Prior-Conditioned Gaussian Discriminants for Generalizable AI-generated Image Detection

Shashank Kotyan, Makoto Shing, Yuki Imajuku, Rujikorn Charakorn, and Tarin Clanuwat

Sakana AI
shashank@sakana.ai
https://sakana.ai

Abstract. Diffusion-based generators have made synthetic images ubiquitous, but detectors often fail under simultaneous shifts in generator, prompt/style, and source-domain. We study AI-generated image detection as a transfer system described by training prior, frozen representation, and decision rule, and ask when classifier head training adds value beyond what is already separable in modern features. As a controlled diagnostic, we fit a prior-conditioned Gaussian discriminant ladder: closed-form heads built from first- and second-order feature statistics under nested covariance assumptions. On Percept-Lens, a unified protocol over 39 public datasets (7.1 million images), the best rung is frequently competitive with, and sometimes exceeds, released detector heads when matched on both prior and encoder. We further quantify strong sensitivity to the training prior, data-efficiency of moment-based heads, and representation dependence of Gaussian shift metrics, motivating (prior, encoder, head)-level reporting and stronger analytical baselines for AIGI transfer.

Keywords: AI-generated Image Detection · Distribution Shift · Gaussian Discriminant Analysis · Few-Shot Transfer · Transfer Learning

1 Introduction

Diffusion-based generative models [47,53–55] have made high-quality AI-generated imagery widely accessible. These models have practical benefits, but they also complicate trust in visual media and enable misuse of synthetic content [17,24,46]. As a result, AI-generated image (AIGI) detection has become a core recognition problem with direct implications for safety and privacy.

Modern detectors must transfer across shifts in *both* generator family and image-source domain. In practice these shifts co-occur (new generators, new prompting styles, new platforms, and post-processing), and performance can collapse even when in-distribution metrics look strong. Moreover, collecting large-scale labeled data from newly emerging generators is often expensive or infeasible. Together, these conditions make AIGI detection a concrete *transfer / low-shot recognition* problem, where a deployed detector must operate under joint shift and sometimes with only a small support set for calibration.

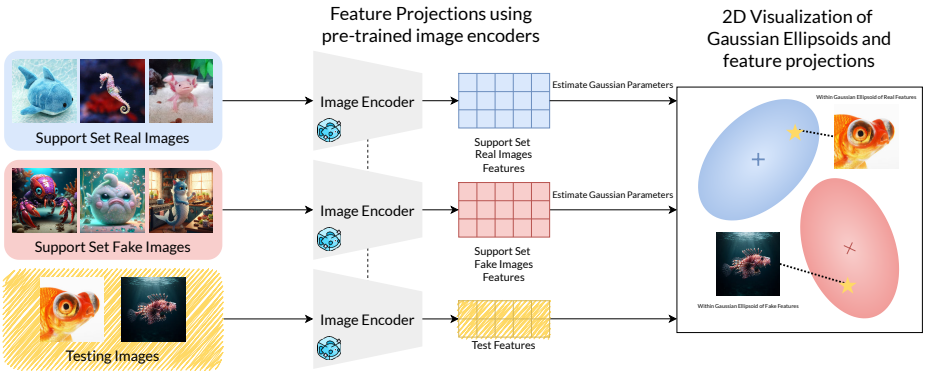

Fig. 1: Training-prior-conditioned Gaussian discriminants in a frozen feature space. A frozen image encoder ϕ maps images x to embeddings z. Given a (possibly few-shot) support set $\mathcal{S}$ drawn from a *training prior*, we estimate class-conditional moments (μ_c, Σ_c) and instantiate a ladder of *classical* closed-form heads under nested covariance assumptions. By keeping the encoder and prior fixed, the ladder isolates what the decision rule contributes.

Existing evaluations [16, 19, 80] typically probe only part of this space, frequently reusing narrow prompt regimes and closely related sources. Such evaluations can conflate genuine generalization with prior-specific shortcuts. We take a stricter, system-level view. A reported number is a property of a *system* consisting of a training prior, a frozen image encoder, and a decision rule on top. Here, *training prior* denotes the data distribution induced by a public training dataset (e.g., `CommunityForensics` [50]); it is not a Bayesian prior over model parameters. If generalization fails, it is not obvious whether the bottleneck is the representation, the head optimization, or the prior itself.

To isolate the decision-rule factor while keeping the prior and representation fixed, we fit *training-prior-conditioned Gaussian discriminants*, which are classical closed-form heads obtained from first- and second-order statistics of frozen encoder features (Figure 1). We organize these rules into a *Gaussian discriminant ladder* spanning nested covariance assumptions (isotropic, diagonal, shared, class-specific). Because each rung has an analytic solution, the ladder is a reproducible, hyperparameter-light baseline. The identity of the best-performing rung also indicates which low-order feature statistics transfer across generator and domain shifts. Although Gaussian discriminants are classical, we argue that they are underused as diagnostic baselines in AIGI detection, and we find that they can be competitive under matched (prior, encoder) conditions.

Empirically, across 39 public test sets (7.1 million images), at least one rung is often competitive with trained heads under *matched* priors and frozen encoders. As a retrospective diagnostic upper bound, the best rung can even surpass released heads in several settings. Our decomposition is not fully symmetric across the three factors. The most direct intervention is a matched head replacement, while

the encoder and prior experiments are controlled one-factor sweeps. We use the procedure as a source-side audit. If a trained head does not outperform the best rule in the ladder on the same final representation and support prior, classifier training is unlikely to be the source of OOD transfer gains. If it succeeds, it signals structure beyond second moments.

Contributions.

Prior-conditioned Gaussian discriminant ladder (diagnostic baseline). We instantiate a ladder of *classical* closed-form Gaussian heads on frozen encoder features, varying covariance assumptions (isotropic → diagonal → shared full → class-specific full). The best-transfer rung provides an interpretable diagnostic of which low-order feature statistics are preserved under shift.
Post-hoc controlled evaluation under matched (prior, encoder) conditions. On Percept-Lens, a unified evaluation *protocol* built entirely from existing public datasets, we compare released detector heads against the ladder using the same training prior and the same frozen encoder features. This isolates when learned heads add value beyond low-order geometry.
Empirical drivers of generalization under joint shift. We quantify sensitivity to the training prior under a fixed encoder, show that moment-based heads can adapt with few labeled samples on strong frozen encoders, and demonstrate that Gaussian shift metrics are strongly representation-conditioned.

2 Related Works

2.1 AI-generated Image (AIGI) detection under Distribution Shift

The literature on AIGI detection spans artifact-driven detectors [29, 64, 73] and representation-based detectors [16, 48, 50]. Early work focused on GAN-era artifacts and domain-specific settings such as face forgeries [29, 64, 73]. Diffusion models increased photorealism and prompt diversity, and exposed a persistent failure mode: detectors trained on narrow synthetic prior exploit shortcuts that do not transfer across generators or domains [16, 19, 68, 80]. One response is to broaden training priors by scaling generator coverage [19, 50]. Another is to use foundation-model features and lightweight heads [16, 48, 78]. Concurrent calibration work adjusts detector logits or thresholds after training under shift [72]. Our intervention instead replaces the classifier in feature space while matching the prior and encoder. Thus our emphasis is not another trained detector, but what is already separable in frozen encoder space and what the learned head adds under the *same* prior. This is closely related to the observation that many OOD detectors can be expressed as generative scoring rules on features, but the implications for AIGI detection under joint shift have not been carefully isolated.

2.2 Analytical Gaussian Discriminants and Feature-Space Scoring

Gaussian discriminant analysis and Mahalanobis scoring are classical tools for uncertainty and OOD detection [41]. Recent AIGI work has connected few-

shot / prototypical inference to Euclidean nearest-centroid rules [67], and has explored orthogonal subspace decompositions to preserve pretrained structure under adaptation [70].

We build on these ideas, but use them as a *diagnostic ladder*: isotropic, diagonal, tied, and class-specific covariance assumptions correspond to increasing geometric expressivity. The decision rules themselves are not new. Our contribution is to condition them on explicit training priors and use the ladder covariance model to interpret what transfers, and what does not, across generator and domain shifts.

