QID: QUANTILE INCEPTION DISTANCE FOR GENERATIVE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative model development requires repeated comparisons with a dataset of real images, often using only a small batch of generated samples at each checkpoint. We introduce Quantile Inception Distance (QID), a fast, sample-efficient metric that represents feature distributions by projected quantile displacements from a common reference. This optimal transport embedding provides coordinates for comparing generated distributions with the real dataset and with one another. We derive a finite-dimensional representation of this embedding, enabling efficient comparison of small generated batches with the full real dataset. We also derive fQID, a Gaussian approximation for faster scoring. We evaluate QID on over one million real and generated images across ImageNet, COCO and CIFAR-10. In small-batch ImageNet comparisons with Inception features, QID reduces pairwise ranking errors nearly fivefold relative to FID and roughly halves the variance of score differences between models relative to MIND. On ImageNet, QID achieves higher pairwise ranking agreement than polynomial and RBF kernel estimators with 2,000–10,000 generated images in both Inception and DINOv2 space. Speedups reach $89\times$ over full-reference kernel estimation and $234\times$ over FID.

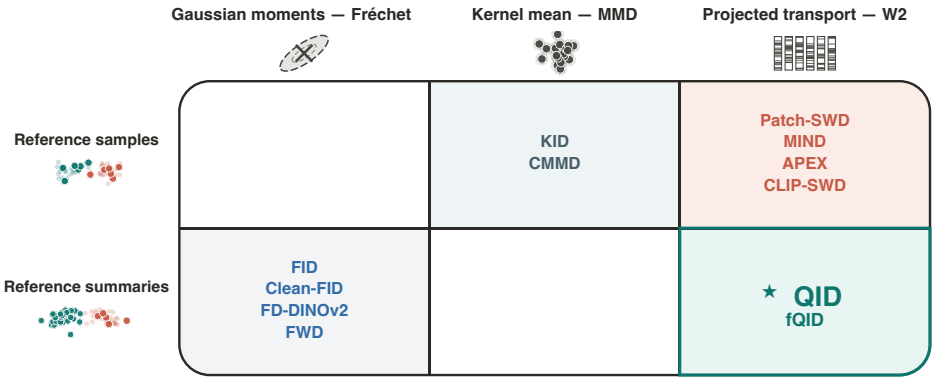

Figure 1: Generative evaluation metrics grouped by their distribution comparison and representation of the real reference.

1 INTRODUCTION

A generative model is evaluated repeatedly as its parameters and sampling procedure change. Selecting a checkpoint, comparing sampling algorithms and choosing the number of sampling steps require reliable comparisons from the generated images available at each evaluation. The real dataset usually contains many more images and remains available throughout this process, making it natural to prepare its representation once and reuse it. Fréchet Inception Distance (FID) already follows this approach by storing the real feature mean and covariance (Heusel et al., 2017; Parmar et al., 2022). However, its Gaussian comparison cannot distinguish distributions with the same first two moments, and its finite-sample estimates can change model rankings (Chong and Forsyth, 2020; Jayasumana et al., 2024; Berthet et al., 2026). Repeated evaluation therefore calls for a reference representation that retains the distributional information lost by its Gaussian moments.

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.

Optimal transport embeddings represent distributions by their transport displacements from a common reference (Wang et al., 2013; Kolouri et al., 2016a). In one dimension, the displacement is the difference between quantile functions, and its squared norm is the Wasserstein cost. Applying this construction along feature directions yields sliced-Wasserstein coordinates (Kolouri et al., 2016b), which describe changes in the projected distributions beyond their means and variances. PSWE uses quantile interpolation when set sizes differ, while FourierSWE integrates projected quantiles against sampled cosine functions (Naderializadeh et al., 2021; Amir and Dym, 2025). MIND, APEX and CLIP-SWD use projected quantiles for generative evaluation and demonstrate their advantages in sample efficiency, computation and sensitivity (Berthet et al., 2026; Gallegati et al., 2026; Kwon et al., 2026). The common-reference formulation makes these comparisons part of one coordinate space: each generator is represented relative to the real benchmark, and differences between its coordinates and those of another generator give their projected transport distance. We therefore seek a finite representation of this reference that remains efficient when a generated batch is much smaller than the real dataset.

We introduce Quantile Inception Distance (QID) by constructing that finite representation. A batch of n equally weighted samples has an empirical quantile that is constant on n equal intervals, whereas a larger reference partitions the quantile axis more finely. We project the reference quantile onto the generated batch's step-function space to obtain interval averages, which also form its optimal equally weighted n-point Wasserstein approximation (Xu and Berger, 2019). The component orthogonal to this space contributes a constant squared residual. By retaining both the averages and residual, we preserve the full-reference transport score for every generated batch of size n. Together they provide finite displacement coordinates ($Mn + 1$ for M fixed projection directions) for reference scoring and comparisons between equal-size generated batches, with reference storage proportional to the generated batch size (Table 2). After preparing the reference, each evaluation requires projection, sorting and a sum of squared coordinate values. We also derive fQID, which replaces the generated feature law by a diagonal Gaussian and reduces this computation to feature moments and a closed-form expression.

