STANDARD EVALUATION METRICS FOR GENERATIVE IMAGE MODELS OVERLOOK HALLUCINATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper reveals an important problem with the standard evaluation metrics used for image generation: they cannot distinguish between two kinds of flawed samples, *poor inliers* and *hallucinations*. Poor inliers preserve the semantic properties that define the target distribution but are rendered imperfectly, with defects such as blur or warped edges. Hallucinations violate those properties, while often remaining visually plausible. The distinction is semantic rather than visual, and it significantly affects how the model should be improved. We show that the scores of the standard metrics are consistent with many combinations of the two failures, so a model that hallucinates more but produces better inliers can score the same as, or even higher than, one that hallucinates less. Having identified the problem, we analyse what determines how strongly these metrics respond to hallucinations relative to inlier quality. Experiments in controlled settings and on samples from trained generative models show that hallucinations largely go unnoticed, because the feature extractors the metrics rely on often embed them close to inliers. Even when hallucinations are embedded far from inliers, averaging over all samples hides them, while a statistic that reports only the most atypical samples recovers them and ranks models differently from the standard metrics.

1 INTRODUCTION

Despite their success, generative models still produce samples that do not belong to the target data distribution. Originally coined for large language models (LLMs), the term *hallucination* describes instances where models generate confident outputs that diverge from input facts or reality (Zhang et al., 2025). Many works now evaluate how much of it they produce (Janiak et al., 2025) and how to reduce it (Tonmoy et al., 2024). The term has since transferred to vision-language models (VLMs), with the same concerns and objectives (Liu et al., 2024). In these language models the failure is semantic: the output is fluent, confidently asserted and false against a reference standard, such as the input documents in summarisation or world knowledge in open-ended generation (Maynez et al., 2020; Li et al., 2023).

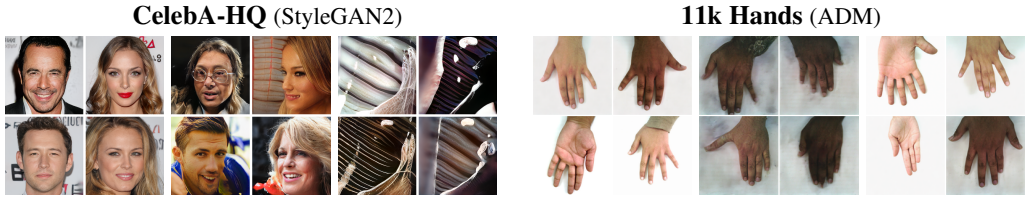

Figure 1: **Poor inliers and hallucinations.** Rendering quality and semantic correctness are independent. A poor inlier satisfies the semantic properties of the target distribution but is rendered badly. A hallucination violates a semantic property of the target and can be either a degenerate sample with no content (hallucinated, left) or a flawlessly rendered one (hallucinated, right). Standard metrics conflate the two failures and can even reward hallucinations.

Unconditional generative image models have a standard of their own: the semantics of the target distribution they aim to learn. A sample can fail in two ways, only one of which violates this

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.

standard. It can deviate from these semantics, or it can respect them but be rendered badly. These failures need different remedies, yet current practice scores both with the same quantity and reports them as a single number. We separate them (Figure 1). A *poor inlier* keeps the properties that define membership in the target distribution and is only rendered imperfectly, for instance with blur or distorted edges. A *hallucination* mixes attributes the target distribution keeps apart, drops one it requires, or adds one it does not contain, and may be rendered perfectly while doing so. The remedies are not interchangeable, and the ones that suppress the semantic failure are paid for in diversity, whether by truncation (Brock et al., 2018), guidance (Ho & Salimans, 2022), or outright rejection (Aithal et al., 2024), while poor rendering is what better training and sampling design improve at no such cost (Karras et al., 2022). Acting on either failure therefore requires knowing which one a model has. This is a question for the evaluation, and this paper asks what the standard metrics answer.

In this paper we show that the standard metrics do not distinguish the two failures. Each reduces to a feature extractor, a score assigned to every generated sample, and an average of these scores. Even when a metric penalises hallucinations more than poor inliers, a single value is consistent with many combinations of hallucination rate and inlier quality, and a model that hallucinates more but renders better can score as well or better. We then analyse how much inlier quality a hallucination is worth to the score and identify three limits. The feature extractor may not represent the semantic deviation, the per-sample score may miss it even when it is represented, and the average dilutes it because hallucinations are rare. Averaging only the most atypical samples removes the dilution but not the first two limits. We validate our claims with both controlled experiments and trained generative models, supported by manual evaluation.

2 Related Work

Evaluating generative models. A generated set is judged on two properties. *Fidelity* asks whether its samples resemble instances from the target distribution, and *diversity* whether they cover it. The two fail independently, since a model that emits one perfect image has full fidelity and no diversity (Sajjadi et al., 2018). The Fréchet Distance (FD) (Heusel et al., 2017) scores both at once, comparing the first two moments of the generated and reference feature distributions, so a single value cannot say which of the two a model lacks (Sajjadi et al., 2018). Although metrics like the Maximum Mean Discrepancy (MMD) mitigate FD's rigid normality assumptions, they similarly yield a single aggregate distance score (Jayasumana et al., 2024). Precision and Recall were introduced to separate them. The first measures the share of generated samples that fall inside an estimate of the reference support and the second the share of reference samples covered by the generated one (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). Density and Coverage replace the binary support test by a count of covering neighbourhoods to reduce the effect of outliers on the estimate (Naeem et al., 2020). In this paper we show that the separation stops one level too early. A fidelity metric depends on two things at once: how well the inliers are rendered and how often the model hallucinates. Depending on the hallucination, the metric ignores it, trades it against inlier quality, or even rewards it.

Off-target generation. Samples that do not belong to the target are treated as a problem to be engineered away, from two origins. One line assumes the training set is contaminated with a fraction of out-of-distribution samples, or *outliers*, and recovers the clean distribution. Most of it makes optimal transport robust by down-weighting or reassigning the mass the outliers contribute (Balaji et al., 2020), absorbing the resulting mass mismatch through the semi-dual formulation of unbalanced transport (Choi et al., 2023), or estimating Wasserstein distances under contamination by duality (Nietert et al., 2022) and median-of-means aggregation (Staerman et al., 2021); the rest builds the robustness into the generative process itself, adapting the denoising process to a corrupted training set (Dao et al., 2024) or separating contaminant mass from unlabelled data in a positive-unlabelled setup (Tian et al., 2023). The other line assumes clean data and locates the failure in the model. Aithal et al. (2024) give the one explicit definition we are aware of, calling a sample a hallucination when it lies outside the support of the training distribution, and attribute it to mode interpolation, the smoothed score function interpolating between nearby training modes; the definition is stated for diffusion models and tied to that mechanism. The mitigations that follow correct the sampling trajectory, by sharpening the score along artifact directions (Triaridis et al., 2026), modulating attention during early denoising (Oorloff et al., 2025), amplifying its tangential component (Cho et al., 2026),

or detecting the steps at which the trajectory leaves the manifold (Tian et al., 2025; Sobieski et al., 2026). The two origins differ but the goal of the interventions does not: reduce the generation of samples that deviate semantically from what the model was meant to produce. From the standpoint of evaluation, generated outliers are a specific case of hallucinations, and we use the latter term for both. Both lines report progress using the standard metrics above. In this work, we show that an improved score may reflect better-rendered inliers alongside as many or even more hallucinations, or even hallucinations that the metric rewards.

3 TWO FAILURES, ONE NUMBER

In this section we show that standard metrics conflate the quality of the inliers with how many hallucinations a model produces. Trading one against the other traces curves of constant metric value, and the slope of these curves, the exchange rate between the two failures, determines whether a metric tracks hallucinations, inlier quality, or both.

3.1 METRICS EMBED, SCORE AND AGGREGATE

Suppose we had to build a metric. We are looking for generated samples that deviate from a normality only the reference set defines, and we expect them to be rare. The standard instrument is anomaly detection (Nassif et al., 2021): fit a model of normality on the reference data, score each new sample by how far it departs, and summarise the scores. That is three choices. For generated images $x \sim P_g$, a feature extractor ϕ maps each image to a representation $z = \phi(x)$, whose distribution we denote $\mathcal{Z}_{P_g}$; applied to reference images $x \sim P_r$, it yields $\mathcal{Z}_{P_r}$. A scorer A, fitted on $\mathcal{Z}_{P_r}$, assigns each generated sample a scalar $s(x) = A(z)$, larger meaning more anomalous, and a statistic T reduces the distribution of these scores, $\mathcal{S}_{P_g}$, to one number:

$$M(P_g) \;=\; T\big[\mathcal{S}_{P_g}\big]. \tag{1}$$

Standard metrics already follow this recipe, which lets us analyse them under one formulation. Precision, Density, FD and MMD all take T to be the mean and differ only in A. Precision scores a sample by whether it falls outside the reference support estimated from k-nearest-neighbour balls, and Density by how many of these balls cover it, so both coarsen the neighbourhood geometry a k-nearest-neighbour scorer reads directly. FD scores a sample by its squared distance to the reference centroid and MMD by its negative mean kernel similarity to the reference. FD and MMD also add a term in Equation 1 that depends on the spread of the generated distribution, tracking diversity. We omit this term because hallucinations and poor inliers are fidelity failures: both are properties of the samples a model generates, not of the ones it fails to generate. Any response of this diversity term to them would credit a fidelity failure as diversity, so omitting it discards no signal that should be attributed to them. We still report Recall and Coverage, which measure diversity, to expose such cases. The reductions are in Appendix A.

