WHAT FID HIDES: DETECTING, RANKING, AND DIAGNOSING DEVIATIONS IN GENERATIVE EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models are commonly ranked by Fréchet Inception Distance (FID) and Kernel Inception Distance (KID), yet FID's first-two-moment summary can miss distributional differences, and a reported scalar gap alone is not a calibrated test against sampling variation. FID's moment restriction has concrete consequences: on ImageNet, visually unrecognizable images optimized only to match the reference mean and covariance in Inception feature space obtain FID 24.7 versus 58.6 for held-out real images (lower is better). Moreover, FID and KID are scalar discrepancies that are unchanged when the two samples are exchanged and therefore do not encode the direction of a dispersion change: under-dispersion, as can occur in mode collapse, versus over-dispersion. We introduce **ZID** (*Z-resolved Integrated Diagnostic*), which combines six standardized components from a rank-based graph test and Gaussian-kernel tests at two bandwidths. Rather than asking one scalar to serve incompatible roles, ZID reports three linked outputs: a score for ranking candidate models, a permutation p-value for testing distributional equality, and a signed dispersion readout for diagnosis. In a 19-method comparison at a fixed dimension and sample size, ZID is the only method with detection power of approximately .70 or higher on each of six departures. Its score is positively associated with increasing severity across all six ranking sweeps, including cases in which FID is flat or reversed. On DiT-XL/2 and SiT-XL/2 guidance sweeps, ZID's score tracks departure from real data, and its signed readout distinguishes low-guidance over-dispersion from high-guidance under-dispersion.

1 INTRODUCTION

Generative models are increasingly used across science and industry, making model-selection metrics consequential: a misleading ranking can favor a worse model. Two widely used embedding-based scalar summaries are the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID). FID (Heusel et al., 2017) approximates each set's embedded feature distribution by the Gaussian determined by its empirical mean and covariance and reports the Fréchet distance between the two Gaussian approximations; KID (Bińkowski et al., 2018) reports the maximum mean discrepancy (MMD) associated with a degree-3 polynomial kernel on the same embeddings.

Because FID depends only on the first two moments of the embedding distribution, optimizing a noise-initialized batch of 64×64 images to match the mean and covariance of an ImageNet reference in Inception feature space (Deng et al., 2009; Russakovsky et al., 2015) yields visual noise to which FID assigns a lower, nominally better value than to held-out real images: 24.7 versus the 58.6 finite-sample real–real baseline (Inception-2048, n=512; Fig. 1 and Sec. D), demonstrating FID's matched-moment blind spot. Additional feature-space experiments corroborate the same blind spot. With full-rank Inception-2048 sample covariances (m=n=2500), a bimodal feature sample adjusted to closely match the reference mean and covariance yields FID 0.003, versus 19.5 for held-out real features. The same inversion appears across CIFAR-10 and Tiny-ImageNet[1], under both Inception and DINO features (Caron et al., 2021), using Gaussian alternatives fitted to the reference features projected onto their leading 128 principal components (PCA-128) (App. E.1).

[1]Tiny-ImageNet is the 200-class dataset distributed by Stanford CS231n.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

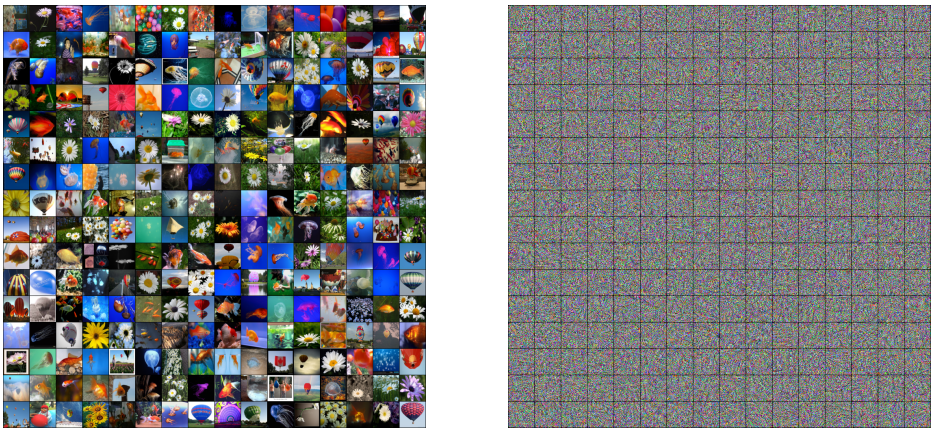


Figure 1: **Pixel-space matched-moment optimization reverses the FID ranking.** The montages show 256 images from each 512-image set; metrics use all 512. Left: held-out ImageNet images. Right: noise-initialized images optimized only to match the reference mean and covariance in Inception feature space. Despite being visual noise, the optimized set receives a lower (nominally better) FID (24.7 vs. 58.6). ZID assigns scores of 17.2 to the optimized comparison and 0.86 to held-out real. With 499 random relabelings, permutation tails are p=.002 for the optimized comparison (descriptive because it was optimized against the reference) and p=.328 for held-out real.

A second limitation appears when the departure has a direction. We sweep classifier-free guidance (CFG) (Ho & Salimans, 2022) in DiT-XL/2 (Peebles & Xie, 2023), a standard control of the fidelity–diversity tradeoff. Along this sweep, FID and KID are scalar discrepancies that are unchanged when the two samples are exchanged and therefore do not distinguish low-guidance off-manifold dispersion from high-guidance diversity collapse.

Finite-sample sensitivity provides an additional concern: at dimension $d = 100$ and per-sample size $n = 100$, KID power remains below .20 over the displayed Gaussian scale-shrinkage sweep, while FID reaches only .75 power for a 5% shrinkage even at $n = 1000$ (App. E.6).

ZID (*Z-resolved Integrated Diagnostic*) addresses these issues by combining complementary graph- and kernel-based statistics, calibrating their aggregated response by permutation, and retaining a signed dispersion contrast. The aggregated response provides an index of departure magnitude, and its permutation distribution yields an equality-test p-value. The signed contrast permits an under- or over-dispersion label when the direction is identifiable.

Contributions.

1. **Matched-moment analysis and calibrated detection.** FID's manipulability is known (Alfarra et al., 2022; Kynkäänniemi et al., 2023), and prior work has exhibited distinct distributions with matching first two moments and hence zero FID (Benny et al., 2021; Luzi et al., 2023; Jayasumana et al., 2024). Feature-space experiments show that FID can remain at its real–real baseline under matched-moment alternatives that ZID detects with its permutation-calibrated equality test. In the pixel-space stress test, optimizing actual images solely to match the reference Inception mean and covariance makes visual noise score better than held-out real images under FID, whereas ZID assigns it a much larger departure score.
2. **Detection coverage and diagnostic readouts.** ZID combines a rank-based graph test with Gaussian-kernel tests at median and fine-scale bandwidths to cover departures that favor different sensitivities. Across six departures spanning location, dispersion, linear and nonlinear dependence, a multimodal construction, and off-manifold support, it is the only method in the displayed comparison with power of approximately .70 or higher in all six columns ($d = 128, n = 200$; Fig. 3). Across a 54-cell dimension–sample-size grid it is most often among the most powerful methods. Beyond detection, ZID's signed readout distinguishes under- from over-dispersion when the direction is identifiable.

3. **Ranking along diverse severity sweeps.** Across the six feature-space departures, the ZID score is consistently positively associated with increasing severity, including settings where FID is flat or reversed. On DiT-XL/2 and SiT-XL/2 (Ma et al., 2024), it also tracks departure across the practically relevant guidance range, while the signed dispersion readout identifies the associated collapse direction.

ZID is intended for evaluation when the departure type is not known in advance. It jointly supports model ranking, calibrated equality testing, and, when identifiable, under- versus over-dispersion diagnosis. Our experiments use both Inception and DINO embeddings and include five pretrained generator families: BigGAN (Brock et al., 2019), StyleGAN2-ADA (Karras et al., 2020a;b), a CIFAR denoising diffusion probabilistic model (DDPM) (Ho et al., 2020), DiT-XL/2, and SiT-XL/2.

The paper proceeds as follows. Section 2 reviews related metrics. Section 3 defines the ZID score, permutation test, and signed readouts. Section 4 presents the controlled feature-space evaluation. Section 5 evaluates pretrained generators. The appendices provide proofs, methods and protocols, the pixel-space FID stress test, and expanded controlled and generator analyses.

2 Background and Related Metrics

Gaussian-kernel MMD provides a kernel two-sample discrepancy (Gretton et al., 2012), while CLIP Maximum Mean Discrepancy (CMMD) applies a Gaussian radial-basis-function (RBF) kernel to CLIP representations for generative-model evaluation (Jayasumana et al., 2024). FID variants address estimation and representation but retain the two-moment restriction. FID_∞ extrapolates finite-sample bias (Chong & Forsyth, 2020); R-FID replaces the standard Inception representation with a robust one (Alfarra et al., 2022); and Stein et al. (2023) recommend DINOv2 features for Fréchet distance. Kynkäänniemi et al. (2023) analyze FID's representation sensitivity. Compound FID instead aggregates Fréchet distances across multiple feature levels (Nunn et al., 2021). These variants remain unsigned summaries of feature means and covariances. Monge Inception Distance (MIND) uses sliced Wasserstein distances for sample efficiency and robustness to moment matching (Berthet et al., 2026), while the Embedded Characteristic Score (ECS) uses characteristic-function transforms to probe moments and tails (Tam & Engelhardt, 2026). The classifier two-sample test (C2ST) asks whether a held-out classifier distinguishes the samples (Lopez-Paz & Oquab, 2017).

Other metrics focus on diversity, coverage, or dispersion. Vendi Score summarizes sample diversity (Friedman & Dieng, 2023); Precision/Recall separates fidelity from coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019); and Density/Coverage refines their neighborhood-based estimates (Naeem et al., 2020). Permutational analysis of multivariate dispersions (PERMDISP) tests homogeneity of multivariate dispersion using distances to group spatial medians and a Levene-type statistic calibrated by residual permutations (Anderson, 2006).

3 The ZID Diagnostic: Score, Test, and Signed Readout

3.1 Candidate Families and a Common Two-Coordinate Structure

The generalized permutation-based kernel (GPK) test (Song & Chen, 2024), Rank In Similarity graph Edge-count (RISE) test (Zhou & Chen, 2023), and generalized edge-count test (GET) (Chen & Friedman, 2017) encode pairwise similarity differently but share a two-coordinate form. For reference sample X of size m, comparison sample Y of size n, pooled observations $z_1, \ldots, z_N$ with $N = m+n$, and pair weights w_{ij}, define the within-sample totals $U_x = \sum_{i<j} w_{ij}\mathbf{1}\{z_i \in X,\ z_j \in X\}$ and $U_y = \sum_{i<j} w_{ij}\mathbf{1}\{z_i \in Y,\ z_j \in Y\}$. Define the aligned and opposing coordinates and their null-standardized versions by

$$W = \frac{U_x}{m-1} + \frac{U_y}{n-1}, \quad D = U_x - U_y, \qquad Z_W = \frac{W - \mathbb{E}_0 W}{\sqrt{\mathrm{Var}_0(W)}}, \quad Z_D = \frac{D - \mathbb{E}_0 D}{\sqrt{\mathrm{Var}_0(D)}},$$

where $\mathbb{E}_0$ and Var_0 refer to the label-permutation null. For these constructions, $\mathrm{Cov}_0(Z_W, Z_D) = 0$. The corresponding quadratic statistic is $T = Z_W^2 + Z_D^2$. Z_W records aligned displacement of U_x and U_y, whereas Z_D records opposing displacement; exchanging X and Y leaves T unchanged but reverses the sign of Z_D. Full test-specific constructions and interpretation across location, scale, and mixed alternatives are given in App. A.1.

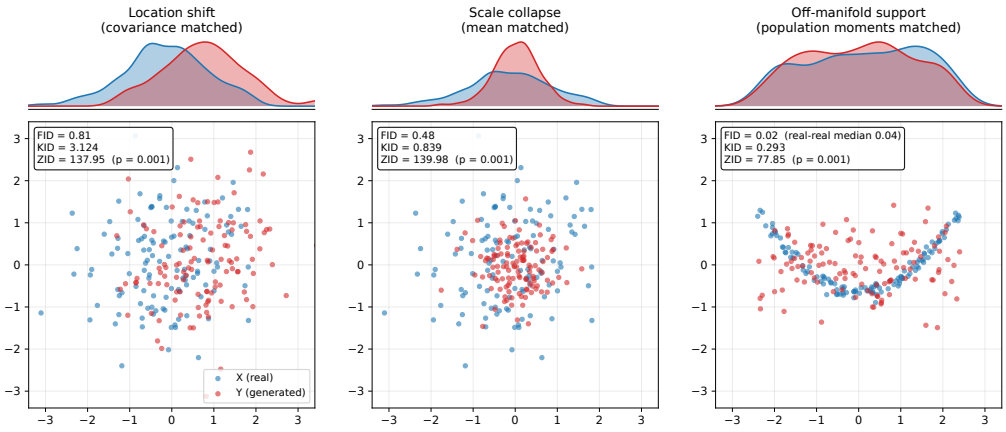


Figure 2: **Generator-relevant departures and score/test readouts.** Two-dimensional examples with $n = 120$ per sample (real in blue, generated in red). Boxes report FID, KID, and the ZID score and p-value. Metric magnitudes are on this illustrative scale and are not comparable to Inception-space values. Left: a location shift with matched covariance. Middle: mean-matched scale collapse illustrates an under-dispersed departure that the displayed scalar scores alone cannot label. Right: independent draws from populations with identical means and covariances but different support geometry; FID is below the median of 1,000 independent real–real comparisons, while ZID's calibrated test detects the departure.

3.2 Connections and contrasts with MMD, FID, and KID

For kernel weights, the usual unbiased empirical squared MMD is a positive multiple of the centered W coordinate:

$$\widehat{\mathrm{MMD}}_u^2 = \frac{2(N-1)}{mn}\big(W - \mathbb{E}_0 W\big).$$

It is therefore symmetric under sample exchange and has no counterpart to Z_D. When U_x is above its permutation-null mean and U_y below its own (or vice versa), their changes can cancel in $W - \mathbb{E}_0 W$ but add in $D - \mathbb{E}_0 D$. App. A.2 provides more details on this possible cancellation.

FID's dependence on only the feature mean and covariance creates a matched-moment blind spot: its population value is exactly zero whenever those moments match, even if the two feature distributions differ. The following elementary proposition records this known non-identifiability in a general form.

Proposition 1 (Blindness under matched means and covariances)**.** *Let $M(P, Q)$ be a discrepancy determined only by the means and covariances of P and Q, with $M(P, Q) = 0$ whenever these moments agree. For every distribution P on $\mathbb{R}^d$ with finite second moments and nonzero covariance, there exists $Q \neq P$ with the same mean and covariance. Hence $M(P, Q) = 0$; in particular,* $\mathrm{FID}(P, Q) = 0$.

(Proof in Appendix E.2: choose either the Gaussian with the same moments or, when P is already that Gaussian, an explicit non-Gaussian distribution with the same moments.)