We also study how the reference affects repeated model comparisons. Reusing the same real sample removes the variation introduced by drawing a new reference, while increasing its size changes the empirical distribution against which models are compared. In the transport coordinates, a model-score difference is a difference of squared norms, allowing us to express how it changes when the reference changes. We derive the corresponding score-gap and variance identities and establish consistency as the reference, generated batch and number of directions grow. We then examine how ranking reliability changes when a real batch is shared within a comparison, reused across comparisons, or replaced by the full real dataset.

Our experiments span over one million real and generated images across ImageNet (Deng et al., 2009), COCO (Lin et al., 2014) and CIFAR-10 (Krizhevsky, 2009). They include eleven ImageNet generator configurations evaluated in Inception and DINOv2 space and six text-to-image configurations evaluated on COCO in CLIP space. With 512 generated images per model, QID reduces ImageNet ranking errors nearly fivefold relative to FID and roughly halves the variance of model-score differences relative to MIND in Inception space, with improvements also observed in DINOv2 space. COCO extends reliable small-batch ranking to text-to-image generation. QID scores generated features up to $234\times$ faster than FID against a 100,000-image reference and, at 1000 generated images, uses 66–99 times less reference storage than a representation based on sorted projections and prefix sums. Controlled experiments complement these model comparisons: QID detects differences between feature sets with identical sample means and covariances, demonstrating the information retained by the quantile representation.

2 Problem Setting and Background

Let P_{data} be the population image law and $\psi : \mathcal{X} \to \mathbb{R}^d$ a fixed feature map. For a generator g_θ and latent $Z \sim \mathcal{N}(0, I)$, the feature laws are

$$X = \psi(g_\theta(Z)) \sim p_\theta, \qquad Y = \psi(Y_{\mathrm{data}}), \quad Y_{\mathrm{data}} \sim P_{\mathrm{data}}, \quad Y \sim p_{\mathrm{data}}. \tag{1}$$

Table 1: Distribution comparisons and reference representations. QID represents the reference by the quantile averages and residual needed for a generated batch of size n.

Metric	Features	Comparison	Reference representation	Batch scoring
FID	Inception	Gaussian W_2	Mean and covariance	Moments and matrix square root
KID, CMMD	Inception or CLIP	Kernel MMD	Real feature pool	Pairwise kernels
MIND	Chosen embedding	Empirical sliced W_2^2	Real feature batch	Project, sort and sum
APEX, CLIP-SWD	CLIP or chosen embedding	Empirical sliced W_2^2	Sorted real projections	Project, sort and integrate
QID	Chosen embedding	Empirical sliced W_2^2	Quantile averages and residual	Project, sort and sum
fQID	Chosen embedding	Generated Gaussian approximation	Three reference slice summaries	Moments and closed form

For conditional generators, p_θ also averages over the benchmark conditioning inputs. We evaluate empirical generated laws $\hat{p}_{n,\theta}$ against a real set $D_{\mathrm{ref}} = \{y_j\}_{j=1}^N$ with feature measure

$$\hat{p}_{N,\mathrm{ref}} = \frac{1}{N}\sum_{j=1}^{N} \delta_{\psi(y_j)}. \tag{2}$$

The reference, feature map and preprocessing are fixed throughout repeated evaluation. The generated batch size n may vary and is typically much smaller than N.

FID represents the two feature distributions by empirical means μ_X, μ_Y and covariances Σ_X, Σ_Y, then computes their squared Gaussian Wasserstein distance,

$$\mathrm{FID} = \|\mu_X - \mu_Y\|^2 + \mathrm{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_Y\Sigma_X)^{1/2}). \tag{3}$$

Sliced Wasserstein retains the projected distributions and averages their one-dimensional transport costs over directions,

$$\mathrm{SW}_2^2(p,q) = \mathbb{E}_{u\sim\mathcal{U}(\mathbb{S}^{d-1})}\left[W_2^2(u^\top p, u^\top q)\right]. \tag{4}$$

Here $u^\top p$ denotes the law of the projection onto u. For equal-size empirical measures on the line, sorting gives

$$W_2^2(\hat{p}_n, \hat{q}_n) = \frac{1}{n}\sum_{i=1}^{n} |x_{(i)} - y_{(i)}|^2. \tag{5}$$

MIND applies this identity along random feature directions,

$$\mathrm{MIND}(\hat{p}_{n,\theta}, \hat{p}_{n,\mathrm{data}}) = \frac{\alpha}{M}\sum_{m=1}^{M} W_2^2(u_m^\top \hat{p}_{n,\theta}, u_m^\top \hat{p}_{n,\mathrm{data}}), \qquad \alpha = 3d. \tag{6}$$

For a fixed set of directions $U = (u_1, \ldots, u_M) \subset \mathbb{S}^{d-1}$, we write the projected distance as $d_U(p,q) = [M^{-1}\sum_m W_2^2(u_m^\top p, u_m^\top q)]^{1/2}$. The scores below use its square, multiplied by a positive scale α held fixed within each experiment.

