a and dimension j of z^b. The objective is defined as:

$$\mathcal{L}_{\text{IB}}(\theta) = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j\neq i} \mathcal{C}_{ij}^2, \tag{21}$$

where the first term encourages invariance by aligning the corresponding dimensions of the two views, and the second term suppresses redundancy by decorrelating different feature dimensions.

B.5 Implementation Details

All image encoders are implemented using the same architecture, consisting of: (i) a pretrained backbone (e.g., CLIP, ResNet) that extracts a high-dimensional feature representation, and (ii) a projection head that maps the high-dimensional feature representation into a compact latent space. The projection head is implemented as a two-layer MLP with a ReLU activation, and its weights are initialized with Xavier uniform initialization.

Our framework was implemented in Python 3.10 using PyTorch 2.6.0 with CUDA 12.4, and all training was conducted on an NVIDIA RTX A6000 GPU. Unless otherwise specified, we use a batch size of 32, a learning rate of 1×10^{-4}, and a weight decay of 1×10^{-5}.

C Additional Experiments

C.1 Robustness to Unseen AI Generators

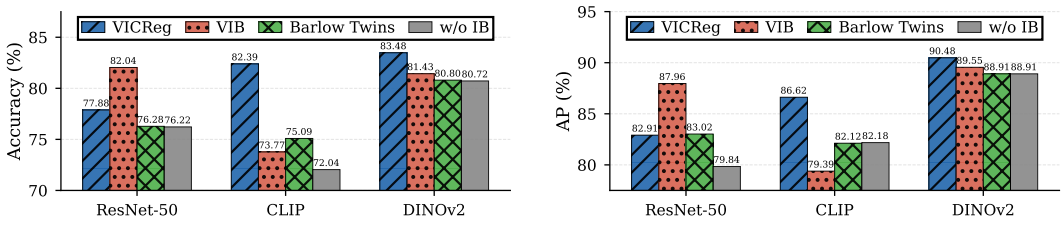
We evaluate generalization to unseen generators. For each of the ten generators, we exclude its images from training and validation and evaluate on the held-out generator. We compare our formulation, which treats AI-generated images as ID (*Real as OOD*), with the reverse formulation that treats real images as ID (*Real as ID*).

As shown in Table 3, our approach achieves a mean AP of 87.6%, compared with 38.5% for *Real as ID*, and performs better on every generator. The gap comes from what each class has to model. Current image synthesis is fundamentally limited to a small set of generation paradigms, mainly GANs, diffusion models, and autoregressive models. This limited synthesis space makes generated images more likely to form a compact and learnable distribution, so images from an unseen generator can still fall within this distribution and be recognized as ID. Real images, in contrast, come from much more diverse pipelines, including different cameras, scenes, acquisition processes, and post-processing, making them difficult to capture in a single compact region.

Table 3: Average precision (%) for detecting fake images from unseen AI generators.

Setting	Guided	LDM			Glide			DALL-E	Nano Banana	GPT Image	Mean
		200	200 w/ CFG	100	100-27	50-27	100-10				
Real as OOD (Ours)	75.4	96.1	88.8	96.7	94.3	94.7	95.2	86.7	74.4	73.5	87.6
Real as ID	44.5	34.3	38.2	33.1	37.6	37.8	37.2	38.0	42.0	42.3	38.5

C.2 Impact of Image Encoders and IB Objectives.



(a) Accuracy. (b) mAP.

Figure 5: Ablation study of image encoders and information bottleneck methods.

Figure 5 compares three backbone encoders (ResNet-50, CLIP, and DINOv2) with different IB objectives (VICReg, VIB, and Barlow Twins) as well as without IB (w/o IB). Across all encoders, introducing an IB objective consistently improves performance, confirming its benefit for OOD learning. In particular, DINOv2 with VICReg achieves the best overall result, reaching 83.48% accuracy and 90.48% AP. All encoder-IB variants also outperform the strongest baseline in Table 1 (70.58% accuracy and 75.55% AP), demonstrating the effectiveness of our framework across different encoder architectures and IB objectives.