Proposition 1 does not directly cover KID, whose degree-3 polynomial-kernel MMD vanishes when all mixed moments through total degree three match.

3.3 Score construction and design choices

ZID combines the rank-weighted RISE representation with Gaussian-kernel GPK at two bandwidths: the pooled median pairwise distance (GPK-med) and 0.175 times that distance (GPK-small). For each representation it keeps both $Z_W^{(j)}$ and $Z_D^{(j)}$, giving six arms, with $j = 1, 2, 3$ indexing RISE, GPK-med, and GPK-small, respectively. GET's binary graph construction remains a standalone comparator rather than an additional ZID member. The 0.175 multiplier was fixed before the reported comparisons; values from 0.10 to 0.25 yield similar minimum power across departures and mean null rejection rates (App. F). Member-removal ablations are reported in App. B.3.

The ZID score: an omnibus departure index. Each of the six arms $\{Z_W^{(j)}, Z_D^{(j)}\}_{j=1}^3$ is centered and scaled to have unit permutation-null variance. For arm k, let Φ denote the standard-normal cumulative distribution function and define the reference-scale transformation $r_k = 2[1 - \Phi(|Z_k|)]$, sort these as $r_{(1)} \le \cdots \le r_{(6)}$, and set

$$q_{\mathrm{rank}} = \min\left\{1, \min_{i=1,\ldots,6} \frac{6r_{(i)}}{i}\right\}, \qquad S_{\mathrm{ZID}} = -\log q_{\mathrm{rank}}.$$

The flat-Simes form (Simes, 1986) makes the score coordinatewise monotone in the observed arm magnitudes and unbounded. A single large arm cannot be canceled by the others, while several moderately large arms can increase the score through the $1/i$ factor in the ordered-tail minimum. App. A.3 gives its bounds and aggregation comparisons. Larger values indicate a more extreme standardized departure from the reference under the fixed evaluation protocol. Because the arms are dependent and the r_k are normal-reference transformations rather than finite-sample arm p-values, q_{rank} is used for ranking rather than test calibration; the outer permutation in Sec. 3.4 supplies the calibrated p-value.

3.4 Permutation test and diagnostic readouts

For a pooled sample, we recompute all six arms and S_{ZID} after each of B label permutations. The reported Monte Carlo permutation tail is denoted $\widehat{p}_{\mathrm{ZID}}$:

$$\widehat{p}_{\mathrm{ZID}} = \frac{1 + \sum_{b=1}^{B} \mathbf{1}\{S_{\mathrm{ZID}}^{(b)} \ge S_{\mathrm{ZID}}^{\mathrm{obs}}\}}{B+1}.$$

Here $S_{\mathrm{ZID}}^{\mathrm{obs}}$ is the score for the observed labeling and $S_{\mathrm{ZID}}^{(b)}$ is the score for the bth random relabeling. At nominal level α, the test rejects when $\widehat{p}_{\mathrm{ZID}} \le \alpha$. The plus-one Monte Carlo permutation p-value (Phipson & Smyth, 2010) includes the observed labeling alongside the B random relabelings, giving $B+1$ exchangeable configurations. Under the exchangeability null, recomputing the six-arm score for each relabeling yields a finite-sample-valid Monte Carlo permutation p-value without requiring the arms to be independent or Gaussian. "ZID test" and "ZID score" always denote the six-arm implementation; member-specific analyses are labeled RISE or GPK.

Component readouts. Applying flat-Simes separately to $\{Z_W^{(j)}\}_{j=1}^3$ and $\{Z_D^{(j)}\}_{j=1}^3$ gives nonnegative aggregated W- and D-component magnitudes, denoted S_W and S_D; let $\widehat{p}_W$ and $\widehat{p}_D$ denote their permutation tails. They locate the six-arm signal but are not separately error-controlled tests; only $\widehat{p}_{\mathrm{ZID}}$ determines rejection of distributional equality.

Dispersion diagnosis. Using $\widehat{p}_D$ as a prespecified diagnostic gate, define the descriptive reference-active set $\mathcal{A} = \{j : 2[1 - \Phi(|Z_D^{(j)}|)] \le \alpha\}$; membership is a descriptive filter, not an arm-level test. Then

$$\mathrm{diagnosis} = \begin{cases} \textit{under-dispersion}, & \widehat{p}_D \le \alpha,\ \mathcal{A} \ne \varnothing,\ Z_D^{(j)} < 0\ \forall j \in \mathcal{A}, \\ \textit{over-dispersion}, & \widehat{p}_D \le \alpha,\ \mathcal{A} \ne \varnothing,\ Z_D^{(j)} > 0\ \forall j \in \mathcal{A}, \\ \textit{member-sign conflict}, & \widehat{p}_D \le \alpha \text{ and signs within } \mathcal{A} \text{ conflict}, \\ \textit{ambiguous}, & \widehat{p}_D \le \alpha \text{ and } \mathcal{A} = \varnothing, \\ \textit{not assigned}, & \widehat{p}_D > \alpha. \end{cases}$$

The coherence-qualified signed readout is $D_{\mathrm{ZID}} = -S_D$ for *under-dispersion* and $D_{\mathrm{ZID}} = +S_D$ for *over-dispersion*; it is undefined for the other three diagnoses. The ungated score $\mathrm{sign}(\sum_j Z_D^{(j)})S_D$ is used only for continuous visualization. The threshold defining $\mathcal{A}$ filters descriptive signs rather than testing them separately. In the controlled scale, truncation, and guidance families studied here, coherent negative and positive signs track diversity contraction and expansion. Conflicting active signs are reported rather than collapsed. Here an identifiable direction means that the diagnostic gate passes and at least one reference-active member gives a coherent sign.

4 Feature-space evaluation under controlled departures

We compare ZID with other methods on detection power and severity-based ranking across six real-feature departures. We also study dispersion diagnosis, ablations, and cross-setting performance.

4.1 DETECTION AND DISPERSION DIAGNOSIS

Each repetition draws two disjoint samples from a fixed pool of real CIFAR-10 Inception features (Krizhevsky, 2009), projects them onto the leading 128 principal components, and applies the specified departure to the second sample. We use $m = n = 200$, 300 independent repetitions, $\alpha = .05$, and 499 permutations for permutation-calibrated methods, rejecting at $p \leq \alpha$ (Sec. 3.4). Each comparator's test statistic, calibration, and permutation tail are fixed across departures (Table 14). We examine six departures: location, dispersion, linear and nonlinear dependence, matched-moment multimodality, and off-manifold support. The multimodality construction targets the mixture's first two moments at the population scale; it does not enforce exact empirical mean and covariance matching in each repetition. These departures isolate feature differences that may coexist in a generator. Signals were selected before formal evaluation and fixed across methods; their departure-specific strengths are not numerically comparable across columns. Appendix E.3 gives the selection protocol, exact transformations and signal levels, and comparator definitions and calibration details. Real-feature null checks are reported in App. E.4.

Fig. 3 compares ZID with four standalone tests and fourteen external comparators. ZID leads nonlinear dependence (.96) and achieves .71 power on matched-moment multimodality, where RISE reaches .74; KID leads location (.89), PERMDISP dispersion (.87), FID linear dependence (.87), and Density off-manifold support (.93). ZID maintains substantial power across all six: it is the only row with power of approximately .70 or higher in every column, while the largest row minimum among the external comparators is Density's .08. This all-six coverage is specific to the Fig. 3 signals.

The standalone tests are complementary: RISE covers dispersion, dependence, and off-manifold support; GPK-med emphasizes location and dispersion; GPK-small emphasizes nonlinear dependence. Matched ablations show drops in six-departure mean power (.806 to .563) without RISE, location power (.830 to .237) without GPK-med, and nonlinear-dependence power (.980 to .860) without GPK-small (App. B.3).

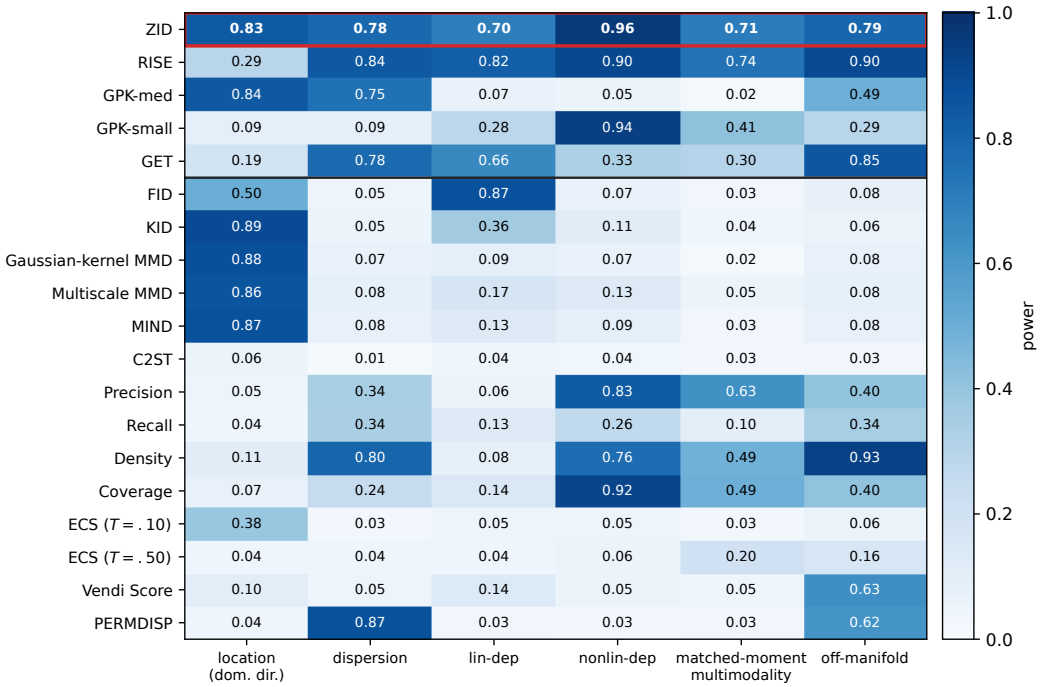


Figure 3: **Detection power across six controlled departures.** Empirical power at level .05 on real CIFAR-Inception PCA-128 features ($n = 200$ per sample; 300 repetitions; 499 permutations for permutation-calibrated tests). PERMDISP uses distances to group spatial medians and residual permutations. The boxed row is the complete six-arm ZID test; the next four rows are standalone RISE, GPK-med, GPK-small, and GET tests; and the remaining fourteen are external comparators.

Breadth across regimes and representations. At $d = 128, n = 200$, ZID's rejection rate increases monotonically across five-level severity ladders (App. G.1). Separately, we vary feature dimension $d \in \{128, 512, 2048\}$ and per-sample size $n \in \{50, 200, 600\}$ across the six departures. ZID's power is at least 80% of the highest power among ZID and fourteen external comparators in 44 of the resulting 54 settings. FID and KID meet the same threshold in 12 and 9 settings, respectively; Density has the highest count among external comparators at 20. Signals vary by cell, so comparisons are within settings rather than power trajectories over d or n. Patterns persist across CIFAR-10 and Tiny-ImageNet and across Inception and DINO. Supporting ablations, sensitivity and calibration checks, and the full grid appear in Apps. B.3, F, E.4, and G.2.

Directional diagnosis. Detection power does not assess whether a metric identifies the direction of a scale change. We therefore apply isotropic contractions and expansions about the feature mean to full-dimensional Inception features and compare the resulting magnitudes and signed readout (Fig. 4). In this full-rank dispersion sweep, the ungated net signed dispersion score changes sign across the reference scale. The formal readout D_{ZID} is negative for the displayed contractions and positive for the displayed expansions; at c=1, it is undefined because no direction is assigned. FID remains unsigned and can assign similar magnitudes to deviations on opposite sides of the reference. Repeated checks assign the correct direction at every tested nonnull scale in both feature settings; App. B.1 reports the assignment counts, null behavior, and runtime.

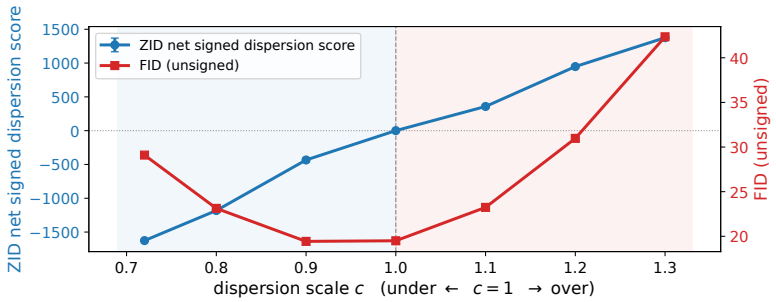


Figure 4: **Signed vs. unsigned dispersion** (real raw Inception-2048, disjoint m=n=2500>d, full-rank covariance; Y=μ+c(base−μ)). The blue curve shows the ungated net signed dispersion score, $\mathrm{sign}(\sum_j Z_D^{(j)})S_D$. For the displayed nonnull points, the formal readout D_{ZID} equals this score; at $c = 1$, the diagnosis is not assigned and D_{ZID} is undefined (Sec. 3.4). FID is unsigned and is nearly equal at c=0.8 and 1.1. FID's displayed minimum, 19.43 at c=0.9 versus 19.50 at c=1, slightly favors mild collapse in this finite sample. Points are means over eight independent repetitions.

4.2 Ranking across increasing severity

Detection at one selected signal level does not establish that a score orders progressively stronger departures. We therefore evaluate 100 independent six-level panels for each departure on CIFAR–Inception features with $m = n = 200$ (Fig. 5). Within a panel, all levels share one real reference sample, while their candidate samples are generated independently. This design targets finite-sample ranking among candidate sets evaluated against a common reference. We report mean panel-wise Spearman ρ: for each panel, we correlate a method's six raw scores with the ordered severity levels, then average the correlations over the 100 panels. Within each panel, we compare all $\binom{6}{2} = 15$ pairs of severity levels. A pair is correct when the higher-severity candidate receives the larger score, incorrect when it receives the smaller score, and receives half credit when tied; all-pairs ordering accuracy is the average over these 15 comparisons.

FID and KID rank location, linear dependence, and off-manifold support well. FID instead reverses the dispersion-collapse ordering ($\rho = -.48$), while KID is only modest on dispersion and nonlinear dependence ($.26/.28$); both are essentially flat on matched-moment multimodality ($-.01/.00$). ZID remains positive across all six ($\rho = .65$–$.98$), with its clearest ranking gains on dispersion, nonlinear dependence, and multimodality.