3 Preliminaries

3.1 Problem setup

We study binary detection of AI-generated images. Given a training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i are images and $y_i \in \{0, 1\}$ are labels for real ($y = 0$) and synthetic ($y = 1$) images, the goal of AIGI detection is to learn a classifier that can accurately predict the label of a previously unseen image.

3.2 Feature Extraction and Trained Classifiers

We use a frozen image encoder $\phi : \mathcal{X} \to \mathbb{R}^D$ from a pre-trained foundation model to map each image x to a D-dimensional feature vector $z = \phi(x)$. A standard approach trains a discriminative head, such as a linear probe, on these features by minimizing a loss function (e.g., binary cross-entropy) via gradient descent: $\min_{w,b} \sum_{i=1}^{n} \text{CE}(\sigma(w^\top z_i + b), y_i)$ where $\sigma(\cdot)$ is the sigmoid activation function and $w \in \mathbb{R}^D$ and $b \in \mathbb{R}$ define a separating hyperplane.

More complex models, such as multi-layer perceptrons (MLPs), can learn non-linear decision boundaries. This process iteratively searches for an optimal boundary based on the training data. Although powerful, these methods can overfit to artifacts specific to the generative models in the training set, potentially limiting generalization to unseen generative processes [50, 70].

We compare (i) *trained* discriminative detectors that optimize a supervised objective on a given prior and (ii) *Gaussian discriminant rule* classifiers fitted on frozen features (Section 4). The audit applies to released systems that expose the feature vector used by a classifier, including frequency/statistical detectors [27, 69], reconstruction-based detectors [18, 19], LoRA-adapted detectors [70], or end-to-end detectors [16, 64] after their final representation is fixed. Methods that expose only a scalar anomaly score can still be evaluated by Percept-Lens as complete detectors, but their internal head cannot be replaced by the ladder. The split is diagnostic: if a trained head underperforms a closed-form rule under OOD shift, the learned decision surface and/or training prior is implicated rather than only the encoder capacity.

4 Prior-Conditioned Gaussian Discriminants as Few-Shot Discriminant Heads

We use Gaussian class-conditional models as a *controlled diagnostic* of what information is already separable in a frozen feature space. Motivated by classical generative classification and recent analyses of contrastive representations using mixture models [15], we approximate

$$p(z \mid y = c) \;\approx\; \mathcal{N}(\mu_c, \Sigma_c), \quad c \in \{0, 1\}. \tag{4.1}$$

Given a support set $\mathcal{S}$ drawn from a training prior, $\mathcal{S} \subseteq \mathcal{D}_{\text{train}}$, we estimate (μ_c, Σ_c) by sample moments. We also define the pooled covariance $\Sigma_p = \frac{(n_0-1)\Sigma_0+(n_1-1)\Sigma_1}{n_0+n_1-2}$, used when assuming a shared covariance structure. In high-dimensional embeddings, covariance inversion can be ill-conditioned. We therefore use standard regularization (diagonal loading and shrinkage) when computing Σ^{-1} and log-determinants (Appendix D). The Gaussian assumption is diagnostic rather than literal. Appendix G shows mostly near-Gaussian marginal summaries on PE-Core features, but also clear failures such as `FourierSpectrumDiscrepancies` [25] and `DiffusionForensics`-fake [65]. Multimodal or heavy-tailed regimes are exactly where learned non-linear heads may add value.

This yields a ladder of *classical* discriminant rules in closed form, and the best-performing covariance assumption becomes an interpretable indicator of which feature statistics transfer across domains, bypassing iterative training. (The same estimation extends to N-way K-shot classification. We focus on binary AIGI detection.)

In this article, we consider Cosine Nearest Centroid Matching (Cos-NCM), Gaussian Naive Bayes (GNB), Mahalanobis Nearest Centroid Matching (Mah-NCM), and Quadratic Discriminant Analysis (QDA) as Gaussian discriminants. We also include squared Euclidean nearest-centroid matching (Euc-NCM), which mirrors the inference rule used in prototypical-network detectors (e.g., [67]) but is applied directly in the frozen encoder space without metric learning. Tab. 1 provides an overview of Gaussian discriminants in analytical form, while Figure 2 provides a visual explanation.

Euclidean Nearest Centroid Matching (Euc-NCM). Assuming an isotropic shared covariance $\Sigma_c = \sigma^2 I$, the Gaussian log-likelihood (up to additive constants) is proportional to the squared Euclidean distance to the class mean. Prediction reduces to nearest-centroid classification in the frozen feature space, matching prototypical inference without metric learning.

Cosine Nearest Centroid Matching (Cos-NCM). When features are ℓ_2-normalized, isotropic Gaussian scoring is monotone in cosine similarity. Cos-NCM predicts the class whose centroid has the highest cosine similarity to the query.

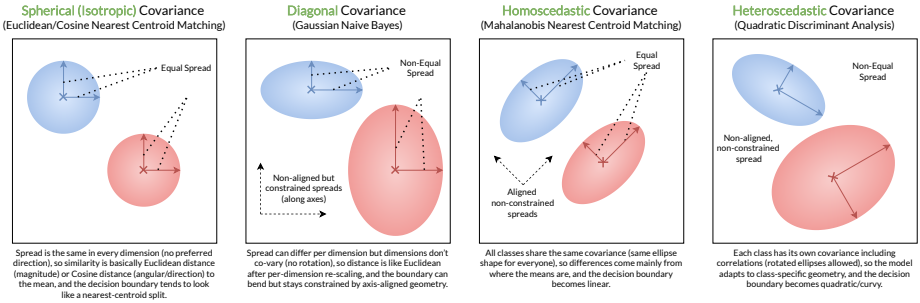

Fig. 2: Gaussian discriminant ladder as nested covariance assumptions. Each rung is a classical generative classifier obtained by restricting the class-conditional covariance: isotropic (Euc-/Cos-NCM), diagonal (GNB), shared full covariance (Mah-NCM), or class-specific full covariance (QDA). This nested family makes it possible to diagnose how much transferable signal is captured by first- and second-order feature statistics.

Table 1: Gaussian discriminant ladder used as analytical baselines. Each rung is a generative classifier obtained by restricting the class-conditional covariance (isotropic, diagonal, homoscedastic, or heteroscedastic), which determines the geometry of the induced decision boundary. We use this ladder to isolate head effects on a fixed frozen representation rather than to propose a new detector.

Gaussian Discriminant	Covariance Structure Assumption		Decision Boundary	Characteristic
Euclidean Nearest Centroid Matching (Euc-NCM)	Isotropic	$\Sigma_{c\in\{0,1\}} = \sigma^2 I$	Linear	Assumes spherical class distributions.
Cosine Nearest Centroid Matching (Cos-NCM)	Isotropic	$\Sigma_{c\in\{0,1\}} = \sigma^2 I$	Linear	Assumes spherical class distributions; decision is based on angular proximity, suitable for normalized embeddings.
Gaussian Naive Bayes (GNB)	Diagonal	$\Sigma_{c\in\{0,1\}} = \mathrm{diag}(\sigma^2_{c,j})$	Non-Linear	Simplest model; assumes feature independence, useful for identifying decorrelated representations.
Mahalanobis Nearest Centroid Matching (Mah-NCM)	Homoscedastic	$\Sigma_0 = \Sigma_1 = \Sigma_p$	Linear	Assumes classes share the same elliptical shape. It accounts for feature correlations by whitening the space, yielding a correlation-aware linear boundary.
Quadratic Discriminant Analysis (QDA)	Heteroscedastic	$\Sigma_0 \neq \Sigma_1$	Quadratic	Most general model; allows each class to have a unique elliptical shape and orientation, capturing complex separations.

Gaussian Naive Bayes (GNB). GNB assumes diagonal covariance $\Sigma_c = \mathrm{diag}(\sigma^2_{c,1}, \ldots, \sigma^2_{c,D})$, which implies conditionally independent features. The assumption simplifies classification but ignores feature correlations.

Mahalanobis Nearest Centroid Matching (Mah-NCM). Mah-NCM assumes homoscedasticity, i.e., a shared covariance matrix for both classes $\Sigma_0 = \Sigma_1 = \Sigma_p$. The classification decision is based on the Mahalanobis distance $(z - \mu_c)^\top \Sigma_p^{-1} (z - \mu_c)$, which accounts for the correlations of the features by transforming the feature space into one in which the pooled covariance is the identity.