3 Quantile Inception Distance

3.1 Common-reference transport coordinates

For $p, p_{\mathrm{ref}} \in \mathcal{P}_2(\mathbb{R}^d)$, write $Q_{p,m}$ and $Q_{R,m}$ for the quantile functions of their projections along u_m. Their difference at each quantile level describes displacement from the reference on that slice (Kolouri et al., 2016b). We collect these displacements over the directions in the common-reference transport embedding

$$\Phi_{p_{\mathrm{ref}},U}(p)(m,t) = \sqrt{\frac{\alpha}{M}}\big(Q_{p,m}(t) - Q_{R,m}(t)\big), \qquad t \in (0,1). \tag{7}$$

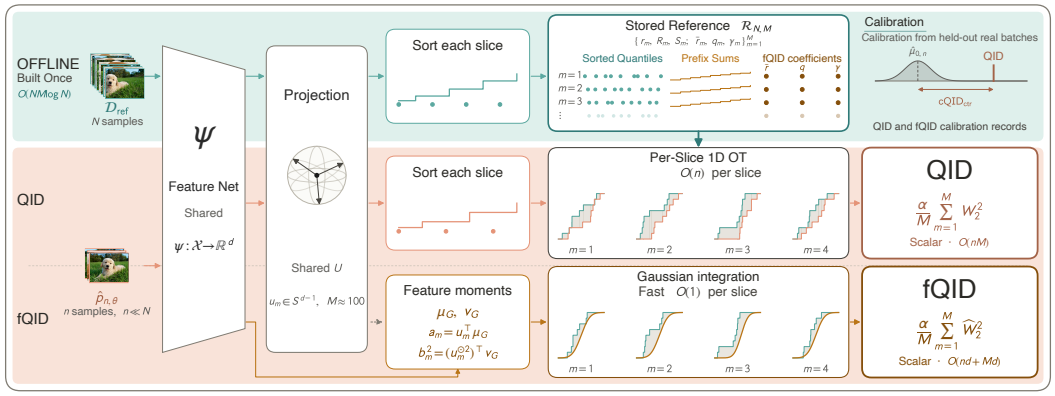

Figure 2: Reference preprocessing and batch evaluation for QID and fQID. Sorted projections and prefix sums supply QID's interval coefficients, while the Gaussian approximation uses three reference summaries per slice. The calibration panel shows the real-data score adjustment described in Appendix F.

The one-dimensional quantile identity relates the norm of these coordinates and their pairwise differences to sliced transport costs,

$$\|\Phi_{p_{\mathrm{ref}},U}(p)\|_{L^2}^2 = \alpha d_U(p, p_{\mathrm{ref}})^2, \qquad \|\Phi_{p_{\mathrm{ref}},U}(p) - \Phi_{p_{\mathrm{ref}},U}(q)\|_{L^2}^2 = \alpha d_U(p,q)^2. \tag{8}$$

The first identity compares a generated distribution with the real reference. In the second, subtracting the two coordinate functions cancels the reference quantile, leaving the sliced transport cost between the generated distributions. The same coordinates therefore support both kinds of comparison. We use the empirical benchmark as this common reference and define QID by

$$\mathrm{QID}_{n,N,M}(\hat{p}_{n,\theta}; \hat{p}_{N,\mathrm{ref}}, U) = \|\Phi_{\hat{p}_{N,\mathrm{ref}},U}(\hat{p}_{n,\theta})\|_{L^2}^2 = \frac{\alpha}{M} \sum_{m=1}^{M} W_2^2(u_m^\top \hat{p}_{n,\theta}, u_m^\top \hat{p}_{N,\mathrm{ref}}). \tag{9}$$

3.2 Finite Empirical Representation

To obtain finite coordinates, we work on the generated batch's quantile partition. An empirical quantile from n equally weighted samples is constant on the intervals $I_i = ((i-1)/n, i/n]$. We use the orthogonal projection Π_n in $L^2(0,1)$ onto this step-function space and retain each reference slice's interval averages and squared projection residual,

$$\beta_{m,i}^{(n)} = n \int_{I_i} Q_{R,m}(t)\mathrm{d}t, \qquad c_m^{(n)} = \sum_{i=1}^{n} \int_{I_i} \left(Q_{R,m}(t) - \beta_{m,i}^{(n)}\right)^2 \mathrm{d}t. \tag{10}$$

The step function $\Pi_n Q_{R,m}$ takes value $\beta_{m,i}^{(n)}$ on I_i. As the L^2 projection with nondecreasing interval averages, it defines the reference slice's optimal equally weighted n-point Wasserstein approximation (Xu and Berger, 2019, Theorem 5.5 and Remark 5.6(ii)). Adding the orthogonal residual recovers the cost of comparison with the full reference.