3.2 A METRIC VALUE IS A CURVE, AND ITS SLOPE IS THE EXCHANGE RATE

We model the generated distribution as a mixture of an inlier component P_δ and a hallucination component Q,

$$P_g \;=\; (1-\alpha)\,P_\delta \;+\; \alpha\,Q. \tag{2}$$

As defined, the split between components is by semantics alone, so the inlier component holds every sample that keeps the defining properties of the target intact, whether it is rendered flawlessly or badly. The level δ indexes a family of corruptions induced in the distribution, ordered so that larger δ renders worse on average, with $P_0 = P_r$. It is not a per-sample quantity, and samples of P_δ vary in how well they render, some of them perfectly. The weight α is the fraction of hallucinations in the generated distribution. The two parameters vary independently. With T the mean, Equation 1 is the expected score of a generated sample, and an expectation is linear in the distribution, so Equation 2 carries through ϕ and A unchanged. Writing $m(\delta) = \mathbb{E}_{\mathcal{S}_{P_\delta}}[s]$ for the mean score of the inliers degraded to level δ and $m_Q = \mathbb{E}_{\mathcal{S}_Q}[s]$ for the mean score of the hallucinations, the metric of the mixture is

$$M(\alpha, \delta) \;=\; (1-\alpha)\,m(\delta) \;+\; \alpha\,m_Q. \tag{3}$$

Hallucination shifts weight between the two means, and degradation moves one of them. The hallucination component does not depend on δ, so m_Q is a constant. The inlier mean $m(\delta)$ is a function of δ, and we assume throughout that it is continuous, strictly increasing and differentiable in δ with $m'(\delta) > 0$, so degrading the inliers raises their mean score. Both parameters enter Equation 3, so a value of M can be reached in more than one way. Proposition 1 makes this precise for the case in which the metric penalises hallucinations more than degraded inliers.

Proposition 1. *Assume m is continuous and strictly increasing, and $m_Q > m(\delta)$ on an interval of levels δ. Then on that interval M is strictly increasing in δ for every $\alpha < 1$ and strictly increasing in α for every δ, and each attainable value M^* is taken at exactly one δ for each α. The level set $\{(\alpha, \delta) : M(\alpha, \delta) = M^*\}$ is therefore the graph of a continuous, strictly decreasing function of α.*

A given score is therefore produced not by a single pair (α, δ) but by a curve of pairs, along which more hallucinations are compensated by less degradation. The score alone cannot tell these pairs apart, and how much this matters depends on the shape of the curve. With α on the horizontal axis and δ on the vertical axis, as in all figures in §5, vertical curves have constant α, so the metric tracks hallucinations alone at any rendering quality. Horizontal curves have constant δ, so the metric tracks degradation alone and says nothing about hallucinations. Between the two, the metric mixes both failures, and the slope of the curves tells us at what rate. We call this slope the *exchange rate* ρ. Differentiating Equation 3 implicitly along a level set gives

$$\rho \;=\; \left.\frac{d\delta}{d\alpha}\right|_M \;=\; -\,\frac{\partial M/\partial\alpha}{\partial M/\partial\delta} \;=\; -\,\frac{\Delta(\delta)}{(1-\alpha)\,m'(\delta)}, \qquad \Delta(\delta) = m_Q - m(\delta). \tag{4}$$

The exchange rate is the change in degradation that keeps the metric fixed per unit increase in the fraction of hallucinations. It is the ratio of two quantities, the *gap* Δ between the mean scores of hallucinations and inliers, and the *drift* m', the rate at which degradation raises the inliers' mean score. Under Proposition 1 both are positive, so $\rho < 0$. More hallucinations are paid for with less degradation, and the more negative ρ, the steeper the trade. As $\rho \to -\infty$ the contours become vertical, and the score tracks hallucinations alone and penalises them. As $\rho \to 0^-$ the contours flatten, and the score tracks degradation alone while saying almost nothing about how often the model hallucinates.

The interval of levels δ on which Proposition 1 holds is decided by the gap at $\delta = 0$. Writing

$$\Delta(\delta) \;=\; \Delta(0) \;-\; \big(m(\delta) - m(0)\big), \tag{5}$$

the gap is the separation $\Delta(0)$ between hallucinations and *clean inliers* (the undegraded component $P_0 = P_r$) minus the drift the inliers have accumulated since $\delta = 0$. The drift is positive, so the gap only shrinks. Starting from $\Delta(0) > 0$, it stays positive until the accumulated drift has used up $\Delta(0)$, at the level δ^* where degraded inliers score exactly as high as hallucinations, $m(\delta^*) = m_Q$. The interval is therefore $[0, \delta^*)$. When the extractor separates hallucinations well, δ^* is large and corresponds to severe inlier degradation. When it barely separates them, δ^* is small and mild inlier degradation is enough to reach it. At δ^* the two components have the same mean score, so mixing them leaves the score unchanged, $\rho = 0$, and the contour is the horizontal line $\delta = \delta^*$. Above δ^* hallucinations score lower than the degraded inliers, so $\Delta < 0$ and $\rho > 0$. The contour rises, and adding hallucinations improves the score. As $\rho \to +\infty$ the contour is again vertical and the score again tracks hallucinations alone, but now rewards them. Both vertical limits decouple the two failures and track hallucinations alone, but with opposite effects. As $\rho \to -\infty$ adding hallucinations worsens the score, and as $\rho \to +\infty$ it improves it. Papers that report a better score as evidence of fewer hallucinations implicitly assume ρ is large and negative, an assumption that nothing in Equation 1 guarantees.

4 What the extractor permits

Section 3.2 reduced the exchange rate to two quantities of the score distribution, the gap $\Delta(\delta)$ and the drift $m'(\delta)$. The sign of the gap decides whether the metric penalises or rewards hallucinations, and its ratio to the drift decides how steeply. Both are computed from means of A over feature distributions, and A sees nothing else, so the extractor limits them before any scorer is chosen.

Proposition 2. *Let A be bounded and normalized to unit range, $\sup A - \inf A = 1$, which leaves ρ unchanged, and let $\mathrm{TV}(\mu_1, \mu_2) = \sup_B |\mu_1(B) - \mu_2(B)|$ be the total variation distance between*

probability distributions μ_1, μ_2 on feature space, the supremum running over regions B of it. Then for every such A,

$$\big|\Delta(\delta)\big| \;\le\; \mathrm{TV}(\mathcal{Z}_Q, \mathcal{Z}_{P_\delta}) \;=\; G(\delta), \qquad m'(\delta) \;\le\; \limsup_{\delta' \to \delta} \frac{\mathrm{TV}\big(\mathcal{Z}_{P_{\delta'}}, \mathcal{Z}_{P_\delta}\big)}{|\delta' - \delta|} \;=\; D(\delta). \tag{6}$$

The proof is in Appendix B. Total variation is the most two distributions can disagree about the probability of any region, zero when they coincide and one when they do not overlap. It is taken here between feature distributions, so neither ceiling mentions A. With the generator fixed, both are properties of the extractor alone. When $G(0)$ is small and D is not, the gap is forced toward zero at the clean inliers while the drift is free to be large. Since $|\Delta(0)| \le G(0)$, every scorer whose drift is well above $G(0)$ reaches δ^* after little degradation and rewards hallucinations from there on. Hallucinations cannot be detected from a representation that does not carry them, whatever the scorer is fitted on and however the scores are aggregated. This is the one case in which rewarding hallucinations after little degradation can safely be attributed to the extractor (§5.1.2). When both are large the representation precludes nothing, and flat contours are not the extractor's doing. Hallucinations can sit far from the inlier component, and a small increase in δ can already move that component appreciably in feature space. The discriminative signal is there, and whether it steepens the contour depends on the two later choices. A scorer fitted on P_r can miss a gap the features carry (§5.1.2), and a mean can dilute a gap the scorer finds (§5.1.3). When $G(0)$ is large and D is small the extractor has done everything it can. A gap is permitted and degradation cannot move the inlier mean score quickly, so a scorer that finds a positive gap keeps it up to a large δ^* and can get a steep contour that penalises hallucinations. When both are small the extractor sees neither failure. Hallucinations and poor inliers land where clean inliers do, so no scorer responds to either parameter and the metric is near constant over the plane.