Quadratic Discriminant Analysis (QDA). QDA is the most general of the four classifiers, allowing each class to have its own full-rank covariance matrix ($\Sigma_0 \neq \Sigma_1$), thus assuming heteroskedasticity. The resulting quadratic boundary can adapt to classes with different shapes and orientations.

4.1 Connection to classical generative classification

The efficacy of these Gaussian discriminants in analytical form can be understood through the lens of Bayesian decision theory, which provides a formal basis for our geometric approach. Under the Gaussian assumption, they approximate or realize the Bayes-optimal classifier, which minimizes the probability of error. Full proofs are provided in Appendix B.

Proposition 1 (Bayes-Optimal Classifier under Gaussian Assumption). *If the class-conditional densities are Gaussian, $p(z \mid y = c) \approx \mathcal{N}(\mu_c, \Sigma_c)$, and the class priors are equal, $P(y = 0) = P(y = 1) = \frac{1}{2}$, the Bayes-optimal decision rule is given by Quadratic Discriminant Analysis (QDA).*

Proposition 2 (Optimality under Homoscedasticity). *If the classes are additionally assumed to be homoscedastic (i.e., they share a common covariance matrix $\Sigma_0 = \Sigma_1 = \Sigma_p$), the Bayes-optimal decision rule simplifies to Mahalanobis Nearest Centroid Matching (Mah-NCM).*

Proposition 3 (Performance Stability under Bounded Drift). *Assume a homoscedastic classifier with shared covariance Σ and equal priors. If the statistical drift between the training and testing distributions is bounded by $\|\mu_c^{test} - \mu_c^{train}\|_2 \leq \epsilon_\mu$ and $\|\Sigma^{test} - \Sigma^{train}\|_F \leq \epsilon_\Sigma$, then $|\mathrm{AUC}_{test} - \mathrm{AUC}_{train}| \leq L(\epsilon_\mu + \epsilon_\Sigma)$ where L is the Lipschitz constant.*

Proposition 4 (Stability of the Fisher Margin under Distributional Drift). *Assume that both the training and testing domains are characterized by homoscedastic Gaussian parameters. If the distributional drift is bounded such that $\|\delta^{test} - \delta^{train}\|_2 \leq \epsilon$ and $\|\Sigma_p^{test} - \Sigma_p^{train}\|_F \leq \eta$, where $\delta = \mu_1 - \mu_0$, then the absolute change in the Fisher Margin $\mathcal{D}_\mu$ is bounded, to a first order, by $|\mathcal{D}_\mu^{test} - \mathcal{D}_\mu^{train}| \leq 2\|(\Sigma_p^{train})^{-1}\|_2 \|\delta^{train}\|_2 \epsilon + O(\eta)$.*

These propositions should be read as *model-conditional*: they characterize the optimal decision rules *if* frozen features are well-approximated by Gaussian class-conditionals. They provide the theoretical foundation for our geometric analysis of generalization. Under this model, Quadratic Discriminant Analysis is Bayes-optimal in the heteroscedastic case (Proposition 1) and simplifies to a linear rule under a shared covariance (Mahalanobis-NCM; Proposition 2). Proposition 3 shows that, under the homoscedastic Gaussian model, the AUC of Mah-NCM is a monotone function of the Fisher margin $\mathcal{D}_\mu$. Proposition 4 bounds how $\mathcal{D}_\mu$ changes under drift in low-order statistics. Together, these results imply smooth degradation when class means and covariances drift. In this model, reliable generalization requires geometric alignment between training and testing data.

The linear decision boundary of Mah-NCM has the same functional form as a linear probe trained with cross-entropy. Both induce a single separating hyperplane in z-space. The key difference is optimization rather than representation. A neural network iteratively searches for the boundary, whereas Gaussian discriminants compute it analytically from the data's first- and second-order statistics. Thus, when a trained linear head does not outperform Gaussian discriminants under matched priors and encoders, the result is consistent with the hypothesis that second-order feature geometry is already sufficient for separation. Consistent gaps instead suggest non-Gaussian or higher-order structure that the ladder cannot capture.

Limitations. By design, the Gaussian ladder is restricted to a Gaussian class-conditional approximation based on first- and second-order statistics. The result is a diagnostic baseline rather than a universal detector, one that leaves out higher-order, multimodal, or heavy-tailed cues from strong post-processing, unusual content, or generator-specific artifacts. In those regimes, learned non-linear heads or end-to-end adaptation can legitimately outperform the ladder.

5 Investigating Generalization with Percept-Lens suite

We use the Gaussian ladder to investigate which components of a detection system drive OOD performance under realistic distribution shift. We focus on the training prior, the frozen encoder, and the decision rule on top. To assess the reliability of current AI-generated image detectors under realistic generative variability, we begin by evaluating pre-trained models on the Percept-Lens test suite. These results help to localize whether failures are attributable to the training prior, the representation, or the learned decision surface, motivating more principled system-level investigations.

5.1 Evaluation Protocol and Data

Percept-Lens denotes a unified evaluation *protocol* constructed entirely from existing public datasets to stress-test AIGI detectors under joint distribution shift. Unlike settings where only the generator family changes while the underlying image source and prompt distribution remain fixed, Percept-Lens explicitly couples shifts in generator family, prompt/style, and acquisition pipeline (Figure 3). To avoid conflating in-distribution performance with transfer, the large training priors used for conditioning (e.g., `CommunityForensics`, `GenImage`, `DRCT-2M`, `ELSA-D3`) are treated as support-only and are not part of the evaluation suite.

Evaluation Datasets. The Percept-Lens suite aggregates 39 existing public datasets totaling 7.1 million images. It spans real-only, synthetic-only, and mixed regimes, with joint shifts in (i) generator family, (ii) prompt/style, and (iii) image-source domain and post-processing. Concretely, it combines in-the-wild

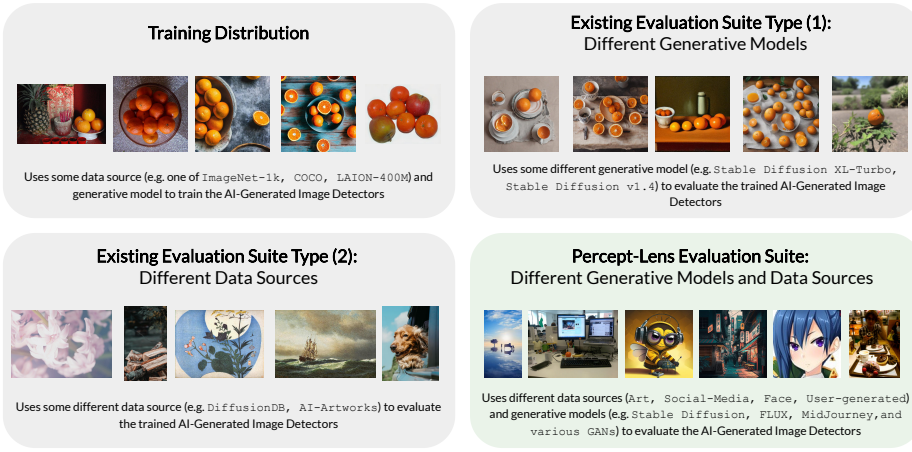

Fig. 3: Percept-Lens stress-tests detectors under joint distribution shift. Percept-Lens is an evaluation *protocol* built entirely from existing public datasets. Many AIGI evaluations vary either the generator family *or* the image source in isolation; we instead evaluate under *joint* shifts in generator family, prompt/style, and source domain, closer to deployment where these factors drift together.

real sources, synthetic-only prompt/style collections, and mixed forensic benchmarks spanning diverse generators and post-processing.

Percept-Lens standardizes evaluation over existing public datasets to ensure head–prior–encoder comparisons are tested under a realistic mixture of shifts. The complete manifest of the dataset and the composition of the class are provided in Appendix D.