Proposition 3.1 (Reference compression)**.** *For arbitrary positive n, N, let $x_{m,(i)}$ be the sorted projections of an equally weighted generated batch, and let $\hat{p}_{n,m}, \hat{p}_{N,m}$ be its projected law and the projected reference law. Then*

$$W_2^2(\hat{p}_{n,m}, \hat{p}_{N,m}) = \frac{1}{n} \sum_{i=1}^{n} \left(x_{m,(i)} - \beta_{m,i}^{(n)}\right)^2 + c_m^{(n)}, \qquad c_m^{(n)} \geq 0. \tag{11}$$

The generated quantile is constant on each I_i, so its squared distance from the reference separates into its squared distance from the interval average and the reference variation within that interval. The cross term vanishes because the residual has zero integral there. Summing over intervals gives the stated Pythagorean identity, with the same residual cost for every generated batch of size n.

Using the slice displacements and their total residual, we form the finite coordinates

$$\Psi_{R,U,n}(\hat{p}_n) = \left(\left\{ \sqrt{\frac{\alpha}{Mn}} \left(x_{m,(i)} - \beta_{m,i}^{(n)}\right) \right\}_{m,i}, \sqrt{C_{R,U,n}} \right), \qquad C_{R,U,n} = \frac{\alpha}{M} \sum_{m} c_m^{(n)}. \tag{12}$$

The squared norm of the coordinate vector is QID, while subtracting the vectors of two batches of size n cancels both the reference averages and the residual coordinate. Their squared Euclidean distance is therefore the scaled empirical sliced-Wasserstein cost between the batches. When $N = n$, the reference averages are its sorted values and the residual vanishes, recovering MIND's empirical score under the feature, direction and scale conventions above. Batches of different sizes can be represented on a common quantile partition, as described in Appendix B.2.

Preparation and scoring. At fixed α, we retain $\mathcal{C}_{R,U,n} = (U, \{\beta^{(n)}_{m,i}\}_{m,i}, C_{R,U,n})$ for repeated evaluation. We prepare this representation by projecting and sorting the real features once, at cost $O(MNd + MN \log N)$, then computing the coefficients from prefix integrals in $O(Mn)$ operations for each n (Algorithm 1). The directions and coefficients occupy $O(Md + Mn)$ storage, while an optional $O(MN)$ prefix workspace supports preparation for additional batch sizes. For each generated batch, we project its features and sort each slice in $O(Mnd + Mn \log n)$ operations and compute its score by an $O(Mn)$ sum, with no need to revisit the real samples.

3.3 Gaussian approximation

To avoid sorting each generated projection, we derive fQID by approximating the generated feature distribution with a diagonal Gaussian while retaining the empirical reference quantiles. The generated projected quantiles then take the form

$$Q_{G,m}(t) \approx a_m + b_m \Phi^{-1}(t), \qquad a_m = u_m^\top \mu_G, \qquad b_m^2 = (u_m^{\odot 2})^\top v_G, \tag{13}$$

where $\mu_G = n^{-1} \sum_i h_i$, $v_G = n^{-1} \sum_i (h_i - \mu_G)^{\odot 2}$, $h_i = \psi(x_i)$, and Φ is the standard normal CDF. Under this approximation, the reference enters through three coefficients per slice: its mean $\bar{r}_m$, second moment q_m, and inner product $\gamma_m = \int_0^1 Q_{R,m}(t)\Phi^{-1}(t)\mathrm{d}t$. Integrating the squared displacement gives

$$\mathrm{fQID}_{n,N,M} = \frac{\alpha}{M} \sum_{m=1}^{M} \left(a_m^2 + b_m^2 + q_m - 2a_m \bar{r}_m - 2b_m \gamma_m \right). \tag{14}$$

The generated mean and coordinate variances suffice to evaluate this expression, giving a scoring cost of $O(nd + Md + M)$. The approximation omits feature covariances when computing projected variances and replaces empirical projected quantiles by Gaussian quantiles. Appendix E derives the score and separates these two contributions to its error (Nadjahi et al., 2021; del Barrio et al., 1999).