5 Experiments

We first verify the analysis of Sections 3 and 4 on a controlled pool in which hallucinations and inlier degradation are set by construction. §5.1.1 experimentally demonstrates that the standard metrics exhibit all three regimes discussed in §3.2, namely $\rho > 0$, $\rho < 0$, and $\rho \approx 0$. §5.1.2 separates a representation that does not carry the hallucinations from a scorer that does not read them, and §5.1.3 shows that in the regime of Proposition 1 replacing the mean with a tail statistic removes the trade-off. The same effects appear on real outputs, where nothing is set by construction. §5.2.1 shows that standard extractors represent a degenerate failure distinguishably but carry no information about a well-rendered semantic hallucination. §5.2.2 applies the tail statistic of §5.1.3 to three models and finds that it inverts the ranking the standard metrics give: the model with the stronger inlier component is the one that hallucinates the most. We report FD, MMD, Precision, Density, Recall and Coverage, each computed in the feature spaces of InceptionV3 (Szegedy et al., 2016), CLIP (Radford et al., 2021) and DINOv2 (Oquab et al., 2023).

5.1 Controlled Experiments

In the following experiments, we construct the generated distribution according to the mixture defined in Equation 2. The inlier component P_δ is formed by drawing samples from a clean dataset and applying an elastic distortion (Simard et al., 2003) of severity δ. The hallucination component Q is populated with samples from another dataset, or from the same dataset with different attributes, to mimic both severe and slight semantic deviations from the target distribution. We mix these components using a fraction α, representing the proportion of hallucinations in the generated set. Severe hallucinations are represented by `faces` (injecting animal faces into human faces) and `scenes` (injecting bedrooms into churches). Slight hallucinations use attribute-based subsets, such as `bald` (where inliers are men with hair and hallucinations are bald men, representing the removal of a feature the model always observed) and `glasses` (where inliers lack glasses and hallucinations wear them, representing the addition of a never-before-seen attribute). We also include a compositional `shapes` scenario that tests counting hallucinations, where inliers render 2, 4, or 6 colored objects while hallucinations render 3 or 5 objects. Finally, the reference set P_r consists of clean inliers corresponding to P_δ at $\delta = 0$, sampled disjointly from the same underlying dataset. Details and examples in Appendix C.

5.1.1 CONTOURS AND STEEPNESS

We evaluate the metrics over a grid of hallucination fractions α and degradation severities δ. At each grid point we build one pseudo-generated set $P_g(\alpha, \delta)$ and evaluate its embeddings against a single, fixed reference set P_r. Each metric thus becomes a surface over the (α, δ) plane, whose level sets are the contours of §3.2. Figure 2 shows these surfaces in DINOv2 space for a severe (`faces`) and a slight (`bald`) hallucination scenario. On `faces` the contours fall across the grid, the case Proposition 1 describes, where a model with more hallucinations and less inlier degradation scores the same as one with fewer hallucinations and more inlier degradation. On `bald` they flatten and the metric tracks inlier degradation almost alone, whatever the fraction of hallucinations.

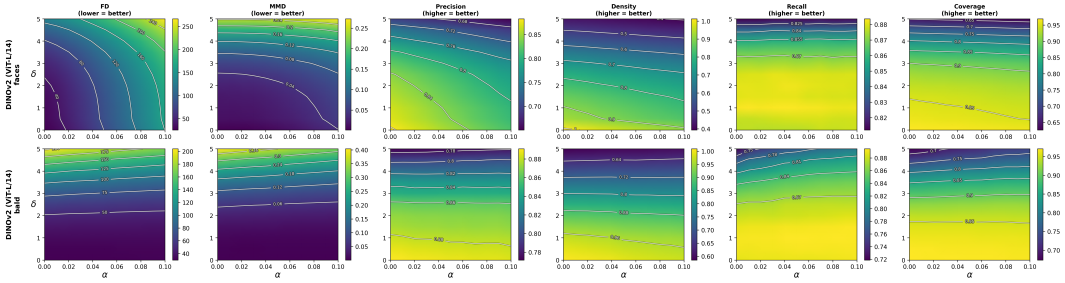

Figure 2: Standard metrics over the (α, δ) grid in DINOv2 space for `faces` (top) and `bald` (bottom). In `faces` the failures trade off; in `bald` the metrics react only to inlier degradation.

Figure 3 reduces each (α, δ) sweep to a single number. Every bar corresponds to one metric in one feature space on one scenario, summarised by the median ρ estimated by finite differences over the grid.

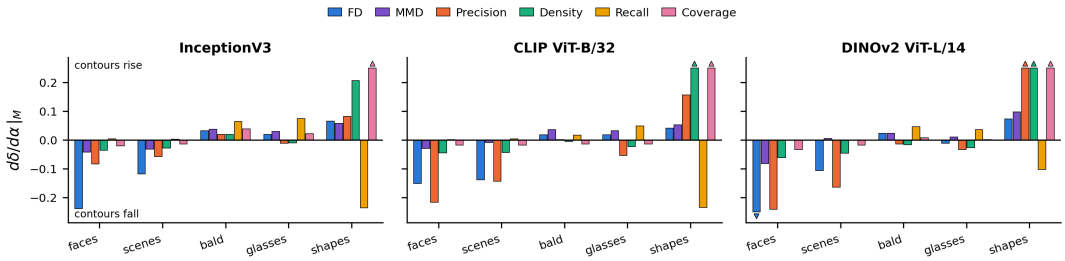

Figure 3: Median ρ per scenario, metric and extractor. $\rho < 0$: failures trade off; $\rho > 0$: hallucinations improve the metric; $\rho \approx 0$: only inlier degradation registers.

On `faces` and `scenes` degradation does not close the gap Δ within the tested range, so δ^* lies beyond it and $\rho < 0$ throughout, which is the regime Proposition 1 describes. The metrics agree on direction here and disagree on depth, with FD and Precision the deepest and MMD and Density much shallower. On `bald` and `glasses` the gap Δ stays near zero across the grid, so every cell behaves as at δ^*, where hallucinations and inliers score alike, $\rho \approx 0$, and the metrics track inlier degradation alone. On `shapes` even the mildest degradation already takes the inliers past δ^*, so throughout the grid the hallucinations score below the degraded inliers and $\rho > 0$. Precision and Density gain the most from hallucinations, FD and MMD the least. Recall and Coverage behave as §3 predicts on `faces`, `scenes`, `bald` and `glasses`. Hallucinations are a fidelity failure. So long as they fall outside the reference, adding them does not change whether it is covered, so ρ stays near zero for both. Figure 2 shows this on `faces`, where their contours stay horizontal while the fidelity metrics run diagonally, and their surfaces stay high over the grid, since inlier degradation leaves the reference modes covered. `shapes` is the exception, because there the hallucinations do not fall outside the reference at all. They cover it themselves, while degradation stops everything else from covering it. Coverage counts any sample landing near the reference, so hallucinations only add to it and ρ turns positive, meaning degradation must be added to hold the score fixed. Recall asks the reverse, where more hallucinations make each one count for less, so ρ turns negative and degradation must be removed instead.

5.1.2 MEASURING THE CEILINGS

Proposition 2 bounds both quantities determining ρ, with ceilings that are properties of the feature extractor alone. By examining those, we can diagnose whether the invariance or rewarding of hallucinations can be attributed to the feature extractor and not to the scorer or the statistic. Both ceilings are Total Variation distances (§4). $G(0)$, between hallucinations and clean inliers, bounds the separation $\Delta(0)$, and since the gap only shrinks with δ, also every positive gap. D, between inliers at consecutive degradation levels, bounds the drift. Total Variation is a supremum over regions of feature space, which we estimate as follows. For $G(0)$, we randomly split the hallucinations and the clean inliers into two equal parts. On the first part we train a logistic regression to tell them apart; each threshold on its output defines a region. On the second part, we compute for each threshold the fraction of hallucinations inside its region minus the fraction of clean inliers inside it, and report the largest value over thresholds, averaged over three random splits. A value of 0.5, for example, means that the best region contains 50 percentage points more of all hallucinations than of all clean inliers, such as 70% against 20%, with 1 for complete separation and 0 for none. Since we only search regions a linear classifier can draw, the estimate is a lower bound on the true Total Variation. D is estimated the same way between inliers at consecutive degradation levels, divided by the spacing between the levels, and we report the median over levels. Table 1 reports both.