Evaluation Protocol. For a training prior, detectors are evaluated on every dataset in the Percept-Lens test suite. We report a dataset-wise class accuracy that equals balanced accuracy when both classes are present and reduces to class-conditional accuracy on one-class sets. Concretely, for an evaluation dataset t with observed classes $\mathcal{C}_t \subseteq \{0, 1\}$, we define Class Accuracy CA as

$$\mathrm{CA}^{(\phi)}(t) \;=\; \frac{1}{|\mathcal{C}_t|} \sum_{c \in \mathcal{C}_t} \Pr(\hat{y}^{\phi} = c \mid y = c), \tag{5.1}$$

and report the macro-average $\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mathrm{CA}(t)$ across evaluation datasets $\mathcal{T}$ so that no single large dataset dominates. For all detectors we use the default $\arg\max$ decision on predicted class probabilities and do not tune thresholds per dataset. On real-only sets, CA corresponds to true-negative accuracy (1–FPR); on synthetic-only sets it corresponds to true-positive accuracy (1–FNR). We additionally compute AUC on the 24 mixed datasets (those containing both real and synthetic images) and report it in Appendix F.

Table 2: Closed-form Gaussian discriminants versus released detector heads under matched (prior, encoder) conditions. Mean class accuracy CA (dataset-wise balanced accuracy where defined; Section 5.1), macro-averaged across Percept-Lens evaluation datasets. For each released detector checkpoint, we report (i) the released decision head (*Out-of-the-Shelf*) and (ii) closed-form baselines fitted on the *same frozen encoder features* using the corresponding public training prior: Euc-NCM [67] and the best-performing rung of the Gaussian ladder. The full ladder is reported in Table 3. (Parentheses indicate the training subset reported by the original work when applicable.)

Detection Model	Out-of-the-Shelf	Euc-NCM [67]	Gaussian ladder (best)
Trained with `CNNSpot` [64] (`ProGAN` Images based on `LSUN`)			
`UnivFD` [48]	54.63%	50.98%	**57.62%** (Mah-NCM) (2.99% ↑)
`AIDE` [69]	56.41%	54.25%	**62.76%** (Mah-NCM) (6.35% ↑)
Trained with `GenImage` [80] (Diffusion Model Images based on `ImageNet-1k`)			
`AIDE` (`GenImage-SDv1`) [69]	48.99%	54.53%	**57.35%** (Mah-NCM) (8.36% ↑)
`Effort` (`GenImage-SDv1`) [70]	72.58%	75.06%	**76.82%** (Mah-NCM) (4.24% ↑)
`DRCT-UnivFD` (Full `GenImage`) [19]	65.77%	66.08%	**73.01%** (Mah-NCM) (7.24% ↑)
`AIDE` (Full `GenImage`) [69]	54.98%	52.24%	**64.64%** (Mah-NCM) (9.66% ↑)
Trained with `DRCT-2M` [19] (Stable Diffusion Model Images based on `COCO`)			
`DRCT-UnivFD` (`DRCT-SDv1`) [19]	63.36%	65.36%	**70.20%** (QDA) (6.84% ↑)
`DRCT-UnivFD` (`DRCT-SDv2`) [19]	62.37%	66.31%	**70.31%** (QDA) (7.94% ↑)
Trained with `ELSA-D3` [16] or `CommunityForensics` [50] (Diffusion Model Images based on `LAION-400M` Dataset)			
`CoDE-kNN` (`ELSA-D3`) [16]	64.01%	63.63%	**66.15%** (Mah-NCM) (2.14% ↑)
`CF-224` (`CommunityForensics`) [50]	81.98%	78.53%	**82.61%** (Mah-NCM) (0.63% ↑)
`CF-384` (`CommunityForensics`) [50]	**87.54%**	82.65%	84.45% (Mah-NCM) (–2.99% ↓)
Frozen image encoder (no detection fine-tuning)			
`PE-Core-bigG-14-448` [12]	—	86.38%	**94.46%** (Mah-NCM)

5.2 Closed-form Gaussian Discriminants versus Trained Heads

Tables 2 and 3 compares released detector heads with Gaussian discriminant rules fitted on the same frozen features and conditioned on the same training prior. Several rows in Tables 2 and 3 share the same underlying backbone. `UnivFD` and `DRCT-UnivFD` both use `ViT-L-14-quickgelu-openai` [49] as the image encoder; differences arise from the public training prior and the learned head. `DRCT` exposes classifier features after reconstruction training. `Effort` applies LoRA adaptation before its head. `AIDE` uses frequency/statistical experts plus semantic features, and `CoDE`/`CF` train backbones and heads end-to-end. All are auditable once a final feature representation is exposed.

Across several priors, the best closed-form Gaussian discriminant improves class accuracy, with the largest gains occurring when separation is well-captured by first- and second-order moments (often Mah-NCM). These patterns are consistent with Propositions 1 and 2: when separation is well-captured by first- and second-order moments in a frozen representation, the corresponding Gaussian discriminants recover most of the available discriminative geometry. A main exception is `CF-384` [50], where the released detector remains ahead of the ladder (87.54% vs. 84.45% CA), consistent with a regime in which the learned decision surface exploits structure not explained by second-order moments alone. Per-dataset results in Appendix F also show large variability. For example, Mah-NCM

Table 3: Comparative Performance of Different Gaussian Discriminants. Mean class accuracy CA (macro-averaged across Percept-Lens evaluation datasets) for each rung of the Gaussian discriminant ladder, conditioned on the same public training prior and fitted on the same frozen features as in Table 2. (Parentheses indicate the training subset reported by the original work; underlining denotes improvement over the corresponding released head as reported in Table 2.)

Detection Model	Cos-NCM	GNB	Mah-NCM	QDA
Trained with CNNSpot [64] (ProGAN Images based on LSUN Dataset)				
UnivFD [48]	51.01%	48.44%	**57.62%**	48.23%
AIDE [69]	54.36%	56.69%	**62.76%**	61.28%
Trained with GenImage [80] (Diffusion Model Images based on ImageNet-1k Dataset)				
AIDE (GenImage-SDv1) [69]	54.23%	54.39%	**57.35%**	57.15%
Effort (GenImage-SDv1) [70]	74.44%	74.92%	**76.82%**	73.21%
DRCT-UnivFD (Full GenImage) [19]	67.97%	68.28%	**73.01%**	69.74%
AIDE (Full GenImage) [69]	52.66%	53.92%	**64.64%**	56.20%
Trained with DRCT-2M [19] (Stable Diffusion Model Images based on COCO Dataset)				
DRCT-UnivFD (DRCT-SDv1) [19]	65.16%	66.37%	65.21%	**70.20%**
DRCT-UnivFD (DRCT-SDv2) [19]	65.53%	66.69%	67.91%	**70.31%**
Trained with ELSA-D3 [16] or CommunityForensics [50] (Diffusion Model Images based on LAION-400M Dataset)				
CoDE-kNN (ELSA-D3) [16]	62.70%	65.98%	**66.15%**	64.00%
CF-224 (CommunityForensics) [50]	78.98%	79.45%	**82.61%**	80.70%
CF-384 (CommunityForensics) [50]	84.46%	80.01%	84.55%	79.52%
Frozen image encoder (no detection fine-tuning)				
PE-Core-bigG-14-448 [12]	86.06%	87.26%	**94.46%**	86.15%

reaches near-ceiling performance on many synthetic-only sets but is much weaker on DeepFakeBench [71], while QDA performs best on CelebA-Spoof [75].

5.3 Encoder Sensitivity of Closed-form Gaussian Discriminants

We investigate the model-agnostic behavior of the ladder by evaluating a range of frozen architectures and pretraining objectives (Tables 4 and 5). Mah-NCM is consistently the top performer across these encoders, indicating that a shared full-covariance geometry captures a stable component of real/fake separability in modern representation spaces. This motivates a practical reporting guideline. Learned heads on frozen encoders should be compared against the best Gaussian rung on the *same* frozen features. Additional models are presented in Appendix E.

5.4 Training Prior Sensitivity of Closed-form Gaussian Discriminants

We isolate training-prior effects by fixing the encoder (PE-Core-bigG-14-448 [12]) and varying only the support prior used to estimate Gaussian parameters (Table 6). The spread is large: the best CA ranges from 77.92% (CNNSpot) to 94.46%

Table 4: Encoder scaling under a fixed prior: representation choice strongly affects transfer. Mean class accuracy CA (macro-averaged across Percept-Lens evaluation datasets) for Gaussian discriminant heads conditioned on the same training prior (`CommunityForensics` [50]), using frozen CLIP encoders of increasing capacity pretrained by OpenAI. Only the head is changed. No detector fine-tuning is performed.