4 Properties

The quantile identities give d_U a Hilbert-space representation and hence a triangle inequality. For a finite bank, d_U is a pseudometric on feature laws: it identifies distributions with the same projections on U (Nadjahi et al., 2020). We use the triangle inequality to compare the empirical and population references,

$$\left| d_U(p_\theta, p_{\mathrm{data}}) - d_U(p_\theta, \hat{p}_{N,\mathrm{ref}}) \right| \le d_U(p_{\mathrm{data}}, \hat{p}_{N,\mathrm{ref}}). \tag{15}$$

Thus the discrepancy between the empirical and population references bounds the change in d_U for every generator. For $p_\theta, p_{\mathrm{data}} \in \mathcal{P}_2(\mathbb{R}^d)$, form their empirical measures from iid sample sequences and take the first M directions of one iid uniform sequence. Then $\mathrm{QID}_{n,N,M} \to \alpha \mathrm{SW}_2^2(p_\theta, p_{\mathrm{data}})$ almost surely as $n, N, M \to \infty$ at arbitrary relative rates (Appendix D).

To study model rankings, we consider differences between squared reference distances. Writing $S_r(p) = \alpha d_U(p, r)^2$ and $\Delta_r(p, q) = S_r(p) - S_r(q)$, a reference shared by the two models cancels the squared-reference term in their score gap. Replacing r by s changes the remaining term according to

$$\Delta_r(p, q) - \Delta_s(p, q) = -\frac{2\alpha}{M} \sum_m \int_0^1 (Q_{p,m} - Q_{q,m})(Q_{r,m} - Q_{s,m})\mathrm{d}t. \tag{16}$$

The change in a model comparison therefore depends on how the reference difference aligns with the difference between the generated quantiles. Reusing r across evaluations fixes this contribution from the real sample, leaving variation over generated batches. Under square-integrability, the law

of total variance decomposes fresh-reference variation into the average conditional variance and the variance of the conditional mean over references, with each chosen reference having its own conditional variance. Appendices D.3 and D.4 give the identities used to study reference sharing, reuse and size in the paired experiments.

5 Experiments

5.1 Model ranking from small samples

Following the sample-size analyses of Berthet et al. (2026); Jayasumana et al. (2024), we evaluate model rankings as the number of generated images decreases. Our ImageNet-256 benchmark contains eleven generator configurations with 50,000 released images each (Stein et al., 2023). We use half of each generated pool to establish a ranking against 100,000 real images and the other half to draw evaluation batches of 64–10,000 images. We compare QID, fQID, Fréchet distance and two kernel estimators in Inception and DINOv2 space, evaluating each metric against its own large-sample ranking over 512 trials. Ranking agreement measures the fraction of correctly ordered model pairs, while selection error measures how often the selected model differs from the large-sample winner (Appendix G).

QID recovers the ImageNet rankings more reliably than Fréchet distance with small batches (Figure 3). At 512 images per model, agreement in Inception space rises from 68.7% for FID to 93.5% for QID, giving nearly fivefold fewer pairwise errors, and the corresponding DINOv2 comparison rises from 91.5% to 97.2%. The kernel estimators have higher agreement at this batch size, but QID overtakes both at 2,000–10,000 images in both feature spaces. Selection of the best model follows a different pattern: Inception KID has lower selection error across the tested sample sizes, even where QID recovers more of the pairwise ordering.