Table 1: Gap ceiling $G(0)$ and drift ceiling D per scenario and feature extractor.

	Inception V3		CLIP ViT-B/32		DINOv2 ViT-L/14	
	$G(0)$	D	$G(0)$	D	$G(0)$	D
`faces`	1.00	0.19	1.00	0.22	1.00	0.23
`scenes`	1.00	0.24	1.00	0.30	1.00	0.18
`bald`	0.69	0.20	0.88	0.22	0.67	0.21
`glasses`	0.71	0.25	0.97	0.30	0.91	0.21
`shapes`	0.19	0.87	0.36	0.77	0.30	0.83

The ceilings tell, for the three regimes of Figure 3, whether a failure lies with the extractor or with the scorer. On `faces` and `scenes` the features separate hallucinations from clean inliers completely ($G(0) = 1$), and stay that way across the sweep, so the metrics react to both failures and the contours are diagonal. On `bald` and `glasses` the separation is smaller but still substantial ($G(0)$ between 0.67 and 0.97), with drift ceilings comparable to `faces` and `scenes`. The information is in the features, yet the metrics barely react to hallucinations and the contours are horizontal: this is a failure of the scorer, not of the extractor. On `shapes` the best region we find separates hallucinations from clean inliers by no more than 0.36, while the drift ceiling is three to four times that of any other scenario. Whatever the scorer, the gap it can find is small, and a little degradation is enough to make inliers look as anomalous as hallucinations. This is a property of the extractor, and no scorer can recover it.

5.1.3 AGGREGATION

We have seen that the extractor or the scorer can hide hallucinations, or even reward them. Even when neither does, the metric value remains underdetermined (Proposition 1), and the mean dilutes the rare hallucinations by giving each the same weight as an inlier. A statistic that reads only the top of the score distribution avoids this. If the scorer ranks the hallucinations above the degraded inliers and $\alpha \leq \tau$, the top τ fraction of scores holds every hallucination and only the worst inliers, so its mean weighs each hallucination $1/\tau$ times more than the full mean does and the contours steepen. We test this with A_{knn}, which scores a generated sample by its mean distance to its $k = 5$ nearest reference samples in feature space, and pool the scores by their mean or by the mean of the top $\tau = 0.10$ fraction. On `faces` and `scenes` A_{knn} separates hallucinations from clean inliers almost perfectly (AUROC ≥ 0.99), and the contours steepen towards vertical in every feature space (Figure 4). On `bald`, `glasses` and `shapes` its AUROC is around 0.5, so $\rho \approx 0$ under both poolings and the metric tracks degradation only. On `bald` and `glasses` the features separate the hallucinations (Table 1) but A_{knn} does not, a failure of the scorer. On `shapes` the standard metrics give $\rho > 0$ (Figure 3), since once degradation moves the inliers outside the reference data, further degradation barely changes these metrics while each hallucination, which still resembles the reference, improves them. A_{knn} instead keeps growing with degradation, so degradation dominates.

Appendix D shows the full surfaces for `faces` and `bald` and the effect of τ, including the loss of sensitivity once $\alpha > \tau$.

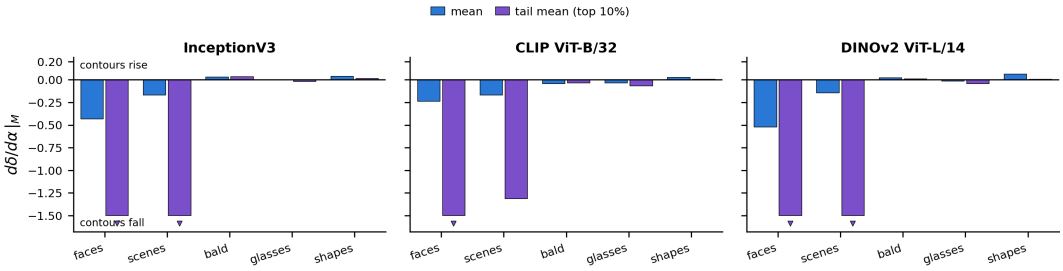

Figure 4: The exchange rate of Figure 3 under mean and tail-mean pooling of A_{knn}, at $\tau = 0.10$.

5.2 Real Generative Models

In the controlled pool, α and δ are set by construction. On real generators neither is known, and we ask whether the same effects appear. §5.2.1 tests whether the extractors represent a known hallucination, and §5.2.2 compares the tail statistic with the standard metrics across three models.

5.2.1 Perfectly rendered hallucinations go unnoticed

Aithal et al. (2024) report that diffusion models trained on 11k Hands (Afifi, 2019) generate hands with the wrong number of fingers, which they attribute to mode interpolation. We follow their setup and train an ADM (Nichol & Dhariwal, 2021) on this dataset. We then manually label 225 generated samples into three classes: correct hands (1), hands with a wrong finger count (2), and samples that are not hands at all (3). This gives two ranking tasks with known ground truth, which let us ask whether each failure is carried by the representation and whether a scorer fitted on the reference finds it. The first ranks wrong finger counts against correct hands (*2v1*), a hallucination rendered as well as the inliers. The second ranks non-hands against correct hands (*3v1*), a degenerate failure.

Table 2: AP for ranking wrong finger counts (2v1) and non-hands (3v1) against correct hands.

	InceptionV3		CLIP ViT-B/32		DINOv2 ViT-L/14	
	AP_{2v1}	AP_{3v1}	AP_{2v1}	AP_{3v1}	AP_{2v1}	AP_{3v1}
Random (prevalence)	0.125	0.279	0.125	0.279	0.125	0.279
kNN	0.154	**0.989**	0.152	0.995	**0.169**	**0.991**
Mahalanobis	0.145	0.940	0.136	0.995	0.161	0.981
LOF (Breunig et al., 2000)	0.142	0.971	0.133	**0.996**	0.162	0.936
IForest (Liu et al., 2008)	**0.158**	0.803	**0.154**	0.992	0.148	0.963
OCSVM (Schölkopf et al., 2001)	0.146	0.917	0.148	0.995	0.148	0.983
PCA recon.	0.142	0.939	0.138	**0.996**	0.155	0.984
ECOD (Li et al., 2022)	**0.158**	0.867	0.150	0.993	0.143	0.970
COPOD (Li et al., 2020)	0.157	0.783	0.153	0.993	0.136	0.975
Supervised Probe	0.202	0.992	0.174	0.996	0.296	0.992

For each task we rank the samples with eight anomaly detectors fitted on the reference embeddings alone, as the scorer of a metric is, and report average precision (AP) in Table 2. Two rows bound the detectors. A random ordering gives the floor, since its AP equals the share of the positive class, and a supervised probe (logistic regression, 5-fold cross-validation on the labels) gives a ceiling on what the representation carries. On *2v1* the probe reaches 0.20 in Inception and 0.17 in CLIP against a floor of 0.125, so these representations carry little of the finger count, as on `shapes`. DINOv2 carries clearly more, with a ceiling of 0.30, yet the best detector, A_{knn}, reaches only 0.17 and closes about a quarter of the gap between floor and ceiling. The signal is in the features but no detector finds it, as on `bald` and `glasses`. On *3v1*, by contrast, the task is easy. Every detector finds the non-hands in every space, with AP of at least 0.78 against a floor of 0.28, and A_{knn} reaches 0.99. As in the controlled experiments, severe hallucinations are readily distinguished, while a slight semantic deviation in an otherwise well-rendered sample goes unnoticed.

5.2.2 RANKING INVERSION

§5.2.1 showed that detectors fitted on the reference find degenerate samples but miss a well-rendered hallucination. §5.1.3 showed that when the scorer ranks hallucinations above the inliers, averaging only the top of the score distribution decouples hallucination from inlier degradation. On real generators, such a tail statistic should therefore at least read the most degenerate samples, and may rank models differently from the standard metrics. We score three models trained on CelebA-HQ against one reference set: StyleGAN2 (Karras et al., 2020), DDPM (Ho et al., 2020) and LDM (Rombach et al., 2022). Next to the standard metrics we report the mean of the top 1% of A_{knn} scores, and Table 3 gives each model's rank under every metric.

Table 3: Rank of each generator under each metric (1 = best), within each feature extractor. $\text{Tail}_{0.01}$ is the mean of the top 1% of A_{knn} scores. Raw scores in Appendix E, Table 4.