Frozen Encoders	Euc-NCM [67]	Cos-NCM	GNB	Mah-NCM	QDA
`ResNet-50` [60]	65.27%	64.36%	65.72%	**72.00%**	63.15%
`ResNet-101` [59]	65.20%	64.70%	64.46%	**70.27%**	62.29%
`ResNet-50x4` [62]	66.24%	65.91%	65.12%	**72.34%**	64.28%
`ResNet-50x16` [61]	67.35%	66.77%	66.21%	**74.98%**	65.71%
`ResNet-50x64` [63]	68.42%	68.19%	65.02%	**74.57%**	68.38%
`ViT-B/16` [39]	64.66%	64.65%	66.37%	**74.18%**	65.70%
`ViT-B/32` [40]	64.53%	64.25%	64.83%	**70.61%**	62.81%
`ViT-L/14` [49]	66.18%	68.44%	69.79%	**76.96%**	68.87%

Table 5: Pretraining objective affects transfer even without detector training. Mean class accuracy CA (macro-averaged across Percept-Lens evaluation datasets) for Gaussian discriminant heads conditioned on the same training prior (`CommunityForensics` [50]), using frozen encoders trained with different objectives.

Frozen Encoders	Euc-NCM [67]	Cos-NCM	GNB	Mah-NCM	QDA
`MAE-Huge` [8]	55.18%	55.22%	58.33%	**69.37%**	64.01%
`BEiT-Large` [5]	58.53%	58.49%	58.80%	**61.78%**	53.63%
`CLIP-XLM-RoBERTa-Large` [4]	70.72%	69.70%	70.61%	**77.67%**	65.79%
`SigLIP-Large` [9]	58.25%	58.25%	58.07%	**64.14%**	60.85%
`BLIP-Large` [10]	55.75%	56.89%	53.87%	**70.17%**	55.42%
`BLIP2` [11]	68.61%	68.46%	65.41%	**77.73%**	62.21%
`DINOv2-giant` [6]	63.15%	63.59%	61.95%	**73.30%**	58.52%
`DINOv3-ViT-7b` [7]	83.85%	85.82%	83.18%	**88.91%**	78.88%

Table 6: Training prior dominates transfer even under a fixed encoder. Mean class accuracy CA (macro-averaged across Percept-Lens evaluation datasets) for Gaussian discriminant heads when varying only the *support prior* used to estimate moments, under a fixed frozen encoder (`PE-Core-bigG-14-448` [12]) (Underlining denotes improvement over the best released head of `CF-384` [50] as reported in Table 2).

Training Dataset	Euc-NCM [67]	Cos-NCM	GNB	Mah-NCM	QDA
`CNNSpot` [64]	74.95%	78.57%	59.21%	**77.93%**	49.24%
`GenImage` [80]	86.68%	88.75%	87.50%	**92.43%**	81.83%
`GenImage-SDv1`	88.74%	89.80%	86.34%	**92.57%**	65.06%
`DRCT-2M` [19]	81.90%	81.76%	86.21%	83.25%	**91.09%**
`DRCT-SDv1` [19]	82.29%	83.81%	88.70%	88.31%	**89.92%**
`DRCT-SDv2` [19]	83.50%	84.38%	89.56%	90.08%	**91.44%**
`ELSA-D3` [16]	88.87%	89.03%	86.98%	**94.45%**	86.40%
`CommunityForensics` [50]	86.38%	86.06%	87.26%	**94.46%**	86.15%

(`CommunityForensics`), despite using the *same* representation and evaluation suite.

Table 7: Data-efficiency of moment-based heads. Mean class accuracy CA (macro-averaged across Percept-Lens evaluation datasets) for Gaussian discriminant heads as the support set used to estimate moments is subsampled from `CommunityForensics` [50], under a fixed frozen encoder (`PE-Core-bigG-14-448` [12]). (Superscript/subscript denote max/min deviations (in percentage points) from the reported mean).

Amount	Euc-NCM [67]	Cos-NCM	GNB	Mah-NCM	QDA
0.001%	$83.97\%^{+1.93}_{-2.74}$	$83.46\%^{+2.25}_{-2.47}$	$\mathbf{87.14\%}^{+1.21}_{-1.68}$	$85.56\%^{+1.84}_{-1.75}$	$85.13\%^{+1.07}_{-0.66}$
0.005%	$86.04\%^{+0.28}_{-0.40}$	$85.78\%^{+0.27}_{-0.53}$	$87.23\%^{+0.76}_{-0.77}$	$\mathbf{90.89\%}^{+0.73}_{-1.34}$	$85.53\%^{+1.51}_{-1.71}$
0.01%	$86.32\%^{+0.67}_{-0.35}$	$86.02\%^{+0.77}_{-0.42}$	$87.42\%^{+1.19}_{-0.90}$	$\mathbf{91.55\%}^{+0.67}_{-0.69}$	$86.48\%^{+1.16}_{-0.82}$
0.1%	$86.26\%^{+0.71}_{-0.44}$	$85.92\%^{+0.74}_{-0.47}$	$87.33\%^{+0.22}_{-0.19}$	$\mathbf{93.18\%}^{+0.42}_{-0.54}$	$84.18\%^{+0.33}_{-0.71}$
1%	$86.39\%^{+0.11}_{-0.12}$	$86.05\%^{+0.11}_{-0.10}$	$87.28\%^{+0.04}_{-0.05}$	$\mathbf{93.52\%}^{+0.28}_{-0.17}$	$85.16\%^{+0.20}_{-0.34}$
100%	86.38%	86.06%	87.26%	**94.46%**	86.15%

Moreover, the prevailing covariance assumption shifts with the prior: `DRCT-2M` favors QDA (heteroscedastic) while all others favor Mah-NCM (homoscedastic). Thus, if two papers use the same backbone but different priors, their OOD behavior can differ more than what is attributable to the classifier family within the Gaussian ladder.

5.5 Data Efficiency of Closed-form Gaussian Discriminants

We investigate data efficiency by subsampling the support set used to estimate Gaussian parameters (Table 7). Performance saturates quickly. With only 1% of the prior (around 43k samples), Mah-NCM nearly matches the full-support estimate, while at extremely small support sizes (0.001% of the prior, or 44 samples) the diagonal GNB assumption can be more stable. The pattern reinforces the central observation from the ladder. For this support prior and encoder, low-order feature statistics capture a large share of the signal used by the best head, without iterative optimization.

With a strong frozen encoder (`PE-Core-bigG-14-448` [12]), Mah-NCM fitted from only 0.005% of the `CommunityForensics` prior (219 labeled samples) reaches 90.89% CA. Although this is not an apples-to-apples comparison to `CF-384` (which uses a different backbone), it indicates that moment-based adaptation can be data-efficient once the representation is strong.

5.6 Representation Sensitivity of Wasserstein-2 Shift Estimates

All geometric shift metrics in Percept-Lens are computed in a learned feature space, hence their numerical values are not purely properties of the data distributions, but of the *representation* used to embed them. This raises a basic reliability question about whether conclusions drawn from a Gaussian shift metric are stable under reasonable encoder changes, or are artifacts of a particular backbone.

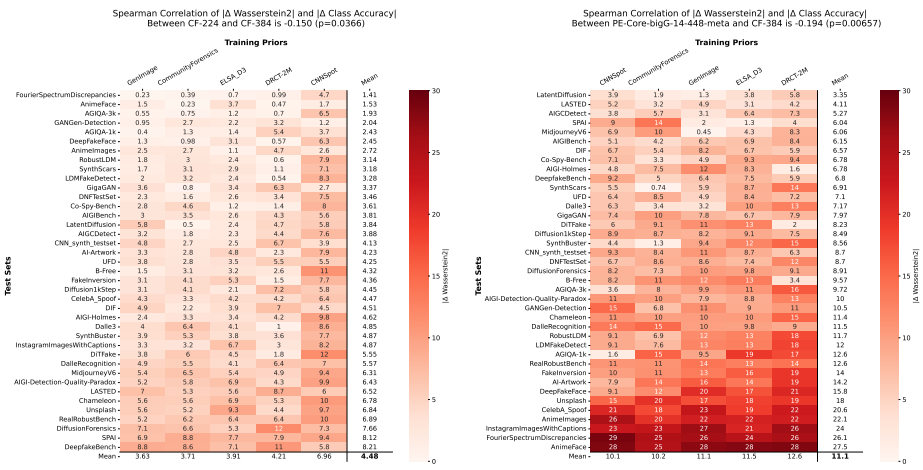

Fig. 4: Gaussian Wasserstein-2 distances are representation-conditioned. Each cell reports the absolute difference in class-averaged Gaussian $\mathcal{W}_2$ for the same train–test pair under two feature spaces, $\Delta\mathcal{W}_2(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}; \phi_a, \phi_b)$ (row/column means on the margins). The sub-figure title reports the Spearman correlation between the flattened cell-wise $\Delta\mathcal{W}_2$ values and the corresponding ΔCA values. Large changes in $\mathcal{W}_2$ do not necessarily coincide with large changes in Mah-NCM CA, so feature-space shift magnitudes should not be interpreted as detector-agnostic indicators without checking stability across representations.