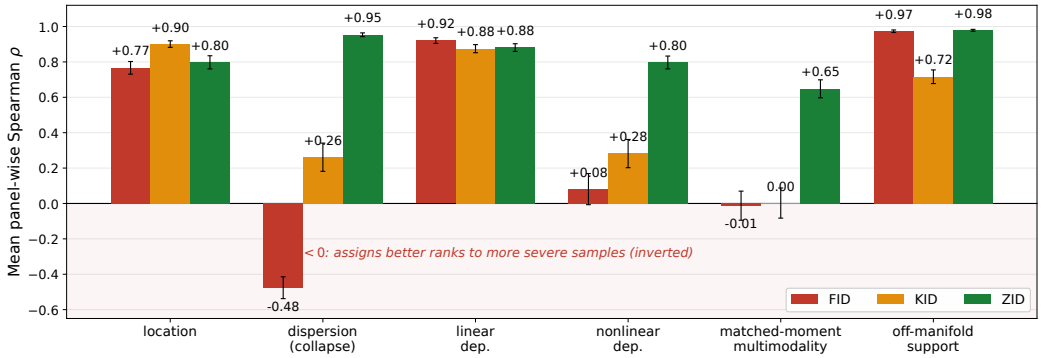


Figure 5: **Ranking across sweeps of increasing severity.** Mean panel-wise Spearman ρ of each raw score against six ordered severity levels. Bars show 95% percentile intervals from 10,000 bootstrap replicates that resample the 100 panels, retaining each panel's six levels. Within each panel, every method uses the same real reference and each severity level has an independently generated candidate sample. Negative ρ indicates inverted ordering.

Across the six departures, ZID's all-pairs ordering accuracy ranges from 76.5% to 97.5%, versus 31.2%–97.1% for FID and 50.5%–90.7% for KID. Against the expanded external-comparator panel, ZID leads the mean and minimum of both ranking metrics; standalone RISE remains close on the averages (App. B.2). Thus the results demonstrate broad ranking coverage.

5 Evaluation of Pretrained Generative Models

We evaluate five pretrained model families spanning GANs, diffusion, and interpolant transformers. We begin with DiT and SiT guidance sweeps, which show how departure magnitude and dispersion direction change along a generator path.

DiT-XL/2: one guidance sweep, two failure directions. We vary classifier-free guidance (CFG) on **DiT-XL/2-256** (Peebles & Xie, 2023) for six ImageNet classes (n=500), producing a trajectory from diffuse, low-precision samples to high-guidance diversity collapse. At $\mathrm{CFG} \in \{1, 2, 4, 8\}$ (Ho & Salimans, 2022), we compare each sample set with matched real images: Higher Precision and Recall indicate better fidelity and diversity, respectively, whereas smaller FID, KID, and ZID scores indicate less departure within this fixed sweep. At low guidance, low Precision suggests fewer generated features lie within the estimated real-data support. Increasing guidance improves fidelity but collapses diversity. In the class-aggregated curves in Fig. 6(a), FID, KID, and the ZID score attain their minima at CFG=2. These scalar summaries identify the same aggregate closest setting, but do not indicate whether larger departures reflect off-manifold over-dispersion or diversity collapse. Precision and Recall must be read together: at CFG=8, Precision can remain as high as 0.89 while Recall falls to 0.025; at CFG=1, mean Recall is 0.75 despite mean Precision of only 0.57. Recall alone can therefore make the low-guidance samples appear strong despite their low Precision, just as Precision alone can make the collapsed high-guidance samples appear strong.

As CFG increases from 2 to 8, ZID's W-component magnitude S_W rises through CFG 8 in all six classes. The final aggregate dispersion readout is over-dispersion at CFG$=1$ and under-dispersion at CFG$\in \{2, 4, 8\}$ for every class; the diagnostic readout D_{ZID} therefore changes from positive at low guidance to negative as guidance rises. Figure 6 shows the two-sided score pattern of Fig. 4 on a pretrained generator, with the signed readout separating the two directions. Per-class results are in App. C.3.

The same guidance transition appears on SiT-XL/2: the signed readout indicates over-dispersion at $\mathrm{CFG} = 1$ and under-dispersion at $\mathrm{CFG} \in \{4, 8, 16\}$ across all six classes. Beyond the DiT and SiT guidance sweeps, we examine truncation in BigGAN and StyleGAN2-ADA and denoising diffusion implicit model (DDIM) sampling steps in a CIFAR DDPM. Table 1 summarizes these five evaluations and their main findings; detailed protocols and numerical results are in App. C.

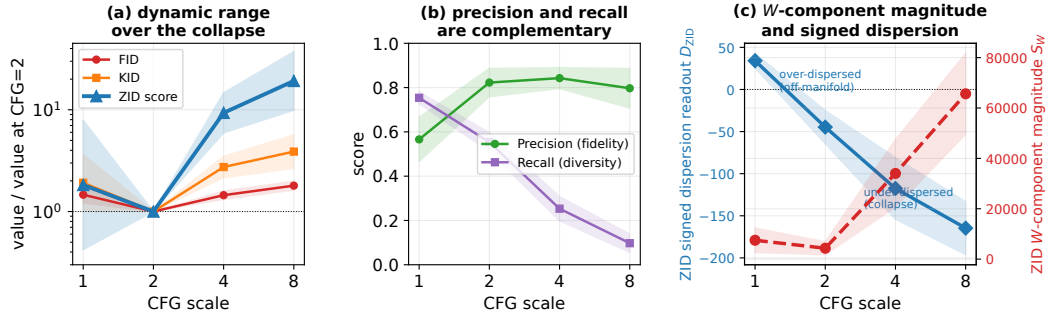


Figure 6: **A guidance sweep from over- to under-dispersion.** Across six DiT-XL/2 ImageNet classes and CFG $\in \{1, 2, 4, 8\}$: (a) thick lines are geometric means across the six class-specific ratios to their CFG$= 2$ values, and shading spans one standard deviation on the log-ratio scale; (b) Precision can remain high while Recall collapses; and (c) ZID's nonnegative W-component magnitude S_W tracks departure magnitude along this sweep, while its diagnostic readout D_{ZID} flips from over-dispersed (> 0) at low guidance to under-dispersed (< 0) at high guidance. At low guidance, low Precision suggests fewer generated features lie within the estimated real-data support.

Table 1: **Evaluations of pretrained generative models.** Each row varies only the listed sampling or truncation parameter; the pretrained generator and evaluation setup are otherwise fixed. StyleGAN2-ADA uses an independent untruncated generator sample as its reference; the other sweeps use real-image references. The final column summarizes the principal finding.

generator	domain / embedding	varied parameter / scale	main finding
BigGAN-deep-128	ImageNet / Inception	truncation, 10 classes, $n = 450$	ZID/KID scores rise as truncation tightens; ZID reports under-dispersion
CIFAR DDPM	CIFAR-10 / Inception	DDIM steps, $n = 200$	more steps reduce FID/ZID; all settings still reject equality
DiT-XL/2	ImageNet / Inception	CFG, 6 classes, $n = 500$	guidance shifts over-dispersion to under-dispersion
SiT-XL/2	ImageNet / Inception	CFG, 6 classes, $n = 500$	the signed guidance transition recurs in a second architecture
StyleGAN2-ADA	FFHQ / Inception	truncation, $n = 5000$	ZID score rises; under-dispersion strengthens with tighter truncation

6 Conclusion

Across six real-feature departure families, ZID maintains broad detection power, and its score tracks increasing severity across panels of independently generated candidates in all six. Matched-moment feature experiments and a reference-adaptive pixel-space stress test show that favorable FID values can coexist with substantial departures from real data.

In pretrained-generator evaluations, the ZID score orders truncation and sampling-step settings and tracks departure on both sides of the closest sampled guidance setting. The signed readout additionally identifies low-guidance over-dispersion and high-guidance under-dispersion in both DiT and SiT. Together, the score, permutation p-value, and supported dispersion label answer distinct questions: how candidates order under a common evaluation protocol, whether equality is rejected, and which direction net dispersion changes when that direction is identifiable.

Scope and limitations. ZID assesses departure in a chosen feature representation, not perceptual quality; it is not a dedicated memorization test. Its score is not a universal effect size, so ranking comparisons require a common reference, embedding, preprocessing, and sample size.

AI USE STATEMENT

We used generative AI tools to provide feedback on research methodology, assist with the construction of controlled synthetic evaluation settings and with software implementation and debugging, interpret and summarize numerical results, draft and edit manuscript text, review related literature and references, and prepare figures and other research artifacts. All AI-assisted code, experiments, citations, numerical results, and prose were checked by us. We take responsibility for the final content of this work, including all text, code, and artifacts produced with AI assistance.

REPRODUCIBILITY STATEMENT

The statistical construction, calibration procedure, controlled alternatives, sample sizes, permutation budgets, and generator-evaluation protocols are specified in the main paper and appendices. The supplementary material includes an anonymized implementation of the complete six-arm ZID test and diagnostic readouts, together with statistical and API tests. It also provides the per-repetition results and scripts that regenerate Figs. 3 and 11, and the panel-level ranking scores with the script that reproduces the ranking summaries.

REFERENCES

Motasem Alfarra, Juan C. Pérez, Anna Frühstück, Philip H. S. Torr, Peter Wonka, and Bernard Ghanem. On the robustness of quality measures for GANs. In *European Conference on Computer Vision (ECCV)*, pp. 18–33, 2022. doi: 10.1007/978-3-031-19790-1_2.

Marti J. Anderson. Distance-based tests for homogeneity of multivariate dispersions. *Biometrics*, 62 (1):245–253, 2006. doi: 10.1111/j.1541-0420.2005.00440.x.

Yaniv Benny, Tomer Galanti, Sagie Benaim, and Lior Wolf. Evaluation metrics for conditional image generation. *International Journal of Computer Vision*, 129:1712–1731, 2021. doi: 10.1007/s11263-020-01424-w.

Quentin Berthet, Yu-Han Wu, Clement Crepy, Romuald Elie, Klaus Greff, and Michael Eli Sander. MIND: Monge Inception Distance for generative models evaluation. *arXiv preprint arXiv:2605.06797*, 2026.

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021.

Hao Chen and Jerome H Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American Statistical Association*, 112(517):397–409, 2017. doi: 10.1080/01621459.2016.1147356.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception Score and where to find them. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6070–6079, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255, 2009.

Ronald A. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 4th edition, 1932.

Dan Friedman and Adji Bousso Dieng. The Vendi Score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pp. 6626–6637, 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6840–6851, 2020.

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

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410, 2019.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 12104–12114, 2020a.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119, 2020b.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, pp. 3929–3938, 2019.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of ImageNet classes in Fréchet Inception Distance. In *International Conference on Learning Representations (ICLR)*, 2023.

Yaowu Liu and Jun Xie. Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529): 393–402, 2020. doi: 10.1080/01621459.2018.1554485.

David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests. In *International Conference on Learning Representations (ICLR)*, 2017.

Lorenzo Luzi, Carlos Ortiz Marrero, Nile Wynar, Richard G. Baraniuk, and Michael J. Henry. Evaluating generative networks using gaussian mixtures of image features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 279–288, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision (ECCV)*, pp. 23–40, 2024.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, 2020.

Eric J. Nunn, Pejman Khadivi, and Shadrokh Samavi. Compound Fréchet Inception Distance for quality assessment of GAN created images. *arXiv preprint arXiv:2106.08575*, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Belinda Phipson and Gordon K. Smyth. Permutation p-values should never be zero: Calculating exact p-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lučić, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 5234–5243, 2018.

R. John Simes. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73 (3):751–754, 1986. doi: 10.1093/biomet/73.3.751.

Hoseung Song and Hao Chen. Generalized kernel two-sample tests. *Biometrika*, 111(3):755–770, 2024. doi: 10.1093/biomet/asad068.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Edric Tam and Barbara E. Engelhardt. Assessing the distributional fidelity of Synthetic Chest X-rays using the Embedded Characteristic Score. *arXiv preprint arXiv:2501.00744v2*, 2026.

Doudou Zhou and Hao Chen. A new ranking scheme for modern data and its application to two-sample hypothesis testing. In *Proceedings of the 36th Conference on Learning Theory (COLT)*, volume 195 of *Proceedings of Machine Learning Research*, pp. 3615–3668, 2023.

A Additional background and method details

A.1 Candidate tests and two-coordinate details

The standalone GPK, RISE, and GET tests use different pairwise representations, but each has a Mahalanobis form. With $U = (U_x, U_y)^\top$ from Sec. 3.1, let μ and Σ be its mean and covariance under the label-permutation null. The joint statistic is

$$T = (U - \mu)^\top \Sigma^{-1} (U - \mu).$$

GPK uses dense Gaussian-kernel weights, GET a sparse binary similarity graph, and RISE a rank-weighted similarity graph. For GET, W is proportional to the weighted edge-count statistic; for RISE, it is the corresponding weighted combination of the two within-sample rank totals. For GPK, $W - \mathbb{E}_0 W$ is proportional to unbiased empirical squared MMD (App. A.2).

The coordinate distinction is not exclusive: at low dimension, scale alternatives can also contribute to Z_W, and mixed location–scale departures may affect both coordinates.

A.2 KERNEL/MMD DECOMPOSITION

Set $w_{ij} = k(z_i, z_j)$ and define $U_{xy} = \sum_{i<j} w_{ij}\mathbf{1}\{z_i \in X,\ z_j \in Y \text{ or } z_i \in Y,\ z_j \in X\}$. Let $\bar{k}$ be the fixed pooled average kernel weight over the $N(N-1)$ ordered pairs of distinct observations:

$$\bar{k} = \frac{\sum_{i \neq j} k(z_i, z_j)}{N(N-1)}.$$

Because $\mathbb{E}_0(W) = N\bar{k}/2$, the usual unbiased empirical squared MMD satisfies

$$\widehat{\mathrm{MMD}}_u^2 = \frac{2U_x}{m(m-1)} + \frac{2U_y}{n(n-1)} - \frac{2U_{xy}}{mn} = \frac{2(N-1)}{mn}\big(W - \mathbb{E}_0 W\big).$$

Thus, $\widehat{\mathrm{MMD}}_u^2$ and $W - \mathbb{E}_0 W$ differ only by the positive factor $2(N-1)/(mn)$. MMD is symmetric under swapping X and Y and supplies no signed-difference coordinate corresponding to Z_D. KID uses the same MMD statistic with a degree-three polynomial kernel. Z_D instead contrasts U_x with U_y; its sign supplies the orientation used by the under- versus over-dispersion readout (Figs. 2 and 4).

The opposite-direction scale pattern is the mechanism exploited by generalized graph and kernel tests (Chen & Friedman, 2017; Song & Chen, 2024). Under the scale alternatives, increasing dimension can concentrate pairwise similarities and place the two within-sample similarities on opposite sides of their pooled permutation reference. When these centered shifts have opposite signs, their contributions largely cancel in $W - \mathbb{E}_0 W$ but add in $D - \mathbb{E}_0 D$. The same pattern can occur when location and scale both change: opposing centered shifts can contribute to Z_D, while aligned shifts simultaneously produce signal in Z_W.