Model	FD	MMD	Precision	Density	Recall	Coverage	$\text{Tail}_{0.01}$
InceptionV3							
StyleGAN2	1	1	1	1	2	1	2
DDPM	3	3	3	3	3	3	1
LDM	2	2	2	2	1	2	3
CLIP ViT-B/32							
StyleGAN2	1	1	1	1	3	1	3
DDPM	2	3	3	3	2	3	1
LDM	3	2	2	2	1	2	2
DINOv2 ViT-L/14							
StyleGAN2	1	1	1	1	3	1	3
DDPM	2	2	2	3	1	2	1
LDM	3	3	3	2	2	3	2

The tail ranks the models differently. FD, MMD, Precision, Density and Coverage place StyleGAN2 ahead of both diffusion models in every feature space, while the tail places it behind DDPM, and in CLIP and DINOv2 also behind LDM. To test which ordering is right, three annotators rated the top 1% of samples of each model under A_{knn} in DINOv2 space, on a four-point scale from 1, a face a real photograph could show, to 4, no person at all. Agreement is substantial, with Krippendorff's $\alpha_K = 0.71$ and mean pairwise Cohen's $\kappa = 0.76$ over all rated samples. All three annotators order the models by mean rating as DDPM, then LDM, then StyleGAN2, the ordering the tail gives in CLIP and DINOv2, and place last the model the standard metrics place first. In Appendix E we provide details about the models and the annotation guidelines, and show samples from the top and the median of each model's ranking. The disagreement is the trade of Proposition 1. The worst samples of StyleGAN2 are mostly not faces, a degenerate failure the extractors separate well (§5.2.1), so the standard metrics penalise them and the proposition applies. The annotations show that StyleGAN2 has the largest α of the three models. Since the standard metrics still rank it first, its inliers must be degraded less, and its median samples are indeed visibly better than those of the diffusion models (Appendix E, Figure 9). The standard metrics pay for the extra hallucinations with better inliers and rank StyleGAN2 first, while the tail reads the hallucinations and ranks it lower.

6 CONCLUSION

We showed that a single value of the standard metrics is consistent with many combinations of hallucination rate and inlier quality, and that their response to hallucinations is limited at three stages: the extractor may not represent them, the scorer may not read them, and the mean dilutes them. A tail statistic removes the dilution and, on trained models, demotes the model the standard metrics rank first, which annotators confirm hallucinates most. Well-rendered hallucinations still go unnoticed, so we recommend reporting a tail statistic alongside the standard metrics and not reading a good score as evidence of few hallucinations.

AI USE STATEMENT

In this work, we used generative AI tools to implement parts of the experimental code and to discuss our ideas and claims in order to obtain feedback on them. Additionally, we used generative AI tools to improve the grammar and readability of the manuscript. We did not use them to search for, summarise or cite related work. AI-generated code was manually reviewed and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The experimental details needed to reproduce our results are given in the appendix. The full code, with instructions to reproduce all experiments, will be released publicly upon publication.

REFERENCES

Mahmoud Afifi. 11k hands: Gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 78(15):20835–20854, 2019.

Sumukh K Aithal, Pratyush Maini, Zachary C Lipton, and J Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. *Advances in neural information processing systems*, 37:134614–134644, 2024.

Yogesh Balaji, Rama Chellappa, and Soheil Feizi. Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems*, 33:12934–12944, 2020.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Hyunmin Cho, Donghoon Ahn, Susung Hong, Jee Eun Kim, Seungryong Kim, and Kyong Hwan Jin. TAG: Tangential amplifying guidance for hallucination-resistant sampling. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=1XJqPy2LhA.

Jaemoo Choi, Jaewoong Choi, and Myungjoo Kang. Generative modeling through the semi-dual formulation of unbalanced optimal transport. *Advances in Neural Information Processing Systems*, 36:42433–42455, 2023.

Quan Dao, Binh Ta, Tung Pham, and Anh Tran. A high-quality robust diffusion framework for corrupted dataset. In *European Conference on Computer Vision*, pp. 107–123. Springer, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Denis Janiak, Jakub Binkowski, Albert Sawczyn, Bogdan Gabrys, Ravid Shwartz-Ziv, and Tomasz Jan Kajdanowicz. The illusion of progress: Re-evaluating hallucination detection in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34728–34745, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1761. URL https://aclanthology.org/2025.emnlp-main.1761/.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315. IEEE, 2024.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116. IEEE, 2020.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems*, 32, 2019.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 6449–6464, 2023.

Zheng Li, Yue Zhao, Nicola Botta, Cezar Ionescu, and Xiyang Hu. Copod: copula-based outlier detection. In *2020 IEEE international conference on data mining (ICDM)*, pp. 1118–1123. IEEE, 2020.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George H Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12181–12193, 2022.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pp. 413–422. Ieee, 2008.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *International conference on machine learning*, pp. 7176–7185. PMLR, 2020.

Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, and Fatima Mohamad Dakalbab. Machine learning for anomaly detection: A systematic review. *IEEE access*, 9:78658–78700, 2021.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PmLR, 2021.

Sloan Nietert, Ziv Goldfeld, and Rachel Cummings. Outlier-robust optimal transport: Duality, structure, and statistical analysis. In *International Conference on Artificial Intelligence and Statistics*, pp. 11691–11719. PMLR, 2022.

Trevine Oorloff, Yaser Yacoob, and Abhinav Shrivastava. Mitigating hallucinations in diffusion models through adaptive attention modulation, 2025. URL https://arxiv.org/abs/2502.16872.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. *Advances in neural information processing systems*, 31, 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.

Patrice Y. Simard, David Steinkraus, and John C. Platt. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR)*, pp. 958–963, 2003. doi: 10.1109/ICDAR.2003.1227801.

Bartlomiej Sobieski, Matthew Tivnan, Dawid Płudowski, Michał Jan Włodarczyk, Pengfei Jin, Przemyslaw Biecek, and Quanzheng Li. Local intrinsic dimension unveils hallucinations in diffusion models, 2026. URL https://arxiv.org/abs/2605.05026.

Guillaume Staerman, Pierre Laforgue, Pavlo Mozharovskyi, and Florence d'Alché Buc. When ot meets mom: Robust estimation of wasserstein distance. In *International Conference on Artificial Intelligence and Statistics*, pp. 136–144. PMLR, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Bowen Tian, Qinliang Su, and Jianxing Yu. Leveraging contaminated datasets to learn clean-data distribution with purified generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 9989–9996, 2023.

Yiqi Tian, Pengfei Jin, Mingze Yuan, Na Li, Bo Zeng, and Quanzheng Li. RODS: Robust optimization inspired diffusion sampling for detecting and reducing hallucination in generative models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 25782–25814, 2025.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*, 2024.

Kostas Triaridis, Alexandros Graikos, Aggelina Chatziagapi, Grigorios G. Chrysos, and Dimitris Samaras. Mitigating diffusion model hallucinations with dynamic guidance, 2026. URL https://arxiv.org/abs/2510.05356.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. Siren's song in the AI ocean: A survey on hallucination in large language models. *Computational Linguistics*, 51(4):1373–1418, December 2025. doi: 10.1162/coli.a.16. URL https://aclanthology.org/2025.cl-4.9/.

A Evaluation metrics in the scoring form

Metrics for which a larger value is better are negated, so that A is an anomaly score throughout. An additive constant in A shifts the metric by the same amount at every (α, δ), so constants below are dropped. Write $z = \phi(x)$ for a feature vector, μ_r, Σ_r for the mean and covariance of the reference

features and μ_g, Σ_g for those of the generated features. Precision and density both use reference neighbourhoods: for a reference feature $z_r = \phi(x_r)$, let $B(z_r)$ be the ball centred at z_r whose radius is the distance from z_r to its k-th nearest reference neighbour.

Precision. With $\mathcal{R} = \bigcup_{z_r} B(z_r)$ the estimated reference support, precision is $\mathbb{E}_{Z_{P_g}} \mathbf{1}[\, z \in \mathcal{R}\,]$. Larger is better, so we negate: taking $A_{\mathrm{Prec}}(z) = \mathbf{1}[\, z \notin \mathcal{R}\,]$ gives $-\mathrm{Precision} = \mathbb{E}_{Z_{P_g}}[A_{\mathrm{Prec}}] - 1$, the mean of A_{Prec} up to an additive constant. The support estimate is built from P_r alone. It is binary: a sample marginally outside the estimated support and one far outside receive the same score.

Density. Density replaces the indicator by the number of covering neighbourhoods, $\frac{1}{k}\,\mathbb{E}_{Z_{P_g}} \sum_{z_r} \mathbf{1}[\, z \in B(z_r)\,]$. Negating,

$$A_{\mathrm{Dens}}(z) = -\frac{1}{k} \sum_{z_r} \mathbf{1}[\, z \in B(z_r)\,], \tag{7}$$

and $-\mathrm{Density} = \mathbb{E}_{P_g}[A_{\mathrm{Dens}}]$ exactly. Density records severity more finely than precision, but still only through how many reference balls overlap a point.