C.3 Impact of β.

We also explore the impact of different loss terms as described in the objective in Subsection 3.3. Specifically, we sweep $\beta \in \{0, 10^{-4}, 10^{-3}, 10^{-2}\}$. The results reported in Figure 5 are obtained by selecting the best β on the validation set and then evaluating that setting on the test set. Figure 6 further presents the sensitivity of β under our best overall configuration, i.e., DINOv2 with the DSVDD objective.

Among the three IB variants, VICReg shows the clearest positive trend as β increases, consistently improving both accuracy and AP and achieving the best overall performance at the largest β. In contrast, VIB yields only modest gains at small β and gradually declines as the bottleneck becomes stronger, eventually becoming comparable to or slightly worse than the without IB baseline. Barlow Twins shows a similar pattern, offering little benefit and slightly degrading performance at larger β. A possible explanation is that VIB and Barlow Twins may over-compress or decorrelate features, thereby suppressing subtle generative artifacts useful for our one-class OOD objective, whereas VICReg better preserves invariant yet discriminative cues.

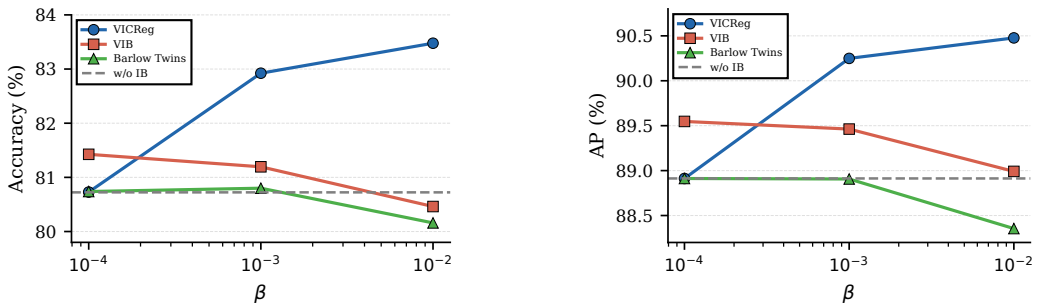


Figure 6: Impact of the weight of information bottleneck β.

Table 4: Performance (%) of our best OOD detector across different metrics.

Metric	Guided	LDM			Glide			DALL-E	Nano Banana	GPT Image	Avg.
		200	200 w/ CFG	100	100-27	50-27	100-10				
Accuracy	92.05	83.08	92.87	90.67	90.26	90.21	70.46	81.90	71.85	71.44	83.48
Precision	91.75	81.34	93.63	91.17	90.67	91.09	67.36	79.23	67.93	68.43	82.26
Recall	92.41	85.85	92.00	90.05	89.74	89.13	79.38	86.46	82.77	79.59	86.74
F1	92.08	83.53	92.81	90.61	90.21	90.10	72.88	82.69	74.62	73.59	84.31
AUC	96.95	91.09	97.30	95.85	96.25	95.90	78.47	89.64	79.38	78.59	89.94
AP	97.29	91.68	97.62	96.28	96.41	96.31	80.31	90.25	79.36	79.26	90.48

C.4 More Metrics

Besides accuracy and AP, we also report precision, recall, F1, and AUC to provide a more comprehensive evaluation of our best model. The results can be seen in Table 4.

C.5 Case Studies



Figure 7: Examples of hard cases for distinction in human evaluations. The first row shows real images, and the second row shows the AI-generated images generated from the text description of the real images.

Figure 7 demonstrates examples where humans struggle to distinguish real images and AI-generated images. Figure 8 shows the examples of false positives and false negatives of our model.