A.3 FLAT-SIMES PROPERTIES

Because different departures can concentrate signal in different arms, we use flat-Simes to preserve a strong signal in any one arm without allowing another arm to cancel it. The map $-\log\{2[1-\Phi(|Z|)]\}$ puts unit-variance arms on a common, unbounded reference scale, assigns zero severity at $Z = 0$, and satisfies $-\log\{2[1-\Phi(|Z|)]\} = Z^2/2 + \log|Z| + O(1)$ as $|Z| \to \infty$. Flat-Simes is symmetric in the arms, coordinatewise monotone, and sparse-adaptive: one large arm can drive the score, while several moderate arms reduce its multiplicity penalty. In particular, writing $s_k = -\log\{2[1-\Phi(|Z_k|)]\}$ and $s_{\max} = \max_k s_k$,

$$s_{\max} - \log 6 \ \le\ S_{\mathrm{ZID}} \ \le\ s_{\max}, \tag{1}$$

and if all six severities are equal to s, then $S_{\mathrm{ZID}} = s$. Thus no arm can be hidden by cancellation. Appendix F compares flat-Simes with max and sum-of-squares aggregation, Fisher's combination rule (Fisher, 1932), and the Cauchy combination rule (Liu & Xie, 2020). Flat-Simes is among the leading rules for both detection and ranking across increasing severity.

B ADDITIONAL FEATURE-SPACE EVALUATIONS

B.1 REPEATED SCALE DIAGNOSIS AND RUNTIME

We repeated the isotropic scale construction of Fig. 4 at $c \in \{.72, .80, .90, 1, 1.10, 1.20, 1.30\}$ using independent reference/base pairs in each repetition. The same pair was evaluated at all seven scales within a repetition. Both settings used ZID with 499 outer relabelings and $\alpha = .05$. The PCA-128 setting used $m = n = 200$ and 300 repetitions; the full Inception-2048 setting used $m = n = 2500$ and 100 repetitions. Table 2 reports correct direction counts across the three contractions and three expansions and, at $c = 1$, counts of direction assignments and equality-test rejections.

At the tested nonnull levels there were no opposite-direction, ambiguous, conflict, or unassigned outcomes. At $c = 1$, a direction was assigned without overall equality rejection in 11/300 (3.7%) PCA-128 repetitions and 1/100 (1%) full-dimensional repetitions. Such assignments are descriptive, not separately error-controlled; the direction readout has its own diagnostic gate. These observations do not establish directional accuracy for weaker or mixed departures.

We timed two public-API calls separately: computing the observed ZID score without outer permutations, and running the complete test with 499 random relabelings to obtain its p-value. With

Table 2: **Direction assignments and equality-test rejections in the repeated scale experiment.**

setting	contraction correct	expansion correct	null assigned	null equality rejected
PCA-128, $m = n = 200$	900/900	900/900	19/300	12/300
Inception-2048, $m = n = 2500$	300/300	300/300	3/100	3/100

one process and one BLAS thread, their median per-call times were .028 and .071 s for PCA-128 ($m = n = 200$), and 13.75 and 16.85 s for Inception-2048 ($m = n = 2500$), respectively. Each median uses three null and three contraction sample pairs; both calls include distance construction and package overhead.

B.2 Ranking across increasing severity

We use the common-reference, independent-candidate panels defined in Sec. 4.2 (100 six-level panels per departure; CIFAR–Inception PCA-128, $m = n = 200$). For each method, a constant-score panel receives Spearman $\rho = 0$, and tied score pairs receive half credit in all-pairs concordance. Bootstrap resampling keeps each panel's levels and the methods' pairing intact. The transformations and complete sweeps are specified in App. E.3. For the off-manifold ranking panels, each candidate's lower half of coordinates by empirical standard deviation is perturbed; the fixed 64-coordinate construction in App. E.3 specifies the detection comparison.

Figure 5 shows that FID and KID track increasing severity on location, linear dependence, and off-manifold support. FID instead reverses the dispersion-collapse ordering ($\rho = -.48$) and is essentially flat on matched-moment multimodality; KID is likewise flat on the multimodality sweep. The ZID score maintains positive mean panel-wise rank association on every sweep, ranging from .65 on matched-moment multimodality to .98 on off-manifold support. No method produces a constant-score panel.

Table 3 separates the individual sweeps. FID or KID leads on location and linear dependence, whereas ZID has the highest all-pairs concordance on dispersion collapse, nonlinear dependence, matched-moment multimodality, and off-manifold support. ZID does not lead on every departure, but maintains substantial ordering accuracy across all six.

Table 3: **Panel-wise ordering of graded departures.** Entries are all-pairs concordance accuracies (percent) across 100 independent six-level panels; ties receive half credit and chance accuracy is 50%. Each panel shares one real reference across levels and draws the six candidate samples independently. Bold denotes the row best.

	FID	KID	ZID
location	82.6	**90.7**	84.5
dispersion (collapse)	31.2	59.7	**95.1**
linear dependence	**92.9**	90.0	89.9
nonlinear dependence	53.3	61.3	**84.7**
matched-mom. multimodality	49.7	50.5	**76.5**
off-manifold support	97.1	80.5	**97.5**

To broaden the comparison beyond FID and KID, we apply the same panel-ranking protocol to these four standalone tests and four additional external comparators: multiscale MMD, MIND, and ECS at $T = .10$ and $T = .50$ (App. F). Among the six external comparators, ZID has the highest mean and minimum Spearman correlations across the six departures, as well as the highest mean and minimum all-pairs concordance. Among the standalone tests, RISE remains close on mean correlation and concordance, whereas ZID has higher mean and minimum values for both summaries.

B.3 Construction ablations and sensitivity

To examine the roles of RISE, GPK-med, and GPK-small and the two coordinates within ZID, Table 4 compares the full six-arm construction with variants that omit RISE, GPK-med, or GPK-small and

two coordinate-only variants. Table 5 isolates the contribution of GPK-small by comparing the full six-arm ZID with its variant without GPK-small across four severity levels. Table 6 studies a moment-restored resampling-with-perturbation construction. Starting from two disjoint real samples R and H, with R as the reference, the construction replaces a fraction f of the observations by perturbed resamples from H and then restores the sample mean and covariance of H (App. E.3). The table examines how the full six-arm ZID, RISE, GPK-med, and GPK-small, together with FID and KID, order increasing values of f across perturbation scales. This is a pooled association across

Table 4: **Paired construction ablation at the Fig. 3 signal levels.** Results use real CIFAR-Inception features at $d = 128, n = 200, \alpha = .05$, with 300 repetitions and 499 outer permutations. Rejection uses $p \leq .05$. Within each repetition, every specification uses the same samples, six computed arms, and label permutations. Among the rows omitting RISE, GPK-med, or GPK-small, bold marks the largest power loss from the full six-arm ZID for each departure.

specification	location	dispersion	lin.-dep.	nonlin.-dep.	multimodal	off-manifold	mean
full six-arm ZID	.830	.750	.727	.980	.717	.833	.806
drop RISE	.843	**.630**	**.167**	.893	**.317**	**.527**	.563
drop GPK-med	**.237**	.760	.800	.983	.767	.870	.736
drop GPK-small	.840	.800	.740	**.860**	.697	.853	.798
Z_W arms only	.850	.073	.793	.983	.777	.100	.596
Z_D arms only	.063	.837	.127	.467	.130	.910	.422

Table 5: **Paired contribution of GPK-small across severity levels.** Entries are paired rejection-rate differences, $\Delta = \mathrm{Power}(\text{full six-arm ZID}) - \mathrm{Power}(\text{ZID without GPK-small})$. Positive values favor including GPK-small. Five departures use perturbation multipliers $0.50, 0.75, 1.00, 1.25$; matched-moment multimodality uses replacement fractions $.25, .50, .75, 1.00$. All cells use 300 paired repetitions, 499 outer permutations, and rejection at $p \leq .05$. **Bold** marks $|\Delta| \geq .05$.

departure	setting 1	setting 2	setting 3	setting 4
common multiplier	0.50	0.75	1.00	1.25
location	$-.013$	$-.027$	$-.013$	$+.003$
dispersion	$-.040$	$\mathbf{-.060}$	$-.040$	$-.017$
linear dependence	$+.010$	$-.023$	$-.017$	.000
nonlinear dependence	$\mathbf{+.343}$	$\mathbf{+.193}$	$\mathbf{+.077}$	$+.013$
off-manifold support	$+.017$	$-.020$	$-.027$	.000
multimodality replacement fraction	.25	.50	.75	1.00
matched-moment multimodality	.000	$+.023$	$+.033$	$+.030$

Table 6: **Moment-restored resampling with perturbation.** Spearman ρ between each score and the replacement fraction f (real CIFAR-Inception, d=128, n=200), evaluated across perturbation scales ε (fractions of the per-coordinate standard deviations). Five perturbation scales $\varepsilon \in \{.05, .10, .20, .30, .50\}$ and six replacement fractions $f \in \{0, .1, .2, .3, .4, .5\}$ are evaluated with 100 repetitions at each (ε, f) setting. For each ε, the correlation pools the 600 independently sampled comparisons across all six f levels; it is distinct from the mean within-panel correlation in Fig. 5. Within each repetition, every method is evaluated on the same reference sample R and transformed sample B. Their standalone columns use $T = Z_W^2 + Z_D^2$. **Bold** marks the largest correlation in each row.

ε	**ZID**	FID	KID	RISE	GPK-med	GPK-small
0.05	$+.97$	$-.10$	$-.13$	$+.74$	$-.10$	$\mathbf{+.98}$
0.10	$+.96$	$-.04$	$+.02$	$+.71$	$+.00$	$\mathbf{+.97}$
0.20	$+.96$	$+.04$	$-.01$	$+.66$	$-.07$	$\mathbf{+.97}$
0.30	$+.93$	$+.02$	$+.06$	$+.55$	$+.05$	$\mathbf{+.95}$
0.50	$\mathbf{+.86}$	$+.06$	$+.04$	$+.23$	$-.07$	$+.84$

independently sampled settings, distinct from the panel-wise ranking measure in Fig. 5. Finally, we check sensitivity to the Gaussian-reference tail probabilities by replacing them with permutation estimates (App. E.5).

Table 4 shows complementary roles within the full six-arm ZID. Omitting RISE produces the largest power loss on four of the six departures; omitting GPK-med reduces location power from .830 to .237; and omitting GPK-small has its clearest effect on nonlinear dependence, reducing power from .980 to .860. Among the coordinate-only variants, the Z_W-only variant is stronger on location and dependence, whereas the Z_D-only variant is stronger on dispersion and off-manifold support.

GPK-small targets unusually close-pair structure, complementing the rank-based and median-bandwidth members. Table 5 shows that GPK-small contributes most clearly on nonlinear dependence: its incremental power contribution grows from $+.013$ at multiplier 1.25 to $+.343$ at $.50$, with the largest gain at the lower-signal end of the tested range. Its effects on the other five departures are smaller and occasionally negative. Table 6 provides a second view of the fine-scale role: under a moment-restored resampling-with-perturbation construction, the full six-arm ZID score remains strongly associated with the replacement fraction in the pooled settings across perturbation scales ($\rho = .86$–$.97$). GPK-small supplies the clearest member-level signal (.84–.98), while RISE weakens from .74 to .23 as the perturbation scale increases; FID, KID, and GPK-med remain near zero.

Gaussian-reference versus permutation arm tails. For flat-Simes aggregation, ZID assigns each standardized arm the two-sided Gaussian-reference tail probability $2[1-\Phi(|Z_k|)]$. To check sensitivity to this choice, we compare these probabilities with estimates obtained from inner label permutations. Both resulting scores use the same outer calibration and therefore control Type-I error. They give qualitatively similar power profiles, although the Gaussian-reference construction is more powerful in several departures, by up to about .10 (App. E.5). Because permutation estimation requires an additional inner loop, we use the computationally simpler Gaussian-reference formula.

B.4 Sensitivity across severity and (d, n) regimes

We evaluate the full six-arm ZID over five ordered severity levels with $d = 128$, $n = 200$, 300 repetitions, and 499 outer permutations per cell. Five departures use multiples $\{0, .5, .75, 1, 1.25\}$ of the Fig. 3 signal. Matched-moment multimodality instead varies the replacement fraction $\lambda \in \{0, .25, .5, .75, 1\}$ from the unmodified real sample to the full bimodal alternative. Across each five-level ladder, the empirical rejection rate increases monotonically with the perturbation multiplier (App. G.1, Fig. 10).

To compare relative sensitivity across several dimension–sample-size regimes, we repeat the six-departure comparison at $d \in \{128, 512, 2048\}$ and $n \in \{50, 200, 600\}$ (Fig. 11). Signals are selected separately within each (d, n) cell, so comparisons are within cells rather than power trends over d or n. Method definitions and calibration procedures follow Fig. 3. Rows requiring permutation calibration use 99 permutations per repetition. We define a cell-wise leading group among ZID and the fourteen external comparators using a threshold of 80% of the highest power within that set. Across the 54 cells, ZID enters this group 44 times, followed by Density (20), PERMDISP (17), and FID (12); multiscale MMD enters nine times, and each other external comparator enters it at most nine times. When assessed against the same threshold, the four standalone rows reach it in 38 cells for RISE, 28 for GET, 18 for GPK-med, and 12 for GPK-small.

B.5 Additional dataset and representation comparisons

The preceding evaluations use CIFAR-10 Inception features. We next examine whether their response patterns persist under a different dataset or feature representation. Table 7 evaluates variance collapse, over-dispersion, a leading-PC location shift, and class mode-drop on CIFAR-10 and Tiny-ImageNet (200 classes) in Inception features[2]. The two datasets show closely matched response profiles: ZID has high power on collapse and location in both datasets, Density leads over-dispersion, and FID leads mode-drop; ZID power remains .62–.97 across all eight comparisons.

[2]The Inception representation is identified in App. E.3. Absolute FID/KID magnitudes are on an implementation-specific scale; the analysis uses their relative rankings, and Fig. 7 reports the corresponding DINO evaluation.

Table 7: **Cross-dataset comparison across four perturbations.** Entries report empirical power at the 0.05 significance level on CIFAR-10 and Tiny-ImageNet (200 classes), with dataset-specific PCA-128 maps fitted to 5,000 Inception features, $m = n = 200$, 300 repetitions, and 499 relabelings. For mode-drop, CIFAR-10 retains 80% of its classes and Tiny-ImageNet retains 50%; the other three perturbations use the same displayed strengths in both datasets. Bold marks row maxima.

dataset	alternative	ZID	FID	KID	MIND	Density	Recall	Coverage	C2ST
CIFAR-10	collapse ($s = .05$)	**.97**	.08	.06	.18	.96	.68	.18	.03
	over-disp. ($\tau = .25$)	.68	.07	.07	.05	**.86**	.26	.39	.05
	location ($\delta = .40$)	.91	.67	**.97**	.93	.15	.05	.10	.11
	mode-drop (80% retained; 2 classes dropped)	.85	**.93**	.88	.83	.06	.10	.86	.15
Tiny-IN	collapse ($s = .05$)	.89	.08	.05	.18	**.92**	.62	.15	.05
	over-disp. ($\tau = .25$)	.62	.07	.05	.08	**.80**	.27	.37	.05
	location ($\delta = .40$)	**.94**	.54	.93	.88	.25	.09	.05	.06
	mode-drop (50% retained; 100 classes dropped)	.80	**.94**	.82	.65	.05	.28	.33	.30

Figure 7 traces power across perturbation levels for the same four departures on CIFAR-10 using Inception-2048 and DINO ViT-S/16 CLS embeddings (Caron et al., 2021). The DINO representation is identified in App. E.3. Across the full curves, ZID responds to all four departures in both representations: power rises rapidly with variance collapse, over-dispersion, and class omission, and increases smoothly with the leading-PC shift. At selected levels, Density leads over-dispersion, KID leads location shift, and FID leads mode-drop.