Fréchet Distance. The Fréchet Distance is

$$\mathrm{FD} = \|\mu_r - \mu_g\|^2 + \mathrm{Tr}\,\Sigma_g + \mathrm{Tr}\,\Sigma_r - 2\,\mathrm{Tr}\left((\Sigma_r \Sigma_g)^{1/2}\right). \tag{8}$$

Take $A_{\mathrm{FD}}(z) = \|z - \mu_r\|^2$. Writing $z - \mu_r = (z - \mu_g) + (\mu_g - \mu_r)$,

$$\mathbb{E}_{Z_{P_g}} \|z-\mu_r\|^2 = \mathbb{E}_{Z_{P_g}} \|z-\mu_g\|^2 + 2\big\langle \mathbb{E}_{Z_{P_g}}[z]-\mu_g,\, \mu_g-\mu_r \big\rangle + \|\mu_g-\mu_r\|^2 = \mathrm{Tr}\,\Sigma_g + \|\mu_r-\mu_g\|^2, \tag{9}$$

since $\mathbb{E}_{Z_{P_g}}[z] = \mu_g$ removes the cross term. Hence

$$\mathrm{FD} = \mathbb{E}_{Z_{P_g}} A_{\mathrm{FD}}(z) + C(P_g), \qquad C(P_g) = \mathrm{Tr}\,\Sigma_r - 2\,\mathrm{Tr}\left((\Sigma_r \Sigma_g)^{1/2}\right). \tag{10}$$

The first term is the expected score of a scorer fitted on the reference and tracks fidelity. C depends on the generated set as a whole, through Σ_g, rather than on single samples. It falls as the generated features spread, so it penalises a concentrated generated set, as when modes are dropped: it is the diversity term. It is not negligible: at $P_g = P_r$ it equals minus the first term, and FD vanishes only because the two cancel. FD therefore conflates fidelity with diversity.

Maximum Mean Discrepancy. Following CMMD (Jayasumana et al., 2024), we use the squared MMD with a Gaussian kernel k. Write $\kappa_g = \mathbb{E}_{z,z' \sim Z_{P_g}} k(z, z')$ and $\kappa_r = \mathbb{E}_{z_r, z_r' \sim Z_{P_r}} k(z_r, z_r')$ for the mean similarity between two independent generated samples and between two independent reference samples. Then

$$\mathrm{MMD}^2 = \kappa_g - 2\,\mathbb{E}_{Z_{P_g}} \mathbb{E}_{Z_{P_r}} k(z, z_r) + \kappa_r. \tag{11}$$

Only the middle term is an average over single generated samples. Taking $A_{\mathrm{MMD}}(z) = -2\,\mathbb{E}_{Z_{P_r}} k(z, z_r)$,

$$\mathrm{MMD}^2 = \mathbb{E}_{Z_{P_g}} A_{\mathrm{MMD}}(z) + C(P_g), \qquad C(P_g) = \kappa_g + \kappa_r. \tag{12}$$

With the Gaussian kernel, $\mathbb{E}_{Z_{P_r}} k(z, z_r)$ is, up to normalisation, a kernel density estimate of the reference at z, so A_{MMD} is density with the covering balls replaced by a smooth kernel.

As for FD, the first term tracks fidelity, and C depends on the generated set as a whole, through κ_g. κ_g is largest when the generated samples are concentrated, as when modes are dropped, so C is again the diversity term, and at $P_g = P_r$ it cancels the first term.

Recall and Coverage. Recall and Coverage measure diversity, whether the generated samples spread over the whole reference, rather than how well each generated sample fits it. Like C in FD and MMD, they are diversity terms, so we do not write them in the form of Equation 1.

B Proof of Proposition 2

The proof relies on the variational representation of the Total Variation (TV) distance. For any two probability measures μ_1, μ_2 defined on a measurable space $\mathcal{Z}$, the TV distance acts as an Integral Probability Metric (IPM) over the class of all measurable functions bounded between 0 and 1:

$$\mathrm{TV}(\mu_1, \mu_2) \;=\; \sup_{f:\mathcal{Z}\to[0,1]} \big|\mathbb{E}_{\mu_1}[f] - \mathbb{E}_{\mu_2}[f]\big|. \tag{13}$$

We use this property to bound both the score gap and the score drift.

Invariance of the exchange rate ρ. The proposition assumes without loss of generality that the scorer A is bounded and normalized to unit range ($\sup A - \inf A = 1$). To see why this leaves the exchange rate ρ unchanged, consider an arbitrary bounded scorer A_{raw}. Any affine transformation $A = \lambda A_{\text{raw}} + \beta$ (with $\lambda > 0$) shifts the mean scores by β and scales them by λ.

In the gap $\Delta(\delta) = \mathbb{E}_{Z_Q}[A] - \mathbb{E}_{Z_{P_\delta}}[A]$, the shift β cancels out by the linearity of expectation, leaving $\Delta(\delta) = \lambda\Delta_{\text{raw}}(\delta)$. In the drift $m'(\delta) = \frac{d}{d\delta}\mathbb{E}_{Z_{P_\delta}}[A]$, the derivative of the constant β is zero, leaving $m'(\delta) = \lambda m'_{\text{raw}}(\delta)$.

Recall from Equation 4 that the exchange rate is $\rho = -\frac{\Delta(\delta)}{(1-\alpha)m'(\delta)}$. Substituting the scaled components, the factor λ cancels out perfectly from the numerator and denominator. Thus, any scaling or shifting of the scorer leaves the geometric slope ρ identical, allowing us to strictly normalize A to a unit range to interact with the Total Variation bounds.

Bounding the Gap $\Delta(\delta)$. With A bounded and normalized to $\sup A - \inf A = 1$, let $c = \inf A$. We define a shifted scoring function:

$$\tilde{A}(z) \;=\; A(z) - c. \tag{14}$$

By construction, $\tilde{A}(z) \in [0, 1]$ for all $z \in \mathcal{Z}$, which qualifies $\tilde{A}$ as a valid test function in the supremum of Equation 13.

The score gap is $\Delta(\delta) = \mathbb{E}_{Z_Q}[A] - \mathbb{E}_{Z_{P_\delta}}[A]$. Substituting $A(z) = \tilde{A}(z) + c$, the constant c cancels out:

$$\Delta(\delta) \;=\; \mathbb{E}_{Z_Q}[\tilde{A} + c] - \mathbb{E}_{Z_{P_\delta}}[\tilde{A} + c] \;=\; \mathbb{E}_{Z_Q}[\tilde{A}] - \mathbb{E}_{Z_{P_\delta}}[\tilde{A}]. \tag{15}$$

Taking the absolute value, we bound this specific difference by the supremum over all functions bounded in $[0, 1]$:

$$\begin{aligned} \big|\Delta(\delta)\big| \;&=\; \big|\mathbb{E}_{Z_Q}[\tilde{A}] - \mathbb{E}_{Z_{P_\delta}}[\tilde{A}]\big| \\ &\le\; \sup_{f:\mathcal{Z}\to[0,1]} \big|\mathbb{E}_{Z_Q}[f] - \mathbb{E}_{Z_{P_\delta}}[f]\big| \\ &=\; \mathrm{TV}(Z_Q, Z_{P_\delta}) \;=\; G(\delta). \end{aligned} \tag{16}$$

This completes the proof for the gap bound.

Bounding the Drift $m'(\delta)$. The drift is the rate at which the mean inlier score changes. Since $m'(\delta) > 0$ by assumption, it equals its absolute value, and by the definition of the derivative

$$m'(\delta) \;=\; \lim_{\delta'\to\delta} \frac{|m(\delta') - m(\delta)|}{|\delta' - \delta|} \;=\; \lim_{\delta'\to\delta} \frac{\big|\mathbb{E}_{Z_{P_{\delta'}}}[A] - \mathbb{E}_{Z_{P_\delta}}[A]\big|}{|\delta' - \delta|}. \tag{17}$$

The numerator is the absolute difference in expectations of the bounded scorer A over two different inlier distributions. Using the exact same shifting argument as above, the shifted scorer $\tilde{A}$ belongs to the class of functions bounded in $[0, 1]$. Therefore, the numerator is bounded by the TV distance between the two nearby distributions,

$$\big|\mathbb{E}_{Z_{P_{\delta'}}}[A] - \mathbb{E}_{Z_{P_\delta}}[A]\big| \;\le\; \mathrm{TV}(Z_{P_{\delta'}}, Z_{P_\delta}). \tag{18}$$

Substituting this upper bound into the limit gives

$$m'(\delta) \;\le\; \limsup_{\delta'\to\delta} \frac{\mathrm{TV}(Z_{P_{\delta'}}, Z_{P_\delta})}{|\delta' - \delta|} \;=\; D(\delta). \tag{19}$$

□

C Controlled Experiments

Figure 5 shows samples from the inlier and hallucination pools of each scenario. We pull the datasets from Hugging Face[1] (HF). In `faces`, inliers are CelebA-HQ faces (HF: `korexyz/celeba-hq-256x256`) and hallucinations are AFHQ animal faces (HF: `huggan/AFHQ`). In `scenes`, inliers are LSUN churches (HF: `tglcourse/lsun_church_train`) and hallucinations are LSUN bedrooms (HF: `pcuenq/lsun-bedrooms`), centre-cropped to 256 pixels. `bald` and `glasses` split CelebA (HF: `tpremoli/CelebA-attrs`) by its attribute labels, stretched from 178×218 to 218×218 so that no crop removes the top of the head. Both pools are restricted to men labelled not young and not wearing a hat, so the held-out attribute is the only systematic difference between them. In `bald`, inliers have hair and exclude receding hairlines, so baldness is not the end of a range the inliers already span. In `glasses`, inliers wear no eyeglasses and hallucinations do. `shapes` is a synthetic dataset we created to test counting hallucinations. Each 256×256 image shows non-overlapping circles, squares and triangles in random colours on a dark background. Inliers contain 2, 4 or 6 objects and hallucinations contain 3 or 5, so every hallucinated count lies between two counts the inliers contain. The two pools have the same mean count, and object size is drawn independently of the count, so the area the objects cover does not separate them.