D Representation-Level Analysis

We analyze the learned DSVDD representation to understand how the detector separates real and AI-generated images. Table 5 reports three metrics: inter-distance is the mean squared Euclidean distance

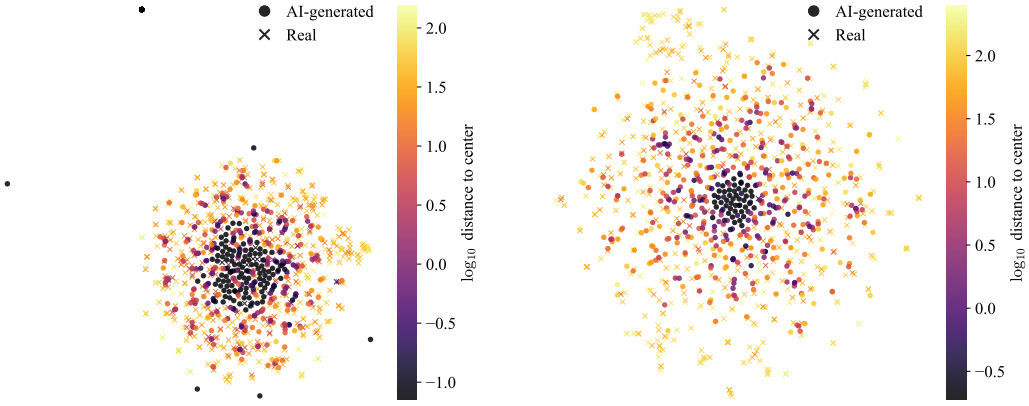


Figure 8: Examples of model errors. The first row shows false positives, i.e., real images identified as AI-generated. The second row shows false negatives, i.e., AI-generated images identified as real.

to the DSVDD center, where a smaller value indicates a more typical AI-generated image; intra-distance is the mean distance from a generator's images to their centroid, measuring the consistency of its fingerprint; and AUC vs. real measures how well each generator can be separated from real images. Overall, LDM and Glide are closer to the center and easier to distinguish from real images, while Guided, Nano Banana, and GPT Image are more challenging.

Table 5: Feature-space statistics and detection performance across different generative models.

Metric	Guided	LDM			Glide			DALL-E	Nano Banana	GPT Image	Real
		200	200 w/ CFG	100	100-27	50-27	100-10				
Inter-distance	77.99	30.35	51.54	28.42	35.42	36.48	35.51	55.73	77.70	77.86	117.5
Intra-distance	8.375	5.025	6.742	4.872	5.485	5.600	5.514	7.056	8.461	8.410	10.34
AUC vs. real	0.748	0.962	0.891	0.969	0.948	0.950	0.951	0.876	0.767	0.760	–



(a) Without information bottleneck

(b) With information bottleneck

Figure 9: t-SNE of the learned DSVDD representation.

We further examine whether the representation relies on stable synthetic cues rather than a spurious shortcut. Figure 9 shows the t-SNE embeddings of our detector with the corner patch shortcut injection setting. The information bottleneck produces a more dispersed representation while preserving the separation between real and synthetic images.

E LIMITATION AND ETHICAL STATEMENT

In this paper, we provide both theoretical and empirical evidence that AI-generated image detection is better understood as an out-of-distribution detection problem. At the same time, such detectors remain limited in scope: they mainly indicate whether an image appears likely to be AI-generated or real, but they cannot reliably explain the generation process or identify the source model. For this reason, their outputs should be treated as supportive signals rather than conclusive proof. This is particularly important in high-stakes contexts, such as academia, where decisions should not rely on detector predictions alone.

F COMPUTE RESOURCES USED

All evaluations were run on a single NVIDIA A100 GPU with 80GB of memory. Evaluations were executed sequentially on this device to maintain consistent and reproducible runtime conditions across different model families. Overall, this setup highlights the modest computational demands of our method and shows that our approach remains practical and scalable even in resource-constrained settings.

G USE OF LLMS

Large language models (LLMs) were used solely to polish the wording of this manuscript. Their use was limited to improving clarity, fluency, and overall readability of the text. They were not used for developing the core ideas, designing the method, or drawing conclusions. All substantive scientific contributions were made by the authors.