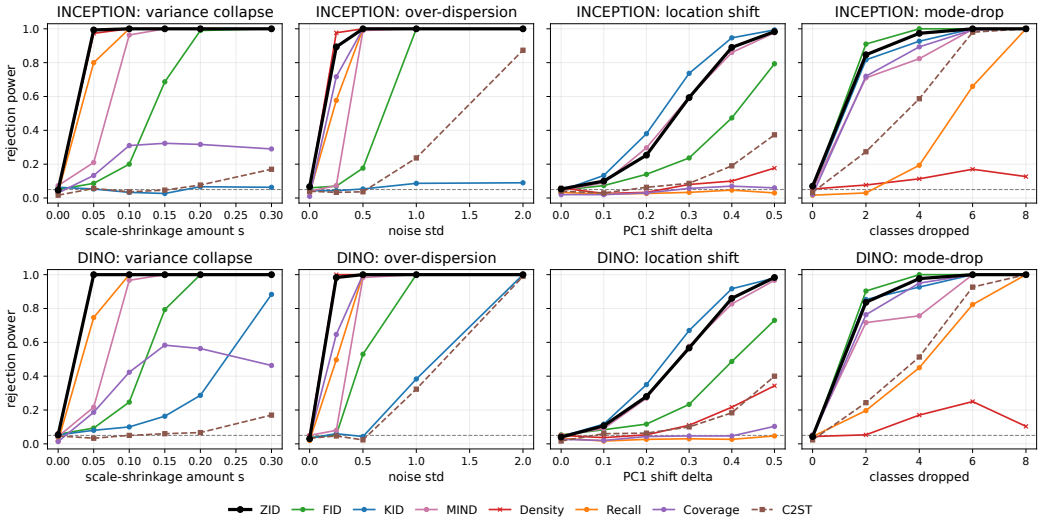


Figure 7: **Detection power across perturbation levels on CIFAR-10.** Rows use Inception-2048 and DINO ViT-S/16 CLS embeddings; columns show variance collapse (scale-shrinkage amount s, where deviations from the reference-feature mean are multiplied by $1 - s$), over-dispersion (adding independent Gaussian noise with coordinate-wise standard deviation τ times the reference standard deviation), location shift (translating by δ reference standard deviations along the leading PC), and mode-drop (omitting the displayed number of classes). Curves report power at the 0.05 significance level with $m = n = 200$ and 300 repetitions per level. Permutation-calibrated methods use 99 relabelings.

C COMPLETE EVALUATIONS OF PRETRAINED GENERATIVE MODELS

Table 1 summarizes the five evaluations in the main paper. We report the numerical sweeps first, followed by their sampling and calibration protocol and detailed BigGAN, DiT, and SiT results.

BigGAN-deep-128: ordering along truncation. On **BigGAN-deep-128** (Brock et al., 2019) (m=n=450, 10 ImageNet classes), FID is lowest at $\psi = .8$ and rises under stronger truncation. KID and the ZID score both increase monotonically as truncation tightens, from .0075 to .0173 and from 328 to 1150, respectively. The ZID test rejects all four generator–real comparisons at $p = .002$. ZID reports under-dispersion throughout, with S_D increasing from 42.2 to 176.7 (App. C.2).

CIFAR DDPM: sampling-step comparison and direction. On a pretrained CIFAR-10 **diffusion** model (Ho et al., 2020), ZID rejects all three generator–real comparisons, and both the ZID score and KID move toward their real–real values as DDIM steps increase (Song et al., 2021). FID likewise decreases from 173.6 to 130.8 toward the real–real value 121.0. ZID reports under-dispersion at five steps and a member-sign conflict at 15 and 25 steps (Table 8).

DiT-XL/2: one guidance sweep, two failure directions. The DiT-XL/2 summary and aggregate curves appear in Sec. 5 and Fig. 6; App. C.3 reports the complete per-class values.

SiT-XL/2: guidance transition in a second architecture. The same transition appears on the interpolant-based transformer **SiT-XL/2** (Ma et al., 2024), over six ImageNet classes with $n = 500$ and CFG $\in \{1, 2, 4, 8, 16\}$. Across the six classes, the signed dispersion readout indicates over-dispersion at CFG $= 1$ and under-dispersion at CFG $\in \{4, 8, 16\}$; at CFG $= 2$, five classes are under-dispersed and one is not assigned. At CFG $= 16$, FID exceeds its CFG $= 1$ value on five classes but remains lower on class 910 (107.2 versus 125.8; App. C.3, Fig. 8).

StyleGAN2-ADA: full-rank evaluation on FFHQ. We examine the collapse diagnosis outside diffusion models and ImageNet by sweeping **StyleGAN2-ADA** (Karras et al., 2020a;b) on **FFHQ** (Karras et al., 2019) over truncation $\psi \in \{1.0, 0.9, 0.7, 0.5\}$ (n=5000>d, giving a full-rank sample covariance and reduced finite-sample bias). Table 9 summarizes the sweep.

FID, KID, and Gaussian-kernel MMD all increase monotonically as truncation tightens, ordering departure from the untruncated generator. ZID likewise separates every truncated setting from the untruncated generator reference and additionally identifies under-dispersion, with a signed readout whose magnitude increases as ψ falls. Precision and Density instead increase as Recall declines, reflecting their emphasis on different aspects of the precision–recall tradeoff. In this full-rank FFHQ setting, FID, KID, Gaussian-kernel MMD, and the ZID score therefore agree on the truncation ordering, while ZID additionally supplies calibrated detection and direction.

Table 8: **DDPM sampling-step sweep.** A pretrained CIFAR-10 DDPM versus real CIFAR-10 in Inception (d=2048, m=n=200). The ZID equality test and diagnostic gate each use 499 outer permutations. The real–real row compares disjoint halves and has no assigned dispersion direction.

DDIM steps	FID	KID	ZID score	ZID test p-value	dispersion diagnosis
5	173.6	.0789	6442	.002	under-dispersion
15	138.3	.0223	519	.002	member-sign conflict
25	130.8	.0148	255	.002	member-sign conflict
real–real	121.0	.0004	.109	.878	not assigned

Table 9: **StyleGAN2-ADA truncation on FFHQ.** Columns compare independent draws at the indicated ψ with an untruncated generator reference; the $\psi = 1$ versus $\psi = 1$ baseline uses two independent untruncated draws. FID and ZID use 5000 samples; KID, Gaussian-kernel MMD, Precision, Density, and Recall use a fixed 2500-sample subset. Arrows give the conventional preferred direction. Within each metric row, bold marks the best value for that metric. The ZID test p-value and D_{ZID} are diagnostic quantities rather than ranking criteria and are not bolded. Larger ZID scores indicate stronger departures, while negative D_{ZID} indicates under-dispersion; the $\psi = 1$ versus $\psi = 1$ baseline has no assigned direction.

metric	$\psi = 1$ vs. $\psi = 1$ baseline	$\psi = .9$	$\psi = .7$	$\psi = .5$
FID ↓	**5.97**	6.97	22.13	57.89
KID ↓	$\mathbf{-4.4\times10^{-5}}$	.00119	.01461	.04596
Gaussian-kernel MMD ↓	$\mathbf{-4.0\times10^{-5}}$	.00130	.01664	.05510
Precision ↑	.817	.876	.931	**.963**
Density ↑	.964	1.169	1.604	**1.944**
Recall ↑	**.750**	.698	.506	.221
ZID score ↓	**.52**	626.6	1.35×10^{5}	1.41×10^{6}
ZID test p-value	.502	.002	.002	.002
D_{ZID}	undefined	-22.7	-441.0	-1215.6

What ZID adds across generators. Across the five generator families, ZID jointly provides scalar ordering, calibrated detection, and dispersion diagnosis. On BigGAN and StyleGAN2-ADA, the ZID score increases as truncation tightens and its signed readout identifies under-dispersion; conventional distances respond as well but do not supply direction. On DiT and SiT, the ZID score measures departure on both sides of the closest sampled guidance setting, while its signed readout distinguishes low-guidance over-dispersion from high-guidance collapse. On DDPM, ZID reports under-dispersion at five steps and a *member-sign conflict* at 15 and 25 steps.

C.1 Sampling and calibration for pretrained generative models

Section 5 evaluates precomputed feature banks from fixed released model weights; each entry is a particular reference–model comparison, not an estimate of variation across training or generation runs. Within each sweep, image preprocessing, feature extractor, reference source, and sample size remain fixed while the parameter in Table 1 changes. BigGAN draws a new 450-feature real subset from the same pool at each truncation value. Inception results use 2048-dimensional penultimate features. For all reported ZID equality-test and D-component p-values, 499 label relabelings act on the pooled features; neither the generator nor the feature extractor is rerun.

For BigGAN-deep-128, each truncation bank pools 450 generated features from ten ImageNet classes (IDs $\{0, 217, 482, 491, 497, 566, 569, 571, 574, 701\}$), compared with 450 real features from those classes; the real–real row uses disjoint real samples. Generated class labels are sampled uniformly, whereas the real subset is drawn from the pooled ten-class bank without fixed per-class quotas. The global relabeling test concerns equality of the pooled feature distributions, not within-class conditional equality. The CIFAR DDPM compares 200 generated features at 5, 15, and 25 DDIM steps with a fixed 200-feature real reference; its real–real row uses a disjoint real sample. The DiT-XL/2 and SiT-XL/2 studies compare 500 generated features with fixed class-matched real-feature banks for each class in $\{10, 190, 370, 550, 730, 910\}$ and guidance setting. DiT uses 25 sampling steps, and SiT uses a 50-step Euler ODE sampler. The StyleGAN2-ADA FFHQ study compares independent 5000-feature draws at $\psi \in \{.9, .7, .5\}$ with an untruncated $\psi = 1$ generator draw; its $\psi = 1$ versus $\psi = 1$ baseline uses a second independent untruncated draw. KID, Gaussian-kernel MMD, Precision, Density, and Recall use a fixed 2500-sample subset. Gaussian-kernel MMD uses the unbiased U-statistic; its RBF bandwidth is the median pairwise distance among the first 300 observations of the fixed untruncated reference subset. Checkpoint identifiers and available revisions, seeds, preprocessing, and software details are documented in the supplementary generator manifest.

C.2 BIGGAN TRUNCATION CONTROL

Table 10 adds dispersion ratios, component magnitudes, and signed readouts to the BigGAN overview earlier in this appendix. The trace ratio remains below one and D_{ZID} is negative in every generator row; the real–real row is not assigned.

Table 10: **BigGAN truncation control.** BigGAN-deep-128 versus real ImageNet from the same 10 classes (Inception-2048, m=n=450). The dispersion ratio is $\mathrm{tr}(\widehat{\Sigma}_{\mathrm{gen}})/\,\mathrm{tr}(\widehat{\Sigma}_{\mathrm{real}})$. For the generator rows, $(\widehat{p}_{\mathrm{ZID}}, \widehat{p}_D) = (.002, .002)$; for real–real, the pair is $(.598, .856)$.

ψ	disp. ratio	FID	KID	ZID score	S_W	S_D	D_{ZID}	dispersion diagnosis
1.0	0.82	90.0	.0075	328	329.1	42.2	-42.2	under-dispersion
0.8	0.79	88.7	.0108	566	566.4	96.1	-96.1	under-dispersion
0.6	0.77	96.8	.0145	813	813.8	119.3	-119.3	under-dispersion
0.4	0.76	101.6	.0173	1150	1151.2	176.7	-176.7	under-dispersion
real–real	1.00	70.3	$-.0004$	.353	1.05	.12	undefined	not assigned

C.3 DETAILED CFG SWEEPS: DIT AND SIT

Table 11 gives the class-specific DiT-XL/2 guidance results underlying the class-aggregated results in Fig. 6.

Table 11: **Per-class DiT-XL/2 classifier-free-guidance sweep.** Six ImageNet classes with class-matched real images (Inception-2048, $m = n = 500$). Parentheses report $\widehat{p}_D$; "over" and "under" abbreviate the dispersion diagnoses. All $\widehat{p}_{\mathrm{ZID}} = .002$ with 499 permutations.

class	CFG	FID	S_{ZID}	KID	Prec	Recall	S_W	D_{ZID}	diagnosis ($\widehat{p}_D$)
10	1	40.9	4.1e3	0.0149	0.652	0.805	4.1e3	41.1	over (.002)
	2	20.2	3.2e3	0.0053	0.890	0.639	3.2e3	-50.3	under (.002)
	4	27.7	26e3	0.0121	0.814	0.251	26e3	-100.2	under (.002)
	8	38.4	69e3	0.0216	0.698	0.065	69e3	-148.3	under (.002)
190	1	42.3	2.2e3	0.0128	0.698	0.745	2.2e3	18.4	over (.002)
	2	35.4	6.8e3	0.0140	0.868	0.500	6.8e3	-61.3	under (.002)
	4	53.1	33e3	0.0320	0.814	0.179	33e3	-83.8	under (.002)
	8	59.3	60e3	0.0398	0.782	0.109	60e3	-141.8	under (.002)
370	1	42.6	3.9e3	0.0153	0.624	0.773	3.9e3	30.8	over (.002)
	2	37.4	8.2e3	0.0176	0.818	0.532	8.2e3	-66.9	under (.002)
	4	56.2	54e3	0.0361	0.868	0.195	54e3	-175.7	under (.002)
	8	64.0	81e3	0.0408	0.892	0.025	81e3	-223.4	under (.002)
550	1	80.9	6.9e3	0.0374	0.498	0.739	6.9e3	39.2	over (.002)
	2	55.2	4.7e3	0.0221	0.850	0.559	4.7e3	-43.4	under (.002)
	4	87.0	45e3	0.0622	0.938	0.285	45e3	-150.0	under (.002)
	8	98.8	65e3	0.0716	0.936	0.152	65e3	-184.4	under (.002)
730	1	85.6	12e3	0.0375	0.514	0.732	12e3	30.6	over (.002)
	2	55.9	2.5e3	0.0152	0.818	0.558	2.5e3	-40.3	under (.002)
	4	87.8	35e3	0.0577	0.810	0.268	35e3	-124.1	under (.002)
	8	121.2	83e3	0.0926	0.702	0.123	83e3	-161.9	under (.002)
910	1	121.0	16e3	0.0467	0.410	0.731	16e3	44.7	over (.002)
	2	76.4	0.6e3	0.0087	0.692	0.535	0.6e3	-5.9	under (.008)
	4	92.0	12e3	0.0322	0.812	0.342	12e3	-71.5	under (.002)
	8	121.3	35e3	0.0567	0.772	0.111	35e3	-128.3	under (.002)