Figure 5: Inliers and hallucinations of each scenario.

Degraded inliers are produced by the elastic distortion of Simard et al. (2003). For an image x of width W pixels, two noise fields $\varepsilon_1, \varepsilon_2$ with i.i.d. $\mathcal{N}(0,1)$ entries are smoothed by a normalised Gaussian kernel g_σ and scaled by an amplitude s. This gives a displacement field in pixels,

$$\mathbf{u}(\mathbf{p}) \;=\; s\,\big((g_\sigma * \varepsilon_1)(\mathbf{p}),\; (g_\sigma * \varepsilon_2)(\mathbf{p})\big), \tag{20}$$

and each pixel of the distorted image is read from the original at its displaced location,

$$\tilde{x}(\mathbf{p}) \;=\; x\big(\mathbf{p} + \mathbf{u}(\mathbf{p})\big), \tag{21}$$

with bilinear interpolation and edge pixels replicated beyond the border. The smoothing width is $\sigma = W/7$, the $\sigma = 4$ pixels that Simard et al. (2003) use on MNIST, rescaled to width W. The distortion therefore bends whole regions of the image at every resolution rather than adding fine texture.

We depart from Simard et al. (2003) in two ways: the noise is Gaussian rather than uniform, and the amplitude s is set by the degradation level δ rather than chosen directly. Each component of $\mathbf{u}(\mathbf{p})$ is a weighted sum of independent standard normals, so it is Gaussian with standard deviation $s\|g_\sigma\|_2$, where $\|g_\sigma\|_2$ is the ℓ_2 norm of the kernel weights. The two components are independent, so $\|\mathbf{u}(\mathbf{p})\|$ is Rayleigh distributed with mean $s\|g_\sigma\|_2\sqrt{\pi/2}$. We define δ as this mean in percent of the image width,

$$\delta \;=\; \frac{100}{W}\, s\,\|g_\sigma\|_2 \sqrt{\pi/2}, \tag{22}$$

[1] https://huggingface.co

which is linear in s and is inverted for each level of the sweep. At $\delta = 0$ the distortion is the identity. At $\delta = 5$, pixels move by 5% of the width on average, about 13 pixels at $W = 256$.

The distortion is applied in pixel space at the scenario's resolution ($W = 256$, or 218 for the attribute scenarios), before any extractor preprocessing, so every extractor receives the same degraded pixels. Each pool image keeps a single noise draw at every level, so moving along δ changes only the amplitude of its deformation. Figure 6 shows the distortion of a face, a church and a `shapes` image. The distortion is applied only to inliers, so it acts on faces in `faces`, `bald` and `glasses`, on churches in `scenes`, and on the rendered objects in `shapes`.

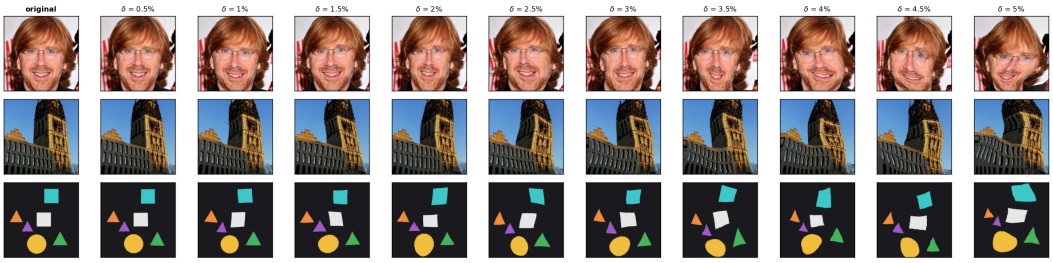

Figure 6: Inliers distorted at increasing degradation levels δ.

We embed images with three extractors. For InceptionV3 we use the TF-ported FID weights of `pytorch-fid` (`pt_inception-2015-12-05`) and take the 2048-dimensional pool3 features of images resized to 299×299, so our FD in this space is the standard FID. For CLIP ViT-B/32 (HF: `openai/clip-vit-base-patch32`) we take the 512-dimensional projected image embedding. For DINOv2 ViT-L/14 (HF: `facebook/dinov2-large`) we take the 1024-dimensional CLS token. Both ViTs receive a 224×224 centre crop, normalised with the model's own mean and standard deviation.

Each cell of the grid compares a pseudo-generated set P_g of $N = 9500$ samples with a fixed reference set P_r of $N = 9500$ clean inliers, drawn disjointly from the inlier pool. $|P_g|$ is the same in every cell, since FD is biased in the sample size. We sweep $\alpha \in \{0, 0.01, \ldots, 0.10\}$ and $\delta \in \{0, 0.5, \ldots, 5\}$, giving 121 cells per scenario. Cell (α, δ) takes the first $N - \mathrm{round}(\alpha N)$ inliers distorted at δ and the first $\mathrm{round}(\alpha N)$ hallucinations of pools drawn once with a fixed seed, so moving along either axis changes only that parameter. Precision, Recall, Density, Coverage and A_{knn} use $k = 5$ neighbours. MMD uses a Gaussian kernel of bandwidth 10 on L2-normalised features, as in CMMD (Jayasumana et al., 2024).

D Choosing the tail fraction

In practice α is unknown, so the tail fraction τ can be set above or below it. Figure 7 sweeps τ from 0.01 up to $\tau = 1$, which is the mean. On `faces` $|\rho|$ is largest for small τ and falls steadily as τ grows, from 13.6 at $\tau = 0.10$ to 0.52 at the mean, where the contours are diagonal again. On `bald` it stays near zero for every τ, because A_{knn} does not separate the hallucinations and no tail can select them.

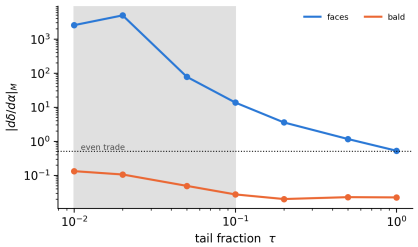

Figure 7: Median $|\rho|$ of A_{knn} in DINOv2 space against the tail fraction τ. $\tau = 1$ is the mean. The gray band marks the range of α in the grid.

Figure 8 shows the full contours at four values of τ. When $\tau > \alpha$ the tail holds every hallucination together with the worst inliers, so it reacts to both axes and the contours steepen as τ shrinks. When $\tau < \alpha$ the tail holds only hallucinations, and since they are not degraded the score becomes a constant plane. In the top 5% column the score is identical at every δ once $\alpha \geq 0.05$ and changes only from 1.350 to 1.370 between $\alpha = 0.05$ and 0.10. A tail narrower than the fraction of hallucinations therefore stops measuring both failures, and a wider one dilutes the hallucinations again.