Setup. For each training prior $\mathcal{D}_{\text{train}}$ and test set $\mathcal{D}_{\text{test}}$, class-conditional Gaussians are fitted to the representation induced by an encoder ϕ, then a class-averaged Wasserstein-2 shift is computed between the training and test distributions:

$$\mathcal{W}_2^{(\phi)}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) \;=\; \frac{1}{2} \sum_{c \in \{0,1\}} \mathrm{W}_2\Big(\mathcal{N}(\mu_{c,\mathcal{D}_{\text{train}}}^{(\phi)}, \Sigma_{c,\mathcal{D}_{\text{train}}}^{(\phi)}), \mathcal{N}(\mu_{c,\mathcal{D}_{\text{test}}}^{(\phi)}, \Sigma_{c,\mathcal{D}_{\text{test}}}^{(\phi)})\Big)\,. \tag{5.2}$$

To isolate representation effects while keeping the data fixed, Figure 4 uses heatmap cells for the encoder-induced change in Gaussian $\mathcal{W}_2$ and panel titles for its Spearman correlation with the corresponding change in Mah-NCM accuracy. Thus each panel asks whether the train–test pairs whose estimated shift changes most after an encoder swap are also the pairs whose detector accuracy changes most. High correlation means the shift estimate is decision-aligned for that encoder pair. Weak correlation means $\mathcal{W}_2$ is responding to feature-space changes that the detector does not use:

$$\Delta\mathcal{W}_2(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}; \phi_a, \phi_b) = \Big|\mathcal{W}_2^{(\phi_a)}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}}) - \mathcal{W}_2^{(\phi_b)}(\mathcal{D}_{\text{train}}, \mathcal{D}_{\text{test}})\Big|\,, \tag{5.3}$$

$$\Delta\mathrm{CA}(\mathcal{D}_{\text{test}}; \phi_a, \phi_b) \;=\; \Big|\mathrm{CA}^{(\phi_a)}(\mathcal{D}_{\text{test}}) - \mathrm{CA}^{(\phi_b)}(\mathcal{D}_{\text{test}})\Big|\,. \tag{5.4}$$

Figure 4 shows strong representation dependence. Within a closely related encoder family (`CF-224` [50] vs. `CF-384` [50]), $\Delta\mathcal{W}_2$ is modest on average. Across larger backbone changes (e.g., `PE-Core-bigG-14-448` [12] vs. `CF-384` [50]), $\Delta\mathcal{W}_2$ increases by an order of magnitude, especially for real-only datasets. The sensitivity is also uneven. Particular prior–test pairs exhibit consistently larger $\Delta\mathcal{W}_2$, indicating that some distribution shifts are more representation-fragile.

A weak association between $\Delta\mathcal{W}_2$ and ΔCA implies that global distributional distances can change under encoder re-parameterizations that do not materially affect the *discriminative* geometry used by Mah-NCM. For example, anisotropic rescaling can inflate Wasserstein distances while leaving the effective decision geometry largely intact. Percept-Lens therefore measures generalization at the level of complete detection systems (training prior $\times$ encoder $\times$ classifier), and the representation axis is not interchangeable. Any claim that uses $\mathcal{W}_2$ to argue for or against a training prior must therefore be stated as *representation-conditioned*. Robust conclusions should rely on patterns that persist across encoders (e.g., rank-consistent prior ordering or decision-aligned shift measures), rather than absolute $\mathcal{W}_2$ values in a single feature space.

Practical reporting checklist. These experiments suggest a minimal audit for future AIGI detector papers. Authors should report the released head and the best Gaussian rung on the same frozen features and support prior, include per-dataset CA/AUC rather than only macro averages, state the encoder used for any geometric shift metric, and flag non-Gaussian feature regimes using diagnostics such as those in Appendix G. This makes head-level OOD claims auditable without treating the Gaussian ladder as a replacement for specialized detectors.

6 Conclusion

This paper studies prior-conditioned Gaussian discriminants as practical baselines and diagnostics for AI-generated image detection under joint distribution shift. Using a unified public-data protocol, the ladder is often competitive with trained detector heads under matched priors and encoders, and sometimes exceeds them. We support a matched head audit plus controlled prior/encoder sweeps, not a fully symmetric decomposition of every detector component. The source-side recommendation is to compare the trained head with the best Gaussian rung on the same representation and support prior before claiming head-level OOD gains. Across priors, the winning covariance model indicates which low-order statistics transfer, and the data-efficiency results show that useful moment estimates can be obtained from small support sets. Future work should characterize failures in multimodal or heavy-tailed feature regimes and design adaptation procedures that improve OOD performance while preserving transferable geometry.

References

1. Instagram images with captions, https://www.kaggle.com/datasets/prithvijaunjale/instagram-images-with-captions
2. terminusresearch/midjourney-v6-520k-raw · datasets at hugging face, https://huggingface.co/datasets/terminusresearch/midjourney-v6-520k-raw
3. Unsplash dataset, https://unsplash.com/data
4. laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k · hugging face (Oct 2019), https://huggingface.co/laion/CLIP-ViT-H-14-frozen-xlm-roberta-large-laion5B-s13B-b90k
5. microsoft/beit-large-patch16-224-pt22k-ft22k · hugging face (Jun 2021), https://huggingface.co/microsoft/beit-large-patch16-224-pt22k-ft22k
6. facebook/dinov2-giant · hugging face (Aug 2025), https://huggingface.co/facebook/dinov2-giant
7. facebook/dinov3-vit7b16-pretrain-lvd1689m · hugging face (Aug 2025), https://huggingface.co/facebook/dinov3-vit7b16-pretrain-lvd1689m
8. facebook/vit-mae-huge · hugging face (Sep 2025), https://huggingface.co/facebook/vit-mae-huge
9. google/siglip-large-patch16-384 · hugging face (Jul 2025), https://huggingface.co/google/siglip-large-patch16-384
10. Salesforce/blip-image-captioning-large · hugging face (Oct 2025), https://huggingface.co/Salesforce/blip-image-captioning-large
11. Salesforce/blip2-opt-6.7b · hugging face (Oct 2025), https://huggingface.co/Salesforce/blip2-opt-6.7b
12. timm/PE-core-bigG-14-448 · hugging face (Sep 2025), https://huggingface.co/timm/PE-Core-bigG-14-448
13. Bammey, Q.: Synthbuster: Towards detection of diffusion model generated images. IEEE Open Journal of Signal Processing **5**, 1–9 (2024). https://doi.org/10.1109/OJSP.2023.3337714, https://ieeexplore.ieee.org/document/10334046
14. Banana_Leopard: Anime images dataset (2023), https://www.kaggle.com/datasets/diraizel/anime-images-dataset
15. Bansal, P., Kavis, A., Sanghavi, S.: Understanding contrastive learning via gaussian mixture models (Oct 2025), https://openreview.net/forum?id=nCAdkkAeR9
16. Baraldi, L., Cocchi, F., Cornia, M., Baraldi, L., Nicolosi, A., Cucchiara, R.: Contrasting deepfakes diffusion via contrastive learning and global-local similarities. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 199–216. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73036-8_12, https://openreview.net/forum?id=tsrYdgxbFM
17. Bushard, B.: Fake image of explosion near pentagon went viral—even though it never happened. Forbes (2023), https://www.forbes.com/sites/brianbushard/2023/05/22/fake-image-of-explosion-near-pentagon-went-viral-even-though-it-never-happened/
18. Cazenavette, G., Sud, A., Leung, T., Usman, B.: FakeInversion: Learning to Detect Images from Unseen Text-to-Image Models by Inverting Stable Diffusion. pp. 10759–10769 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Cazenavette_FakeInversion_Learning_to_Detect_Images_from_Unseen_Text-to-Image_Models_by_CVPR_2024_paper.html
19. Chen, B., Zeng, J., Yang, J., Yang, R.: DRCT: diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Forty-first