Figure 8 summarizes the corresponding guidance transition on **SiT-XL/2** (Ma et al., 2024), an interpolant transformer with a continuous flow objective. The sweep uses the same six classes and n=500, with CFG $\in \{1, 2, 4, 8, 16\}$. FID, KID, and the Frobenius covariance discrepancy are

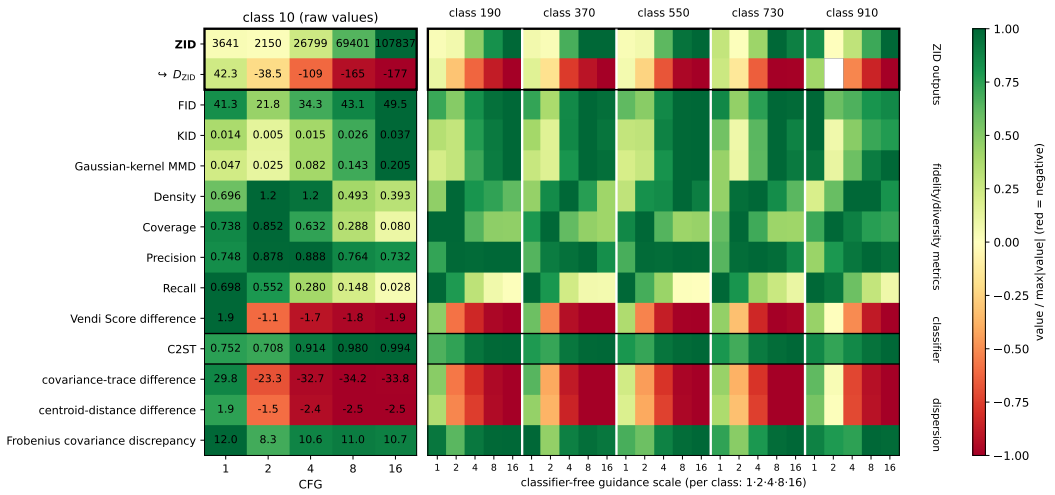


Figure 8: **Classifier-free-guidance sweep on SiT-XL/2.** Six ImageNet classes are compared with class-matched real images (n=500). Class 10 shows raw values; the other classes are normalized within each row and class, so color shows response shape rather than cross-row magnitude (green$>$0, red$<$0). The boxed rows show ZID and D_{ZID}; a blank cell denotes no assigned direction.

minimized near CFG=2 and increase on both sides, but their scalar values do not encode which side is over- versus under-dispersed (FID on class 10: $41 \to 22 \to 34 \to 43 \to 49$). The covariance-trace difference, centroid-distance difference, and Vendi Score difference corroborate that dispersion changes across the sweep. For reference sample X and generator sample Y, the covariance-trace difference is $\sum_k \{\widehat{\mathrm{Var}}(Y_k) - \widehat{\mathrm{Var}}(X_k)\}$, the Vendi Score difference is $V(Y) - V(X)$, and the centroid-distance difference is $\bar{r}_Y - \bar{r}_X$, where $\bar{r}_X = n^{-1}\sum_i \|X_i - \bar{X}\|$ and likewise for Y. The Frobenius covariance discrepancy is $\|\widehat{\Sigma}_Y - \widehat{\Sigma}_X\|_F$. ZID's final aggregate readout supplies the direction: on class 10, its D-component magnitude S_D grows from 42.3 at CFG= 1 to 176.6 at CFG= 16, while the direction changes from over-dispersion to under-dispersion.

D GAMING FID IN PIXEL SPACE

Proposition 1 states the matched-moment blind spot in feature space; here we realize it by optimizing actual pixels. Starting from noise, we optimize a set of $N = 512$ images at 64×64 by gradient descent to match an ImageNet reference's Inception mean and covariance, using loss $\|\hat{\mu}-\mu_r\|_2^2+\lambda\|\hat{\Sigma}-\Sigma_r\|_F^2$. Here $(\hat{\mu}, \hat{\Sigma})$ are the optimized batch's empirical feature mean and covariance, (μ_r, Σ_r) are those of the reference batch, $\|\cdot\|_F$ is the Frobenius norm, and the covariance weight is $\lambda = 1$. The pretrained Inception feature extractor is held fixed; gradient descent updates only the image pixels. After each Adam update, pixel values are clamped to $[0, 1]$. The reference and held-out real sets are drawn from four ImageNet classes (goldfish, jellyfish, balloon, and daisy), the optimization runs for 500 iterations, and all three sets are embedded with `pytorch-fid` v0.3.0.

This optimization directly targets the moments that determine FID. It lowers FID to 24.7, below the real–real baseline of 58.6, while the resulting images remain visually unrecognizable (Fig. 1). On the fixed feature banks, the real–holdout and real–optimized KID estimates are 8.6×10^{-6} and -5.2×10^{-5}, respectively (App. E.1). Against the same reference and using 499 outer permutations, the held-out-real comparison yields a ZID score of 0.86 (p=.328), whereas the optimized-set comparison yields 17.2, larger than all 499 relabeled scores. For the optimized pair, ZID reports component magnitudes S_W=17.9 and S_D=9.8; S_D also exceeds all 499 relabeled values. The RISE dispersion arm is not reference-active (Z_D^{RISE}=.58, normal-reference tail .565). The two active GPK arms have opposite signs (Z_D^{med}=$-$3.5, Z_D^{small}= + 4.3), so D_{ZID} is undefined and the dispersion diagnosis is *member-sign conflict*.

Because the 512 images are jointly optimized against the same reference used for evaluation, the generated batch is reference-adaptive rather than an i.i.d. draw independent of that reference. For this fixed pair, none of the 499 random outer relabelings produces an equally or more extreme score; the plus-one Monte Carlo tail is therefore .002. Because the pair is reference-adaptive, this tail is descriptive, not a calibrated population two-sample p-value. Thus moment matching can drive FID below its real-data baseline for visually unrecognizable samples, while ZID assigns this FID-targeted construction a large departure score under the same reference.

A separate CIFAR-10 replication with the same $N{=}512$ optimization drives the optimized FID to 31.1, below the 76.0 real–real baseline, while the ZID score exceeds all 499 randomly relabeled scores (App. E.1).

E Foundations and controlled validation

E.1 Matched-moment feature experiments

For a bimodal alternative whose mean and covariance match those of real CIFAR–Inception features, empirical FID remains at its real–real baseline as $m = n$ increases from 2k to 25k: the alternative/real–real values are $24.5/24.3$ at $n = 2$k and $2.00/1.99$ at $n = 25$k. At $n = 2$k, the ZID score has median 84.3 for the alternative and .32 for the corresponding real–real comparison; its permutation test rejects in all 20 repetitions at the minimum attainable $p = .01$ with 99 permutations.

Table 12 tests whether the feature-space FID inversion in the Introduction is specific to one dataset or embedding. Each comparison uses $m = n = 2500$ samples after projection to the leading 128 pooled principal components. The Gaussian construction is drawn from the Gaussian fitted to the reference sample.

Table 12: **FID under fitted-Gaussian feature comparisons.** Lower values are nominally better. The real–real column compares the reference with a held-out real sample; the Gaussian column replaces the held-out sample by a draw from the fitted Gaussian. DINO denotes the DINO ViT-S/16 CLS embeddings ($d = 384$) defined in Sec. B.5.

Dataset / embedding	real–real	fitted Gaussian
CIFAR-10 / Inception	2.517	1.157
Tiny-ImageNet / Inception	2.884	1.291
CIFAR-10 / DINO	76.472	33.864
Tiny-ImageNet / DINO	113.849	53.552

Table 13 uses an empirical moment-restoring affine transformation to closely match the reference sample's mean and covariance. By contrast, the preceding large-sample comparison draws from a fixed population-matched alternative, so its empirical FID follows the real–real finite-sample baseline rather than being numerically zero. In that comparison, empirical FID is evaluated through $n = 25$k observations in each sample. The six-arm ZID comparison uses $n = 2$k, 20 repetitions, and 99 outer permutations.

Table 13: **Full-dimensional matched-bimodal feature comparisons.** Each comparison uses raw Inception-2048 features with $m = n = 2500$. The matched-bimodal sample undergoes a moment-restoring affine transformation to closely match the reference mean and covariance. KID is the unbiased degree-3 polynomial-kernel MMD estimate and may be slightly negative.

Dataset	Metric	real–real	matched bimodal
CIFAR-10	FID	19.5146	0.0030
	KID	9.5×10^{-5}	-3.20×10^{-4}
Tiny-ImageNet	FID	25.8680	0.0024
	KID	-5.7×10^{-5}	-3.39×10^{-4}

The CIFAR-10 replication follows the same reference-adaptive stress-test design: a noise-initialized image batch is optimized to match the feature mean and covariance of one real sample, while a

disjoint real sample supplies the real–real baseline. Two disjoint $N = 512$ training samples are bilinearly resized from 32×32 to 64×64 and embedded, together with the optimized batch, by `pytorch-fid` v0.3.0. Starting from 512 independent uniform-noise images, Adam minimizes the loss of Sec. D for 500 steps (initial learning rate .03 with cosine decay), and the lowest-FID iterate is retained. Because the optimized batch is adapted to its evaluation reference, the resulting permutation tail is descriptive rather than a calibrated population two-sample p-value.

E.2 Proof of Proposition 1

Proof. Let μ_P and Σ_P denote the mean and covariance of P, and let $G = \mathcal{N}(\mu_P, \Sigma_P)$. If $P \neq G$, take $Q = G$. It has the same mean and covariance as P and is a different distribution. It remains to handle $P = G$. Write the nonzero part of the spectral decomposition as $\Sigma_P = U_r \Lambda_r U_r^\top$, where $r \geq 1$, and let $\varepsilon \in \{-1, +1\}^r$ have independent Rademacher coordinates. Define Q as the law of

$$\mu_P + U_r \Lambda_r^{1/2} \varepsilon.$$

This distribution has mean μ_P and covariance Σ_P, but it has finite support and therefore differs from the Gaussian law $P = G$ (including when that Gaussian is supported on a proper subspace). Thus in either case there is a $Q \neq P$ with matching first two moments, and $M(P, Q) = 0 = M(P, P)$. □

E.3 Controlled-departure definitions and signal levels

For each repetition, R and H are disjoint samples from the fixed CIFAR–Inception PCA pool, and the comparison is between R and a transformation B of H. Let $\hat{\mu}_j$ and $\hat{\sigma}_j$ denote the coordinatewise mean and standard deviation of H, let e_1 be the leading PCA direction, let $\sigma_1^\star$ be the first-coordinate standard deviation in the fixed pool, and let $\mathcal{M}_H(\cdot)$ denote the regularized moment-restoring affine transformation defined below. It restores the sample mean of H; covariance restoration additionally requires nonsingular covariances with no active eigenvalue regularization. For Fig. 3, $d = 128$ and $m = n = 200$; each cell uses 300 repetitions, and permutation-calibrated methods use 499 permutations. Each signal was chosen in a separate 16-repetition study, using seeds disjoint from the reported repetitions, to give a maximum power of about .85–.95 across the displayed methods (or the closest available setting); it was then held fixed across all methods. In every row, the equality setting leaves $B = H$. Figure 3 uses the following alternatives and signal levels.

departure	construction of B	displayed signal level
location	$B = H + \delta \hat{\sigma}_1 e_1$	$\delta = .379$
dispersion	$B = \bar{H} + (1 + \delta)(H - \bar{H})$	$\delta = .0369$
linear dependence	for the eight disjoint PCA-coordinate pairs $(2r-1, 2r)$, $r = 1, \ldots, 8$, leave coordinate $2r-1$ unchanged and set $B_{2r} = \hat{\mu}_{2r} + \sqrt{1-\rho^2}(H_{2r} - \hat{\mu}_{2r}) + \rho(\hat{\sigma}_{2r}/\hat{\sigma}_{2r-1})(H_{2r-1} - \hat{\mu}_{2r-1})$	$\rho = .42$
nonlinear dependence	for the four disjoint pairs $(2r-1, 2r)$, $r = 1, \ldots, 4$, rank-assign the values of H_{2r} according to a noisy ordering of H_{2r-1}^2, then apply $\mathcal{M}_H$	$\delta = .70$
matched-moment multimodality	for $c > 0$, set $B_{i1} = \sqrt{1-c^2}(H_{i1} - \bar{H}_1) + c s_i \sigma_1^\star$ with equal numbers of $s_i = -1$ and $+1$; leave the other coordinates unchanged	$c = .99968$
off-manifold support	add independent Gaussian noise of standard deviation $\delta \ \mathrm{median}_j \hat{\sigma}_j$ to the 64 lowest-variance PCA coordinates	$\delta = .47$

For nonlinear dependence, the noisy ordering score for pair r is $\sqrt{1-\delta}\,\varepsilon + \sqrt{\delta}\ \mathrm{zscore}(H_{2r-1}^2)$ with $\varepsilon \sim N(0, 1)$. Assigning the ordered values of H_{2r} by the ranks of this score preserves that coordinate's empirical marginal; before the affine transformation, the other coordinates are unchanged.

For a sample matrix A, write $\widehat{\Sigma}_{A,\epsilon}$ for its sample covariance with eigenvalues floored at $\epsilon = 10^{-8}$, and let $\bar{A}$ and $\bar{H}$ denote the sample row means of A and H, respectively. The affine transformation used in the experiments is

$$\mathcal{M}_H(A) = (A - \bar{A})\widehat{\Sigma}_{A,\epsilon}^{-1/2}\widehat{\Sigma}_{H,\epsilon}^{1/2} + \bar{H}.$$

When both covariances are full rank and their eigenvalues exceed ϵ, this map restores $\widehat{\Sigma}_H$. In singular regimes, including $n \leq d$, it need not do so: nonlinear coupling can leave residual covariance

differences after the transformation. Those grid cells evaluate the combined departure rather than isolating nonlinear dependence at exactly matched covariance (App. G.2).

The multimodality construction instead uses a fixed population-scale normalization, without $\mathcal{M}_H$. Its balanced mixture offsets and scale factor target the first two moments of the underlying mixture in the globally centered PCA coordinates; they do not restore H's empirical covariance in each repetition. Sample centering and finite-sample variation also distinguish this construction from the exact empirical moment-restoring affine transformation in Table 13. At $c = 0$, we set $B = H$; the positive-signal formula above is not applied.