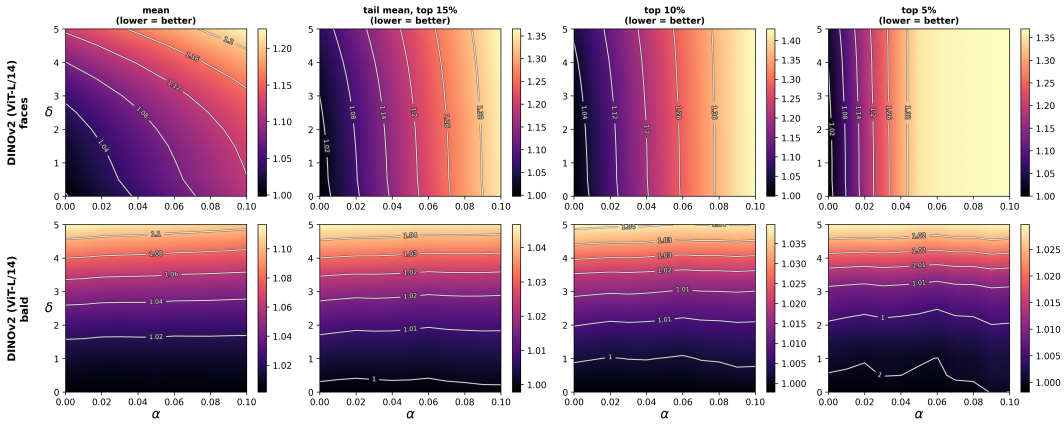

Figure 8: A_{knn} in DINOv2 space pooled by the mean and by the top 15%, 10% and 5% of scores, on `faces` (top) and `bald` (bottom).

E RANKING INVERSION

We score three generators trained on CelebA-HQ at 256×256, all from released checkpoints: StyleGAN2 (`stylegan2-celebahq-256x256.pkl`, from the StyleGAN2 models released with the official StyleGAN3 code[2]), DDPM (HF: `google/ddpm-ema-celebahq-256`) and LDM (HF: `CompVis/ldm-celebahq-256`). We generate 10,000 samples from each model, sampling DDPM with its ancestral sampler at 1000 steps and LDM with DDIM at 200 steps, and compare them with a reference set of 10,000 CelebA-HQ images. The extractors, the metric hyperparameters and A_{knn} are the same as in Appendix C.

Table 4 gives the scores behind the ranks of Table 3. FD, MMD, Precision, Density and Coverage place StyleGAN2 first in every feature space, often by a wide margin. In Inception space its FD is 9.0 against 14.2 for LDM and 26.9 for DDPM, and its Density is 1.10 against 0.82 and 0.50. Recall is the exception and places StyleGAN2 last in CLIP and DINOv2 space. Like Coverage, it asks whether the generated set covers the reference, so it is not an instrument for hallucinations. The tail orders the models differently. DDPM has the lowest tail in every space. StyleGAN2 has the highest in CLIP and DINOv2, where its worst 1% of samples is as far from the reference as the reference set's own worst 1%, and in Inception space it lies between the two diffusion models.

To test which ordering is right, three annotators rated, for each model, the 100 samples with the highest A_{knn} in DINOv2 space, the top 1% of its 10,000, on the following scale:

1 A face a real photograph could show. It may be blurry, grainy or badly rendered, which the rating does not ask about. Slight artifacts on edges or in the background are ignored.

2 A face, but a required part is missing, duplicated, highly distorted or at an unrealistic location, something no real photograph could show. Pose, occlusion, cropping or an object in front of the face do not count.

3 Parts of a face are present (skin, hair, an eye-like region) but do not assemble into a face. No photograph of a person could produce this arrangement.

4 No person in the image: texture, an object, or nothing recognisable.

[2] `https://github.com/NVlabs/stylegan3`

The scale asks whether a sample belongs to the target, not how well it is rendered, so a poor inlier is rated 1 and only hallucinations are rated 2 to 4.

Figure 9 shows, for each model, the 25 samples with the highest A_{knn} in DINOv2 space beside 25 samples from the middle of the same ranking.

Table 4: Raw scores behind Table 3, for every metric in every feature space. Lower is better for FD, MMD and Tail$_{0.01}$, and higher for the others. Tail$_{0.01}$ is the mean of the top 1% of A_{knn} scores.

Model	FD	MMD	Precision	Density	Recall	Coverage	Tail$_{0.01}$
InceptionV3							
StyleGAN2	9.026	0.067	0.780	1.105	0.547	0.892	0.962
DDPM	26.876	0.374	0.657	0.500	0.257	0.519	0.958
LDM	14.204	0.117	0.721	0.815	0.552	0.827	0.981
CLIP ViT-B/32							
StyleGAN2	5.616	0.260	0.865	0.838	0.290	0.738	0.994
DDPM	6.486	0.370	0.775	0.502	0.365	0.664	0.968
LDM	7.062	0.359	0.822	0.662	0.435	0.670	0.984
DINOv2 ViT-L/14							
StyleGAN2	143.304	0.173	0.846	1.012	0.418	0.818	0.999
DDPM	163.662	0.241	0.809	0.699	0.484	0.737	0.956
LDM	245.417	0.412	0.805	0.711	0.455	0.644	0.985

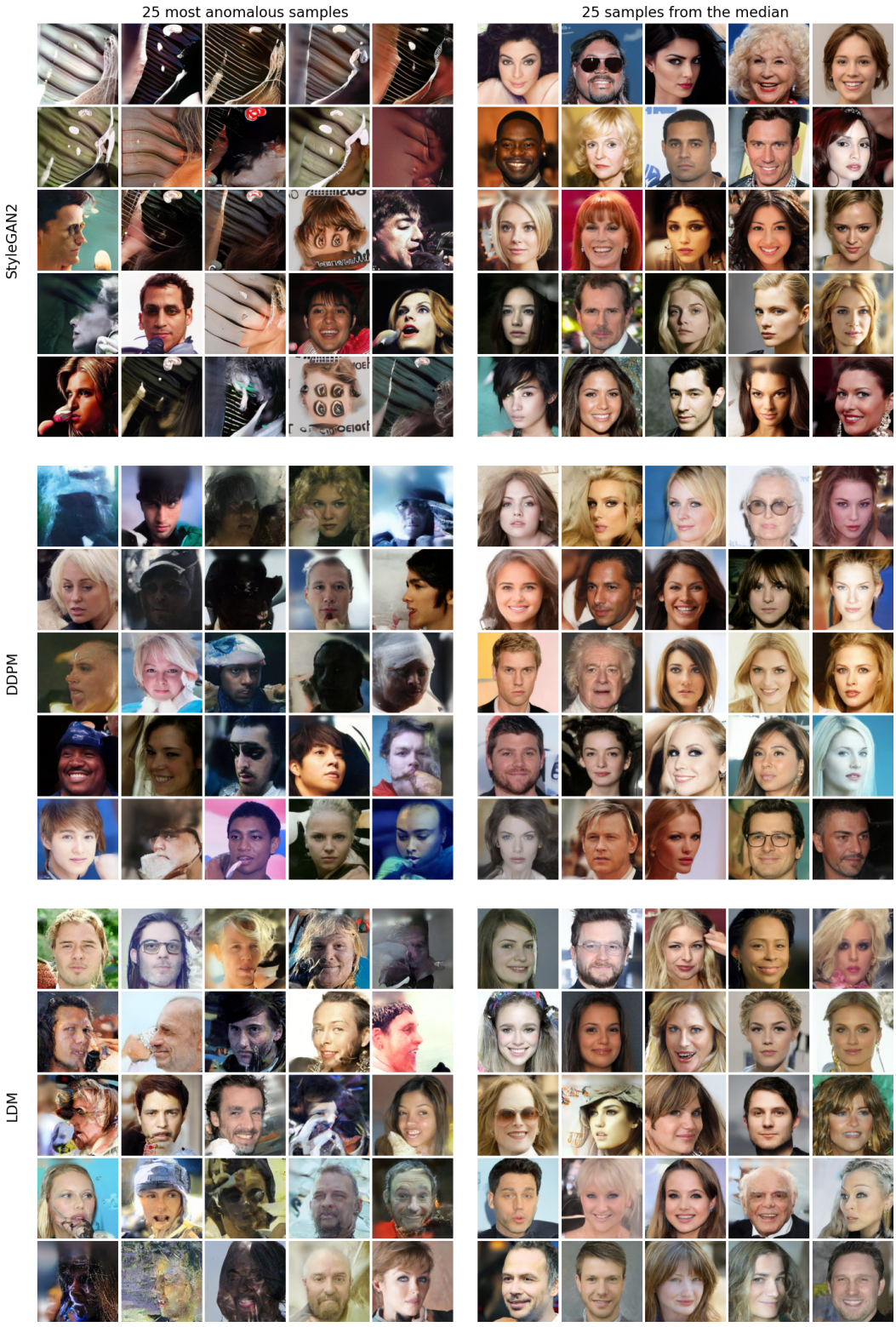

Figure 9: The 25 most anomalous samples of each generator (left) beside 25 samples from the middle of the same ranking (right), ordered by A_{knn} in DINOv2 space. StyleGAN2's most anomalous samples are largely not faces, while both diffusion models' remain coherent, but its median samples are visibly better. StyleGAN2 therefore has more hallucinations (larger α) but better inliers (smaller δ), the trade of Proposition 1, and FD, MMD, Precision and Density rank it above both diffusion models (Table 3).