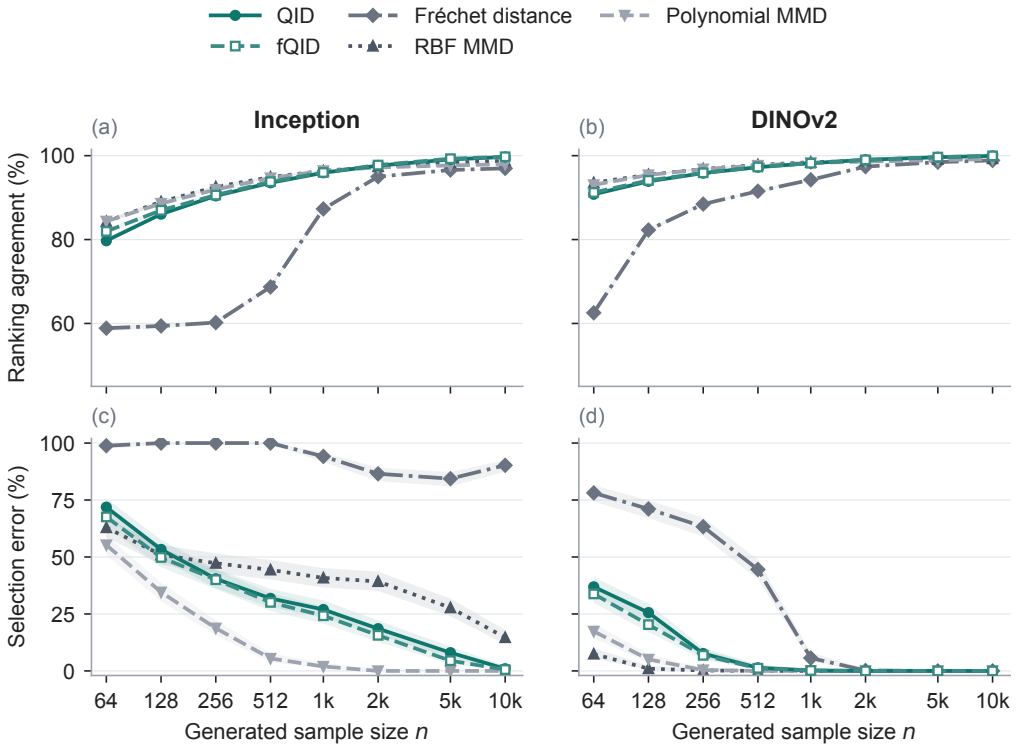

Figure 3: Model ranking on ImageNet-256. Upper panels show agreement with each metric's large-sample ranking, and lower panels show model-selection error. Bands are pointwise 95% intervals over 512 trials. Fréchet distance and polynomial MMD correspond to FID and KID in Inception space.

For text-to-image generation, we compare six SD-Turbo and SDXL-Turbo configurations using one, two and four sampling steps (Sauer et al., 2024), with CLIP features and a 118,287-image COCO

reference. We divide the generated images into ranking and evaluation pools, keeping captions from each source image together, and repeat the comparison after exchanging the pools (Appendix H). In both splits, small batches recover the large-sample QID ordering with high agreement (Figure 4). With 512 images per configuration, QID reaches 99.01% and 98.29% agreement, compared with 98.93% and 98.09% for MIND, evaluated with a fresh real batch shared across models in each trial. Both methods select the large-sample winner in every trial.

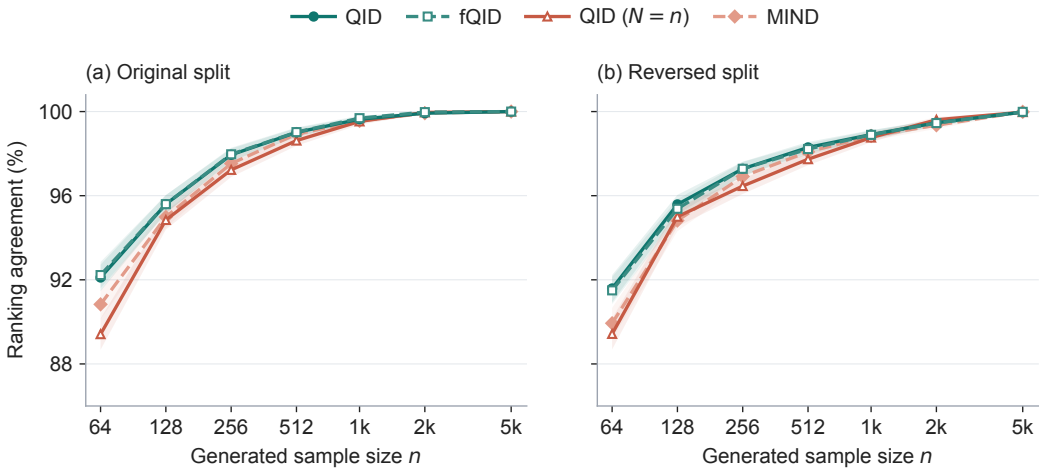

Figure 4: Ranking six text-to-image configurations on COCO with CLIP features. The two panels exchange the image pools used for the large-sample QID ranking and evaluation. Bands are pointwise 95% intervals over 512 trials.

5.2 Effect of reference reuse

To measure how reference sampling contributes to variation in model comparisons, we score the same ImageNet batches with three reference choices. QID uses the full real pool, QID ($N = n$) reuses one fixed subset, and we evaluate MIND with a fresh n-image subset shared across models in each trial. The projected transport distance and directions remain the same, so these choices vary only the real reference. We measure the variance of each model-score difference over 512 trials and average over the 55 pairs, reporting the result relative to MIND at each sample size (Figure 5). At $n = 512$, QID reduces variance by 52.0% in Inception and 21.1% in DINOv2. The small fixed reference gives nearly the same reduction, so reuse accounts for most of the gain. At this size, the full reference recovers more pairs of the large-sample QID ordering, whereas the small reference selects its winner more often. Appendix J examines these ordering differences across reference subsets and projection directions.