Ranking sweeps. The ranking experiments use the following six-level sweeps, with equality first. Each repetition draws one reference sample shared across the six levels and an independent candidate sample for every level. All methods are evaluated on the same complete panels using their raw departure scores; no permutation p-value is used for ranking.

departure	sweep parameter and levels
location	$\delta:\ 0, .10, .20, .30, .40, .50$
dispersion collapse	variance ratio $v:\ 1, .97, .94, .91, .88, .85$; $B = \bar{H} + \sqrt{v}(H - \bar{H})$
linear dependence	$\rho:\ 0, .10, .25, .45, .65, .90$
nonlinear dependence	$\delta:\ 0, .10, .25, .50, .75, .95$
matched-moment multimodality	separation $\delta:\ 0, 4, 8, 16, 32, 64$; $c = (\delta/2)/\sqrt{1 + (\delta/2)^2}$
off-manifold support	$\delta:\ 0, .25, .50, 1, 1.5, 2$

Moment-restored resampling with perturbation. For Table 6, each repetition draws two disjoint sets of $n = 200$ real CIFAR–Inception PCA-128 feature vectors from the fixed pool. The first set, R, is the reference. The second set, $H = (h_1, \ldots, h_n)^\top \in \mathbb{R}^{n \times d}$, supplies the observations that are resampled and perturbed; each h_i is one observation represented by its $d = 128$ Inception feature vector. Let $f \in [0, 1]$ be the replacement fraction, let $\varepsilon \geq 0$ be the perturbation scale, and let $\hat{\sigma}_H \in \mathbb{R}^d$ be the vector of coordinatewise standard deviations of H. Initialize $\widetilde{B} = H$ and choose $\mathrm{round}(fn)$ observations without replacement; let I denote their indices. For each $i \in I$, independently draw J_i uniformly from $\{1, \ldots, n\}$ and set

$$\widetilde{b}_i = h_{J_i} + \varepsilon\, \hat{\sigma}_H \odot \xi_i, \qquad \xi_i \sim \mathcal{N}(0, I_d).$$

Here $\odot$ denotes coordinatewise multiplication and I_d is the $d \times d$ identity matrix. Observations outside I remain unchanged in $\widetilde{B}$. The final transformed sample is $B = \mathcal{M}_H(\widetilde{B})$, so it has the sample mean and covariance of H. Thus f is the fraction of observations replaced by perturbed resamples from H, while ε controls the coordinatewise perturbation scale relative to the standard deviations of H; $f = 0$ leaves $B = H$.

C2ST protocol. For Figs. 3, 7, and 11 and Table 7, including their null controls, each sample is split separately: one half enters training and the other half enters held-out testing, so both training and test halves contain equal numbers from the two samples. A `StandardScaler` is fitted on the training data only, followed by a one-hidden-layer multilayer perceptron (MLP) with 64 hidden units, `max_iter=300`, and fixed classifier seed 0; no cross-validation or hyperparameter selection is performed. After this single fit, the held-out predictions are fixed and the balanced test labels are relabeled. The plus-one upper-tail probability of the held-out correct count uses 499 relabelings in Fig. 3 and Table 7, and 99 in Figs. 7 and 11; conditional on the number of predicted positives, the corresponding complete randomization reference is hypergeometric. This fixed, untuned MLP is included as one learned baseline; its results are not intended to characterize the best achievable power of classifier two-sample tests under architecture or hyperparameter search.

Comparator settings. MIND uses 128 common random projections in Figs. 3 and 11, and 100 in Table 7, Fig. 7, and the ranking analyses. Precision, Recall, Density, and Coverage use $k = 5$. Vendi Score is computed from the cosine-similarity Gram matrix; its controlled test statistic is the difference between the two sample scores and uses a two-sided permutation tail. In Figs. 3 and 11, PERMDISP uses Euclidean distances to each group's spatial median, a one-way Levene-type F statistic, and residual permutations for calibration (499 and 99, respectively). In Figs. 3 and 11, Gaussian-kernel MMD uses the unbiased U-statistic with bandwidth equal to the median of all off-diagonal pairwise

distances in the pooled sample. Multiscale MMD uses the unbiased statistic for the sum of Gaussian kernels at .25, .5, 1, 2, and 4 times that pooled median distance, with an upper-tail permutation test. Each repetition reuses the same random projections, neighbor parameter, Gram or kernel matrix, and pooled sample across the relevant permutations. Table 14 summarizes the statistic, calibration, and tail used for every method in Figs. 3 and 11.

Table 14: **Test construction for each method in Figs. 3 and 11.** The statistic, calibration, and tail are fixed across all departures; no tail or orientation is chosen per alternative. Statistics for which larger values indicate greater departure use the upper tail; unbiased KID and Gaussian-kernel MMD^2 estimates can nevertheless be negative in finite samples. Precision, Recall, Density, and Coverage are nonnegative, but can move in either direction relative to the permutation reference; the Vendi Score difference can also have either sign. These rows use the two-sided permutation tail $\min\{1, 2\min(n_{\geq}, n_{\leq})/(B+1)\}$, where $n_{\geq}$ and $n_{\leq}$ count the $B+1$ labelings, including the observed one, whose statistic is at least or at most the observed value. The first sample is the reference throughout, and rejection uses $p \leq .05$.

method	statistic	calibration	tail
ZID	flat-Simes score S_{ZID}	label permutation	upper
RISE, GET	$Z_W^2 + Z_D^2$	asymptotic χ_2^2	upper
GPK-med, GPK-small	quadratic GPK statistic	label permutation	upper
FID	Fréchet distance between Gaussian fits	label permutation	upper
KID	unbiased degree-3 polynomial-kernel MMD^2	label permutation	upper
Gaussian-kernel MMD	MMD^2, median bandwidth	label permutation	upper
Multiscale MMD	MMD^2, five Gaussian bandwidths	label permutation	upper
MIND	mean squared difference of sorted random projections	label permutation	upper
ECS ($T = .10, .50$)	mean modulus of the characteristic-function difference	label permutation	upper
PERMDISP	Levene-type F on distances to spatial medians	residual permutation	upper
C2ST	held-out correct count	relabeled test labels	upper
Precision, Recall, Density, Coverage	metric of the second sample relative to the first	label permutation	two-sided
Vendi Score	difference of the two sample scores	label permutation	two-sided

Inception feature protocol. We use torchvision Inception-v3 (Szegedy et al., 2016) with `IMAGENET1K_V1` weights. Each RGB image is resized to 299 pixels, converted to a tensor, and normalized with ImageNet channel means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$. We retain the 2048-dimensional penultimate activation before any stated PCA reduction. The pixel-space stress test in Sec. D instead uses `pytorch-fid` v0.3.0 for its Inception features and FID; KID is computed separately on those features.

DINO feature protocol. All evaluations labeled DINO use the DINO ViT-S/16 model of Caron et al. (2021). The model is loaded from the `facebookresearch/dino` Torch Hub source at revision `7c446df5b9f4` with the official `dino_deitsmall16_pretrain.pth` checkpoint. Each RGB image is resized to 224 pixels with bilinear interpolation, center-cropped to 224×224, converted to a tensor, and normalized with ImageNet channel means $(.485, .456, .406)$ and standard deviations $(.229, .224, .225)$. We retain the 384-dimensional CLS embedding.

ECS protocol. We report two fixed ECS specifications, $T = .10$ and $T = .50$, spanning lower- and higher-frequency sensitivity. They are displayed as separate methods; no selection or maximization across T is performed. Because coordinatewise ECS is not invariant to coordinate rescaling or rotation, its reported row is conditional on the globally fitted PCA-128 coordinates and preprocessing used by the controlled comparison.

E.4 OUTER-PERMUTATION CALIBRATION

In the CIFAR–Inception null check at $d = 128, m = n = 200$, ZID rejects in .048 of 500 real-feature null repetitions at $\alpha = .05$ (999 random relabelings; Wilson 95% CI $[.032, .070]$). Table 15 evaluates the complete six-arm ZID test on real-feature null comparisons across two datasets and two embeddings. Its rejection rates remain near their nominal levels, supporting the outer-permutation calibration used in the main experiments.

Table 15: **Outer-permutation calibration on real features.** This independent real-null study applies one common protocol across four dataset–embedding settings: 500 repetitions per setting, $m = n = 200$, and 99 permutations for the six-arm equality test. The CIFAR–Inception row repeats the Fig. 3 null setting as a common reference point for the cross-setting comparison; the separate null check reported above uses 999 permutations for finer p-value resolution. DINO denotes the DINO ViT-S/16 CLS embeddings ($d = 384$) defined in Sec. B.5. Rejection uses $p \leq \alpha$; $p = .01$ is the smallest attainable value.

dataset / embedding	α=.01	.05	.10
CIFAR-10 / Inception	.008	.028	.094
CIFAR-10 / DINO	.010	.052	.104
Tiny-ImageNet / Inception	.010	.054	.098
Tiny-ImageNet / DINO	.014	.048	.090

E.5 GPK, GET, AND RISE SPECIFICATIONS AND SENSITIVITY TO ARM-LEVEL TAIL PROBABILITIES

GPK, GET, and RISE instantiate within-sample similarity differently: GPK uses dense Gaussian-kernel weights; GET uses the binary union of k edge-disjoint minimum spanning trees; and RISE uses a directed k-NN graph weighted by neighbor rank $(k, \ldots, 1)$. In the reported standalone analyses, GPK uses the pooled median-distance bandwidth, RISE uses $k = \lfloor N^{0.65} \rfloor$ nearest neighbors, and GET uses $k = \lfloor N^{0.5} \rfloor$ edge-disjoint minimum spanning trees, where $N = m + n$. Thus GET and RISE differ in graph construction and weighting, whereas GPK is kernel based. The standalone RISE and GET tests use $T = Z_W^2 + Z_D^2$ with their asymptotic χ_2^2 references. For GPK, $W - \mathbb{E}_0 W$ is proportional to $\widehat{\mathrm{MMD}}_u^2$ (Sec. 3.2), and the standalone GPK-med and GPK-small tests calibrate their quadratic statistics by label permutation rather than assigning a χ_2^2 reference from coordinate standardization. The normal-reference tail of an individual GPK Z_D coordinate is an arm-level quantity used in constructing ZID, not the standalone GPK omnibus p-value. These family-specific calibrations apply only to these standalone rows; the complete six-arm ZID p-value is the outer permutation tail of the complete aggregated score.

To check sensitivity to the Gaussian-reference arm tails used in ZID, a nested comparison replaces each arm's two-sided normal tail by an estimate from 299 inner label permutations. The resulting score retains the same bank of 99 outer relabelings for outer calibration. The comparison uses 300 paired repetitions on each of the six Fig. 3 departures; power at the 0.05 significance level is reported in Table 16. The two methods for assigning arm-level tail probabilities yield qualitatively similar departure-wise profiles, but the Gaussian-reference construction is more powerful on location (.823 versus .727), linear dependence, and multimodality, and slightly less powerful on off-manifold support; it is also computationally simpler.

E.6 HIGH-DIMENSIONAL DISPERSION SENSITIVITY AND COORDINATE SEPARATION

We compare median-bandwidth GPK, KID, and FID under one aligned protocol. Figure 9 compares power across dimension and sample size. The scale-shrinkage amount s multiplies standardized deviations from the reference mean by $1 - s$, so $s = 0$ is the null.

Within each repetition, the three methods use the same Gaussian reference–alternative pair and the same 99 label relabelings; power at the 0.05 significance level is estimated from 300 repetitions. The first three panels fix $n = 100$ and vary dimension over the full shrinkage grid; the fourth fixes $d = 100$ and $s = .05$ and varies sample size. At $n = 100$, GPK power for a 1% shrinkage is .98 at

Table 16: **Sensitivity to arm-level tail probabilities.** Each cell reports power at the 0.05 significance level over 300 paired repetitions. The Gaussian-reference specification uses the two-sided normal-reference tail of each standardized arm; the permutation-tail specification estimates each arm tail with 299 inner label permutations. Both specifications use the same bank of 99 outer relabelings for outer calibration.

departure	Gaussian-reference	permutation-tail
location	.823	.727
dispersion	.763	.763
linear dependence	.707	.643
nonlinear dependence	.977	.957
off-manifold support	.760	.777
matched-moment multimodality	.727	.660

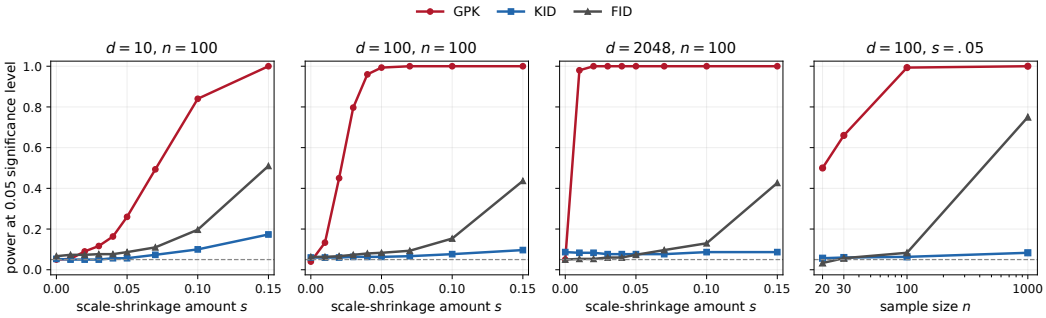


Figure 9: **Aligned sensitivity to Gaussian shrinkage across dimension and sample size.** Empirical power for median-bandwidth GPK, degree-3 KID, and FID; $s = 0$ denotes the null.

$d = 2048$; at $d = 10$, its power is .493 for a 7% shrinkage and .84 for a 10% shrinkage. At $d = 100$ and 5% shrinkage, GPK power is .50, .66, .993, and 1.00 at $n = 20, 30, 100$, and 1000, respectively. FID reaches .75 at $n = 1000$, while KID remains at .083. Across the displayed range, GPK responds sharply to small shrinkages as dimension or sample size increases, KID remains below .20, and FID rises more gradually.

F Additional aggregation, ranking, and bandwidth controls

These controls retain the six-arm definition and its GPK-small member. Unless stated otherwise, they use the six controlled departures, CIFAR-Inception PCA-128 features, and $n = 200$ observations per sample; each table states its repetition and permutation budgets. The ranking controls use the full-range sweeps from Sec. B.2; the power controls use the departure definitions and signal levels from Fig. 3, with independent samples.

Aggregation rule. Table 17 compares five ways to aggregate the same six standardized arms. For the power columns, every rule shares samples, arm matrices, and 199 outer relabelings within a replicate. Flat-Simes, max, and Cauchy have closely matched detection and panel-ranking summaries across the full severity ladders. The differences among flat-Simes, max, and Cauchy are descriptively small, while sum-of-squares and Fisher are weaker in the displayed power cells. Taken together, these results place flat-Simes in the leading group across detection and ranking summaries, supporting its use as ZID's aggregation rule.

Expanded ranking across increasing severity. Table 18 applies the same full-range ranking protocol to the four standalone tests and six external comparators: FID, KID, multiscale MMD, MIND, and ECS at two fixed frequencies. Multiscale MMD uses the sum of Gaussian kernels with bandwidths 0.25, 0.5, 1, 2, and 4 times the pooled median distance. Relative to these external comparators, ZID ranks first on the mean and minimum panel-wise Spearman correlations and all-pairs concordance. The minimum all-pairs accuracy is .765 for ZID, compared with .528 for the strongest external

Table 17: **Paired aggregation-rule comparison.** Mean and minimum summaries are across the six controlled departures. Power and null rejection are evaluated at the 0.05 significance level using 100 repetitions and 199 outer permutations per departure; ranking uses 100 independent six-level panels per departure. All rules are evaluated on identical arms within each replicate.

rule	power		null	panel ρ		all-pairs	
	mean	min	mean	mean	min	mean	min
flat-Simes	.807	.67	.037	.844	.649	.881	.765
max $\lvert Z_k \rvert$	.802	.66	.038	.847	.643	.883	.762
sum of squares	.762	.60	.040	.835	.619	.878	.760
Fisher	.752	.58	.047	.826	.603	.871	.752
Cauchy	.812	.68	.035	.834	.627	.874	.760

comparator and .709 for standalone RISE. Among the standalone tests, RISE remains close on the average summaries, whereas ZID has higher mean and minimum values. Adjacent-level and endpoint accuracies show how the methods behave on the hardest local comparisons and the easiest end-to-end comparison, respectively. Density is not included in Table 18 because converting its directional output into a two-sided departure score requires an additional orientation choice.