International Conference on Machine Learning (Jun 2024), https://openreview.net/forum?id=oRLwyayrh1
20. Cheng, S., Lyu, L., Wang, Z., Zhang, X., Sehwag, V.: CO-SPY: Combining semantic and pixel features to detect synthetic images by AI. pp. 13455–13465 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Cheng_CO-SPY_Combining_Semantic_and_Pixel_Features_to_Detect_Synthetic_Images_CVPR_2025_paper.html
21. Chuangchuang Tan, Renshuai Tao, H.L.: GANGen-detection: a dataset generated by gans for generalizable deepfake detection (2024), https://github.com/chuangchuangtan/GANGen-Detection
22. Churchill, S., Chao, B.: Anime face dataset (2019). https://doi.org/10.34740/KAGGLE/DS/379764, https://www.kaggle.com/ds/379764
23. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (Jun 2023). https://doi.org/10.1109/ICASSP49357.2023.10095167, https://ieeexplore.ieee.org/abstract/document/10095167
24. Daniel, L.: How hurricane helene deepfakes flooding social media hurt real people. Forbes (2024), https://www.forbes.com/sites/larsdaniel/2024/10/04/hurricane-helena-deepfakes-flooding-social-media-hurt-real-people/
25. Dzanic, T., Shah, K., Witherden, F.: Fourier spectrum discrepancies in deep network generated images. In: Advances in Neural Information Processing Systems. vol. 33, pp. 3022–3032. Curran Associates, Inc. (2020), https://proceedings.neurips.cc/paper/2020/hash/1f8d87e1161af68b81bace188a1ec624-Abstract.html
26. Egan, B., Redden, A., XWAVE, SilentAntagonist: Dalle3 1 million+ high quality captions (May 2024), https://huggingface.co/datasets/ProGamerGov/synthetic-dataset-1m-dalle3-high-quality-captions
27. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging Frequency Analysis for Deep Fake Image Recognition. In: Proceedings of the 37th International Conference on Machine Learning. pp. 3247–3258. PMLR (Nov 2020), https://proceedings.mlr.press/v119/frank20a.html
28. Guillaro, F., Zingarini, G., Usman, B., Sud, A., Cozzolino, D., Verdoliva, L.: A bias-free training paradigm for more general AI-generated image detection. pp. 18685–18694 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Guillaro_A_Bias-Free_Training_Paradigm_for_More_General_AI-generated_Image_Detection_CVPR_2025_paper.html
29. He, Y., Gan, B., Chen, S., Zhou, Y., Yin, G., Song, L., Sheng, L., Shao, J., Liu, Z.: ForgeryNet: a versatile benchmark for comprehensive forgery analysis. pp. 4360–4369 (2021), https://openaccess.thecvf.com/content/CVPR2021/html/He_ForgeryNet_A_Versatile_Benchmark_for_Comprehensive_Forgery_Analysis_CVPR_2021_paper.html
30. Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L.: OpenCLIP (Jul 2021). https://doi.org/10.5281/zenodo.5143773, https://doi.org/10.5281/zenodo.5143773
31. Kang, H., Wen, S., Wen, Z., Ye, J., Li, W., Feng, P., Zhou, B., Wang, B., Lin, D., Zhang, L., He, C.: LEGION: Learning to ground and explain for synthetic image detection. pp. 18937–18947 (2025), https://openaccess.thecvf.com/content/ICCV2025/html/Kang_LEGION_Learning_to_Ground_and_Explain_for_Synthetic_Image_Detection_ICCV_2025_paper.html

32. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up GANs for text-to-image synthesis. pp. 10124–10134 (2023), https://openaccess.thecvf.com/content/CVPR2023/html/Kang_Scaling_Up_GANs_for_Text-to-Image_Synthesis_CVPR_2023_paper.html
33. Karageorgiou, D., Papadopoulos, S., Kompatsiaris, I., Gavves, E.: Any-Resolution AI-Generated Image Detection by Spectral Learning. pp. 18706–18717 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Karageorgiou_Any-Resolution_AI-Generated_Image_Detection_by_Spectral_Learning_CVPR_2025_paper.html
34. Kholy, A.E.: AI-artwork (2024). https://doi.org/10.34740/KAGGLE/DSV/7878124, https://www.kaggle.com/dsv/7878124
35. Koliha, N.: AI recognition dataset (2024). https://doi.org/10.34740/KAGGLE/DSV/7501727, https://www.kaggle.com/dsv/7501727
36. Laion: CLIP-convnext_base_w-laion2B-s13B-b82K-augreg (2024), https://huggingface.co/laion/CLIP-convnext_base_w-laion2B-s13B-b82K-augreg
37. Laion: CLIP-convnext_large_d_320.laion2B-s29B-b131K-ft-soup (2024), https://huggingface.co/laion/CLIP-convnext_large_d_320.laion2B-s29B-b131K-ft-soup
38. Laion: CLIP-convnext_xxlarge-laion2B-s34B-b82K-augreg-soup (2024), https://huggingface.co/laion/CLIP-convnext_xxlarge-laion2B-s34B-b82K-augreg-soup
39. Laion: CLIP-ViT-B-16-datacomp.xl-s13B-b90K (2024), https://huggingface.co/laion/CLIP-ViT-B-16-DataComp.XL-s13B-b90K
40. Laion: CLIP-ViT-B-32-256x256-DataComp-s34B-b86K (2024), https://huggingface.co/laion/CLIP-ViT-B-32-256x256-DataComp-s34B-b86K
41. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018), https://papers.nips.cc/paper_files/paper/2018/hash/abdeb6f575ac5c6676b747bca8d09cc2-Abstract.html
42. Li, C., Wang, X., Li, M., Miao, B., Sun, P., Zhang, Y., Ji, X., Zhu, Y.: Bridging the gap between ideal and real-world evaluation: Benchmarking AI-generated image detection in challenging scenarios. pp. 20379–20389 (2025), https://openaccess.thecvf.com/content/ICCV2025/html/Li_Bridging_the_Gap_Between_Ideal_and_Real-world_Evaluation_Benchmarking_AI-Generated_ICCV_2025_paper.html
43. Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai, G., Lin, W.: AGIQA-3K: an open database for AI-generated image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology **34**(8), 6833–6846 (Aug 2024). https://doi.org/10.1109/TCSVT.2023.3319020, https://ieeexplore.ieee.org/abstract/document/10262331?casa_token=d0CSMRfrEDcAAAAA:hMig5SBv7x_oX--KIQDLr_fs5XcGZoKZr7VqsfMjujTKFwMNf32x5WilbDCrBp5l2Ht-REfDSw
44. Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Feng, F.: Improving synthetic image detection towards generalization: an image transformation perspective. In: Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1. pp. 2405–2414. KDD '25, Association for Computing Machinery, New York, NY, USA (Jul 2025). https://doi.org/10.1145/3690624.3709392, https://dl.acm.org/doi/10.1145/3690624.3709392
45. Li, Z., Yan, J., He, Z., Zeng, K., Jiang, W., Xiong, L., Fu, Z.: Is artificial intelligence generated image detection a solved problem? (Oct 2025), https://openreview.net/forum?id=N52U2h9k9o