5.3 Computational cost

We compare the cost of repeated evaluation using a prepared reference of 100,000 real images. Timings start from extracted features, following Berthet et al. (2026); Jayasumana et al. (2024). Figure 6 compares QID and fQID with Fréchet distance and the polynomial and RBF kernel estimators as the generated batch grows. In Inception space, QID is 104× faster than FID at 1,000 images and 234× faster at 5,000 images. Across these two sizes and both Inception and DINOv2 representations, the speedups over the full-reference kernel estimators are 31–89×.

To measure the storage saved by reference compression, we compare representations that preserve the same full-reference score. At $n = 1000$, QID's interval averages and residual use 66–99 times less storage than retaining the sorted projections and their prefix sums, with similar scoring times (Table 2). We prepare these coefficients once for each batch size and report the preparation cost over repeated evaluations in Appendix M.

fQID reduces computation further by replacing the generated quantiles with their Gaussian approximation. Its speedup over QID is 1.05–1.20× on the GPU and 4–20× in the CPU experiments, with larger savings as the number of directions grows. On COCO, fQID disagrees with QID on 0.35% of

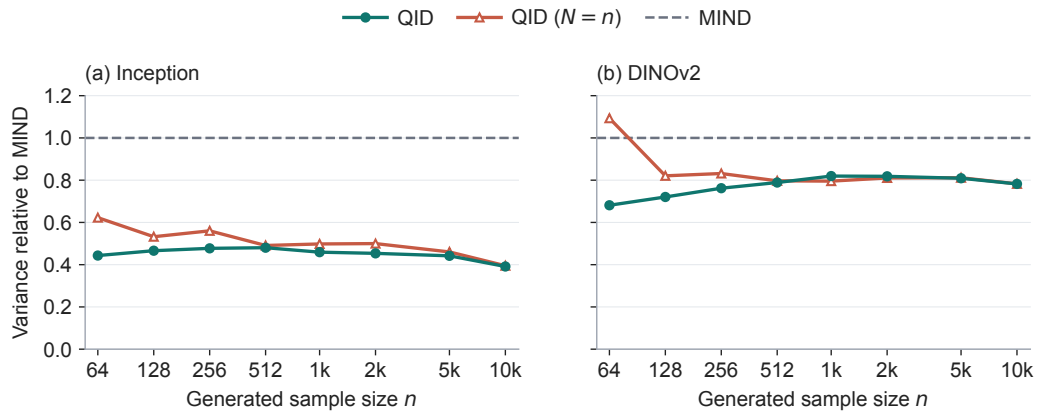

Figure 5: Effect of reference reuse on ImageNet-256. Each curve shows the mean variance over 55 model-score differences divided by MIND's mean variance, using 512 trials. The dashed line denotes equal variance.

model-pair comparisons at 512 images per model, and the two scores select the same model in all 512 trials (Table 19).

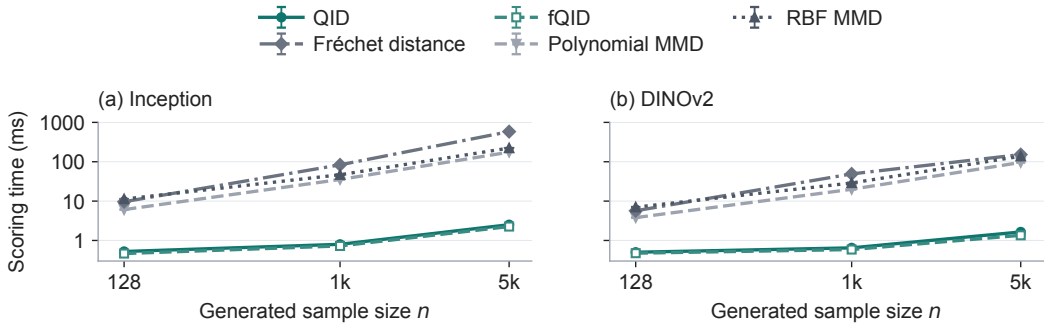

Figure 6: Scoring time from image features with a prepared 100,000-image reference. Points are medians and bars are interquartile ranges over five runs.

Table 2: Computation time and reference storage for 1,000 generated images and 100,000 real images. The lower rows compute QID from sorted projections or prefix sums instead of its interval averages and residual. Times are medians over five GPU runs, with CPU covariance decomposition for Fréchet distance.