Table 18: **Expanded panel ranking across increasing severity.** Each departure uses 100 independent six-level panels with a common real reference and independently generated candidates. Mean and minimum are across the same six departure families; all-pairs and adjacent accuracies give ties half credit. Endpoint accuracy compares the first and final levels within each panel. MIND uses 100 common random projections per panel. The two ECS rows use $T = .10$ and $T = .50$ separately, without selection across them. Rows are grouped as the complete ZID construction, the standalone RISE, GPK-med, GPK-small, and GET tests, and external comparators.

method	mean ρ	min ρ	all-pairs	adjacent	endpoint
ZID	0.844	0.649	0.881	0.778	0.997
RISE	0.816	0.521	0.871	0.776	0.977
GPK-med	0.514	0.019	0.733	0.669	0.813
GPK-small	0.510	0.068	0.717	0.645	0.877
GET	0.619	0.236	0.770	0.674	0.900
FID	0.376	-0.478	0.678	0.646	0.683
KID	0.507	0.002	0.721	0.650	0.835
$\mathrm{MMD}_{\mathrm{ms}}$	0.627	0.063	0.774	0.677	0.902
MIND	0.543	0.017	0.735	0.647	0.842
ECS ($T = .10$)	0.288	-0.068	0.624	0.581	0.685
ECS ($T = .50$)	0.377	0.063	0.664	0.611	0.760

GPK-small bandwidth sensitivity. Within each replicate, Table 19 holds samples and relabelings fixed and changes only the GPK-small multiplier over seven prespecified values from .10 to .25. Across this range, minimum power varies from .70 to .71 and mean null rejection from .032 to .047. The neighboring .15 and .20 ratios differ from .175 by at most .05 and .03, respectively, in power across the six departures and have minimum paired decision agreement of at least .95 under the alternative and .98 under the null. Wider changes produce larger power differences or lower paired agreement on at least one departure. These quantities are descriptive sensitivity summaries rather than a rule for selecting the default.

Table 19: **Seven-point GPK-small bandwidth profile.** Only the GPK-small multiplier changes; the other four arms, samples, and 199 outer relabelings are shared across ratios within each of 100 repetitions per departure. Entries give the minimum power and mean null rejection rate across departures at the 0.05 significance level. The final three columns compare each ratio with .175: the largest absolute departure-wise power difference and the minimum paired decision agreement under the alternative and null.

multiplier	min. power	mean null	max $\lvert\Delta\rvert$	min. alt. agree	min. null agree
.100	.70	.032	.070	.90	.98
.125	.70	.038	.050	.91	.98
.150	.71	.035	.050	.95	.98
.175 (reference)	.70	.038	.000	1.00	1.00
.200	.70	.042	.030	.96	.98
.225	.70	.047	.050	.93	.98
.250	.70	.043	.050	.95	.97

G Detection across severity and across dimension–sample-size regimes

G.1 Detection across increasing severity

To examine how ZID's detection power changes as each departure strengthens, we extend the single signal level used in Fig. 3 to a five-level ladder. Each level uses 300 repetitions and 499 outer permutations per cell, with independent random seeds. Five departures use multiples $\{0, .5, .75, 1, 1.25\}$ of their Fig. 3 signal. For matched-moment multimodality, the ladder varies the replacement fraction $\lambda \in \{0, .25, .5, .75, 1\}$ from the unmodified real sample to the full bimodal alternative. Figure 10 shows monotonically increasing empirical power across all six departures.

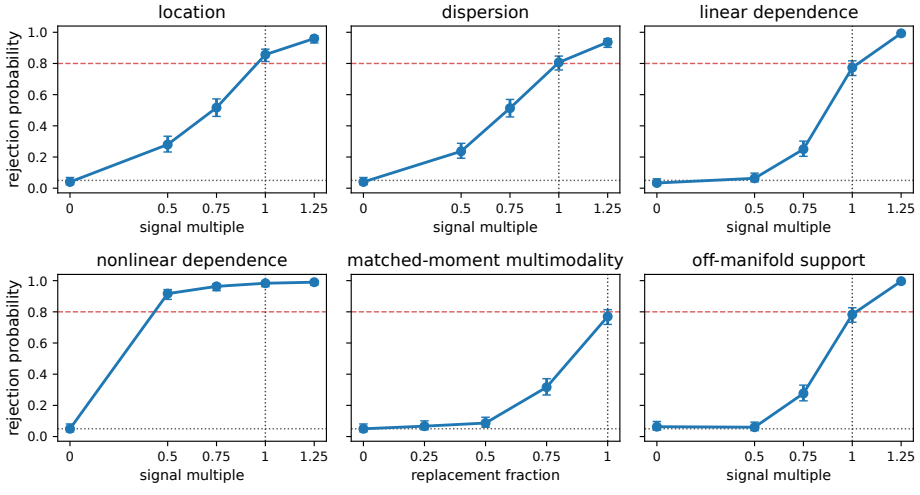


Figure 10: **Detection across increasing severity.** Empirical rejection rates use 300 repetitions and 499 outer permutations per cell at the 0.05 significance level; bars are Wilson 95% intervals. Horizontal dotted and dashed lines mark .05 and .80, and the vertical dotted line marks the Fig. 3 signal. The horizontal coordinate is a signal multiple except in the multimodality panel, where it is the replacement fraction.

G.2 Dimension–sample-size grid

We compare relative method sensitivity across $d \in \{128, 512, 2048\}$ and $n \in \{50, 200, 600\}$ for the same six controlled-departure constructions. Each panel uses the corresponding construction from App. E.3; only d, n, and the cell-specific signal change. Table 20 lists the signal in every cell,

Fig. 11 reports power for the same 19 methods in every panel, and Table 21 reports the maximum panel-specific null rejection estimate for each method and (d, n) cell. Signals were selected as for Fig. 3 (App. E.3), from the multi-method panel rather than ZID alone, and held fixed across methods within each cell. Because signals vary across cells, Fig. 11 compares methods within a cell rather than tracing power over d or n for a fixed alternative. For nonlinear dependence, the six cells with $n \leq d$ use the regularized affine transformation described in App. E.3; residual covariance differences may contribute to their power.

Table 20: **Cell-specific signal values for the dimension–sample-size grid.** Each row uses the corresponding controlled-departure construction from App. E.3. Location is a leading-PC shift in standard-deviation units of the unmodified real sample; dispersion uses multiplicative scale $1+$signal; linear dependence uses an injected correlation; nonlinear dependence uses the noisy rank-assignment coefficient before moment restoration; multimodality uses mode separation $\delta = 2c/\sqrt{1-c^2}$ in population standard-deviation units, without an empirical moment-restoring affine transformation; and off-manifold support uses the lower-variance-coordinate noise multiplier. Values are rounded for display.

	$d = 128$			$d = 512$			$d = 2048$		
departure	$n = 50$	$n = 200$	$n = 600$	$n = 50$	$n = 200$	$n = 600$	$n = 50$	$n = 200$	$n = 600$
Location	.7778	.3793	.2018	.7778	.4151	.2018	.7778	.3790	.2018
Dispersion	.09165	.04360	.02000	.07132	.03816	.01902	.07132	.03646	.01902
Linear dependence	.7860	.4200	.2364	.8223	.4571	.2782	.8223	.4571	.2782
Nonlinear dependence	.9500	.07000	.00300	.9500	.9500	.00300	.9500	.9500	.00300
Multimodality	90.0	90.0	2.449	90.0	90.0	5.903	90.0	90.0	8.000
Off-manifold support	.7843	.4700	.3261	.8039	.5853	.3909	1.510	.9969	.6503

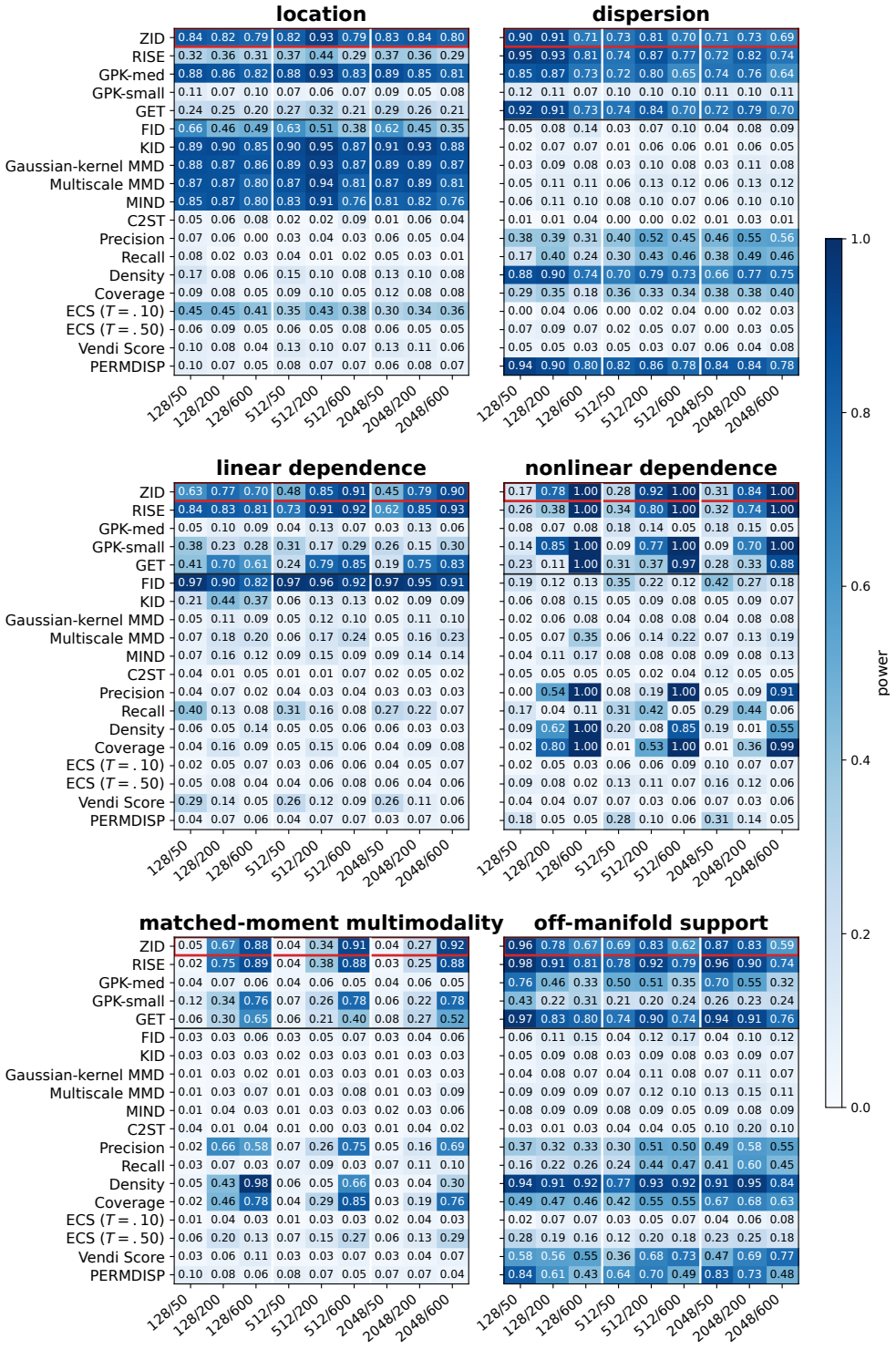


Figure 11: **Detection across departures, sample sizes, and dimensions.** The 19 rows match Fig. 3; columns are nine (d, n) cells and panels are six controlled departures. Each cell uses CIFAR-10 Inception features projected to d pooled principal components and 100 repetitions at the 0.05 significance level; permutation-calibrated rows use 99 permutations. Within each (d, n) cell, all methods use the same selected signal. The boxed row is ZID; GPK-small uses 0.175 times the pooled median pairwise distance.

Table 21: **Null rejection across the aligned robustness grid.** The 19 rows match Fig. 11. For each method and (d, n) cell, the table reports the maximum of the six panel-specific null rejection-rate estimates, providing a worst-panel calibration summary. Each estimate uses real CIFAR-Inception features and 100 repetitions at the 0.05 significance level; rejection is defined by $p \leq .05$, and permutation-calibrated rows use 99 permutations. The displayed values range from .00 to .10; the binomial s.e. of each underlying estimate is approximately .022. GPK-small uses 0.175 times the pooled median pairwise distance.

method	d=128			d=512			d=2048		
	n=50	n=200	n=600	n=50	n=200	n=600	n=50	n=200	n=600
ZID	0.04	0.04	0.06	0.03	0.06	0.04	0.03	0.06	0.04
RISE	0.07	0.03	0.05	0.05	0.05	0.04	0.03	0.04	0.04
GPK-med	0.04	0.05	0.06	0.01	0.05	0.07	0.02	0.04	0.06
GPK-small	0.06	0.04	0.05	0.03	0.04	0.04	0.04	0.04	0.02
GET	0.05	0.05	0.05	0.03	0.05	0.04	0.08	0.06	0.06
FID	0.03	0.05	0.07	0.03	0.08	0.07	0.03	0.08	0.07
KID	0.02	0.04	0.05	0.05	0.04	0.05	0.04	0.04	0.05
Gaussian-kernel MMD	0.05	0.04	0.04	0.03	0.05	0.03	0.03	0.05	0.03
Multiscale MMD	0.03	0.08	0.04	0.03	0.07	0.03	0.02	0.07	0.03
MIND	0.04	0.03	0.04	0.04	0.06	0.04	0.08	0.06	0.03
C2ST	0.01	0.05	0.05	0.01	0.07	0.10	0.01	0.00	0.05
Precision	0.03	0.04	0.06	0.03	0.02	0.05	0.02	0.02	0.05
Recall	0.00	0.01	0.05	0.03	0.03	0.04	0.02	0.04	0.03
Density	0.01	0.04	0.06	0.05	0.04	0.04	0.02	0.04	0.03
Coverage	0.03	0.02	0.06	0.05	0.03	0.03	0.01	0.03	0.01
ECS ($T = .10$)	0.05	0.03	0.04	0.02	0.07	0.04	0.04	0.07	0.07
ECS ($T = .50$)	0.09	0.06	0.03	0.07	0.06	0.06	0.05	0.06	0.07
Vendi Score	0.04	0.03	0.06	0.03	0.06	0.03	0.03	0.05	0.03
PERMDISP	0.08	0.02	0.09	0.07	0.05	0.06	0.05	0.05	0.06