46. McCarthy, B.: Fake pentagon explosion image spreads online. AFP Fact Check (2023), https://factcheck.afp.com/doc.afp.com.33FV4BU
47. MidJourney: MidJourney: AI-based text-to-image generation (2022), https://www.midjourney.com/
48. Ojha, U., Li, Y., Lee, Y.J.: 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), https://openaccess.thecvf.com/content/CVPR2023/html/Ojha_Towards_Universal_Fake_Image_Detectors_That_Generalize_Across_Generative_Models_CVPR_2023_paper.html
49. OpenAI: CLIP-ViT-large-patch14 (2024), https://huggingface.co/openai/clip-vit-large-patch14
50. Park, J., Owens, A.: Community forensics: using thousands of generators to train fake image detectors. pp. 8245–8257 (2025), https://openaccess.thecvf.com/content/CVPR2025/html/Park_Community_Forensics_Using_Thousands_of_Generators_to_Train_Fake_Image_CVPR_2025_paper.html
51. Rajan, A.S., Lee, Y.J.: Stay-positive: a case for ignoring real image features in fake image detection (Jun 2025), https://openreview.net/forum?id=VNLmfMJi3w&noteId=x7FR8R6HMU
52. Rajan, A.S., Ojha, U., Schloesser, J., Lee, Y.J.: Aligned Datasets Improve Detection of Latent Diffusion-Generated Images (Oct 2024), https://openreview.net/forum?id=doBkiqESYq
53. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: Proceedings of the 38th International Conference on Machine Learning. pp. 8821–8831. PMLR (Jul 2021), https://proceedings.mlr.press/v139/ramesh21a.html
54. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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), https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper
55. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Advances in Neural Information Processing Systems **35**, 36479–36494 (Dec 2022), https://proceedings.neurips.cc/paper_files/paper/2022/hash/ec795aeadae0b7d230fa35cbaf04c041-Abstract-Conference.html
56. Sinitsa, S., Fried, O.: Deep image fingerprint: towards low budget synthetic image detection and model lineage analysis. pp. 4067–4076 (2024), https://openaccess.thecvf.com/content/WACV2024/html/Sinitsa_Deep_Image_Fingerprint_Towards_Low_Budget_Synthetic_Image_Detection_and_WACV_2024_paper.html
57. Song, H., Huang, S., Dong, Y., Tu, W.W.: Robustness and generalizability of deepfake detection: a study with diffusion models (Sep 2023). https://doi.org/10.48550/arXiv.2309.02218, http://arxiv.org/abs/2309.02218
58. Tan, C., Zhao, Y., Wei, S., Gu, G., Liu, P., Wei, Y.: Rethinking the up-sampling operations in CNN-based generative network for generalizable deepfake detection. pp. 28130–28139 (2024), https://openaccess.thecvf.com/content/CVPR2024/html/Tan_Rethinking_the_Up-Sampling_Operations_in_CNN-based_Generative_Network_for_Generalizable_CVPR_2024_paper.html

59. Timm: Resnet101_clip.openai (2024), https://huggingface.co/timm/resnet101_clip.openai
60. Timm: Resnet50_clip.openai (2024), https://huggingface.co/timm/resnet50_clip.openai
61. Timm: Resnet50x16_clip.openai (2024), https://huggingface.co/timm/resnet50x16_clip.openai
62. Timm: Resnet50x4_clip.openai (2024), https://huggingface.co/timm/resnet50x4_clip.openai
63. Timm: Resnet50x64_clip.openai (2024), https://huggingface.co/timm/resnet50x64_clip.openai
64. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. pp. 8695–8704 (2020), https://openaccess.thecvf.com/content_CVPR_2020/html/Wang_CNN-Generated_Images_Are_Surprisingly_Easy_to_Spot..._for_Now_CVPR_2020_paper.html
65. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: DIRE for Diffusion-Generated Image Detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22388–22398. arXiv (Oct 2023). https://doi.org/10.1109/ICCV51070.2023.02051, https://ieeexplore.ieee.org/document/10377654/
66. Wu, H., Zhou, J., Zhang, S.: Generalizable synthetic image detection via language-guided contrastive learning. IEEE Transactions on Artificial Intelligence pp. 1–11 (2025). https://doi.org/10.1109/TAI.2025.3641104, https://ieeexplore.ieee.org/abstract/document/11281880
67. Wu, S., Liu, J., Li, J., Wang, Y.: Few-Shot Learner Generalizes Across AI-Generated Image Detection (Jun 2025), https://openreview.net/forum?id=uvU29AfoNT&noteId=76cql2x3jK
68. Xiao, Y., Yang, B., Chen, W., Chen, J., Cao, Z., Dong, Z., Ji, X., Lin, L., Ke, W., Wei, P.: Are high-quality AI-generated images more difficult for models to detect? (Jun 2025), https://openreview.net/forum?id=sKYdVKE1tS&noteId=1BtdJodSwK
69. Yan, S., Li, O., Cai, J., Hao, Y., Jiang, X., Hu, Y., Xie, W.: A sanity check for AI-generated image detection (Oct 2024), https://openreview.net/forum?id=ODRHZrkOQM
70. Yan, Z., Wang, J., Jin, P., Zhang, K.Y., Liu, C., Chen, S., Yao, T., Ding, S., Wu, B., Yuan, L.: Orthogonal subspace decomposition for generalizable AI-generated image detection (Jun 2025), https://openreview.net/forum?id=GFpjO8S8Po
71. Yan, Z., Zhang, Y., Yuan, X., Lyu, S., Wu, B.: DeepfakeBench: A comprehensive benchmark of deepfake detection. Advances in Neural Information Processing Systems **36**, 4534–4565 (Dec 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/0e735e4b4f07de483cbe250130992726-Abstract-Datasets_and_Benchmarks.html
72. Yang, M., Goenawan, G.J., Wang, H., Qin, H., Xu, C., Yang, Y., Fang, F., Sun, Y., Lim, J.H., Zhu, H.: Your AI-Generated Image Detector Can Secretly Achieve SOTA Accuracy, If Calibrated. Proceedings of the AAAI Conference on Artificial Intelligence **40**(14), 11622–11630 (Mar 2026). https://doi.org/10.1609/aaai.v40i14.38146, https://ojs.aaai.org/index.php/AAAI/article/view/38146
73. Yang, X., Li, Y., Lyu, S.: Exposing Deep Fakes Using Inconsistent Head Poses. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8261–8265 (May 2019). https://doi.org/10.1109/ICASSP.2019.8683164, https://ieeexplore.ieee.org/abstract/document/8683164?casa_token=J-Gs5Rxe4HEAAAAA:

67CcBzvyD6o9fsp18ns9tw6bsA4VIADNo5cnBcwvxYl3vHgtPqWCs8k8pWb0eWNRQR0 - HWAg2g

74. Zhang, Y., Xu, X.: Diffusion noise feature: Accurate and fast generated image detection. In: Ecai 2025, pp. 1139–1146. IOS Press (2025). https://doi.org/10.3233/FAIA250925, https://ebooks.iospress.nl/doi/10.3233/FAIA250925
75. Zhang, Y., Yin, Z., Li, Y., Yin, G., Yan, J., Shao, J., Liu, Z.: CelebA-spoof: Large-scale face anti-spoofing dataset with rich annotations. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 70–85. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-58610-2_5
76. Zhang, Z., Li, C., Sun, W., Liu, X., Min, X., Zhai, G.: A perceptual quality assessment exploration for AIGC images. In: 2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW). pp. 440–445 (Jul 2023). https://doi.org/10.1109/ICMEW59549.2023.00082, https://ieeexplore.ieee.org/document/10222021
77. Zhong, N., Xu, Y., Li, S., Qian, Z., Zhang, X.: PatchCraft: Exploring texture patch for efficient AI-generated image detection (Mar 2024). https://doi.org/10.48550/arXiv.2311.12397, http://arxiv.org/abs/2311.12397
78. Zhou, Y., He, X., Lin, K., Fan, B., Ding, F., Zeng, J., Li, B.: Brought a Gun to a Knife Fight: Modern VFM Baselines Outgun Specialized Detectors on In-the-Wild AI Image Detection (Oct 2025). https://doi.org/10.48550/arXiv.2509.12995, http://arxiv.org/abs/2509.12995
79. Zhou, Z., Luo, Y., Wu, Y., Sun, K., Ji, J., Yan, K., Ding, S., Sun, X., Wu, Y., Ji, R.: AIGI-holmes: Towards explainable and generalizable AI-generated image detection via multimodal large language models. pp. 18746–18758 (2025), https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_AIGI-Holmes_Towards_Explainable_and_Generalizable_AI-Generated_Image_Detection_via_Multimodal_ICCV_2025_paper.html
80. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. Advances in Neural Information Processing Systems **36**, 77771–77782 (Dec 2023), https://proceedings.neurips.cc/paper_files/paper/2023/hash/f4d4a021f9051a6c18183b059117e8b5-Abstract-Datasets_and_Benchmarks.html