	Time (ms)		Reference storage (MiB)	
Method	Inception	DINOv2	Inception	DINOv2
Fréchet distance	83.4	49.1	16	4
Polynomial MMD-U	35.4	20	781	391
RBF MMD-U	46.7	28.9	781	391
QID	0.798	0.649	1.17	0.776
fQID	0.726	0.588	1.57	0.786
QID reference representations				
Sorted projections	1.27	1.07	38.9	38.5
Prefix sums	0.846	0.712	77.1	76.7

5.4 Sensitivity to distribution changes

Following Berthet et al. (2026); Jayasumana et al. (2024), we test whether the metrics detect distributional differences that the mean and covariance cannot distinguish. We construct 4,096 feature vectors with the sample mean and covariance of a 50,000-image COCO reference (Appendix K.1), for which FID is zero despite the change in distribution. QID instead increases from 0.62 for a real-image

batch to 190.65 for the constructed features, and MIND shows the same response (Table 3). The two sets receive similar fQID scores because the approximation retains only the generated mean and coordinate variances. The empirical quantiles distinguish the sets through the remaining differences in their projected distributions.

Table 3: Scores for real-image features and constructed features with the reference's mean and covariance. Each metric retains its own units; the MIND entry reports mean ± SD.

Metric	Real-image features	Constructed features
QID	0.62	190.65
MIND	1.13	192.08 ± 0.58
fQID	4.75	4.62
FID	8.48	≈ 0

We next vary class proportions by small amounts, following Kwon et al. (2026), to test whether the scores track a gradual change in the image distribution. For twenty CIFAR-10 class pairs, we compare a balanced reference with evaluation batches whose first-class proportion increases from 50% to 55%. With CLIP features and 1,000 images per batch, the mean Spearman correlation between scores and class proportion is 0.9788 for QID and 0.9527 for Fréchet distance. Both correlations increase with sample size, reaching 0.9994 and 0.9988 at 5,000 images (Table 14).

6 Related Work

FID compares Gaussian feature approximations, while KID and CMMD use kernel discrepancies in Inception and CLIP space (Heusel et al., 2017; Binkowski et al., 2018; Jayasumana et al., 2024). MIND, APEX and CLIP-SWD compare projected empirical distributions through one-dimensional transport, retaining information beyond Gaussian moments and an efficient computation based on sorting (Berthet et al., 2026; Gallegati et al., 2026; Kwon et al., 2026). QID develops this comparison for repeated evaluation against a common real dataset, storing the quantile coefficients required at each generated batch size.

Optimal transport embeddings express distributions through displacements from a common reference (Wang et al., 2013; Kolouri et al., 2016a). Sliced-Wasserstein kernels use the resulting Hilbert-space geometry, and PSWE uses reference coordinates to pool sets of varying size (Kolouri et al., 2016b; Naderializadeh et al., 2021). FourierSWE obtains finite embeddings by integrating projected quantiles against sampled cosine functions (Amir and Dym, 2025). QID's coordinates are determined by the generated batch size: quantile averages and a residual retain the full-reference score for every batch of that size. The Gaussian approximation in fQID connects this representation to the moment-based approximation of Nadjahi et al. (2021), while retaining the empirical real reference.

7 Discussion and Conclusion

We introduced QID for repeated generative model evaluation against a common real reference. We represent each generated feature distribution by its projected quantile displacements, so the same transport coordinates give its score against the real dataset and distances to other generated distributions. For empirical batches, we project the reference quantiles onto the generated step-function space and retain the squared residual. These quantities preserve the full-reference transport comparison while making the retained representation and subsequent scoring work depend on the generated batch size. This allows us to prepare the reference once and reuse it throughout checkpoint selection and comparisons of sampling procedures, retaining distributional information beyond Gaussian moments.

On ImageNet, QID recovers small-batch rankings more reliably than Fréchet distance in both feature spaces, and reference reuse reduces the variance of model-score differences relative to MIND with resampled references. On COCO, QID and MIND both recover the text-to-image model ordering with high accuracy. These comparisons are inexpensive: scoring speedups exceed two orders of magnitude over FID, and reference storage is substantially smaller than with sorted projections and prefix sums. fQID reduces computation further through a Gaussian approximation, while the controlled experiments show the distributional differences retained by empirical quantiles.

AI USE STATEMENT

We used generative AI tools to assist with research, implementation, analysis and manuscript preparation. The authors take responsibility for the final content.

REFERENCES

Tal Amir and Nadav Dym. Fourier sliced-Wasserstein embedding for multisets and measures. In *International Conference on Learning Representations*, 2025.

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 2605.06797, 2026.

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

Sebastian Bischoff, Alana Darcher, Michael Deistler, Richard Gao, Franziska Gerken, Manuel Gloeckler, Lisa Haxel, Jaivardhan Kapoor, Janne K. Lappalainen, Jakob H. Macke, Guy Moss, Matthijs Pals, Felix C. Pei, Rachel Rapp, A. Erdem Sağtekin, Cornelius Schröder, Auguste Schulz, Zinovia Stefanidi, Shoji Toyota, Linda Ulmer, and Julius Vetter. A practical guide to sample-based statistical distances for evaluating generative models in science. *Transactions on Machine Learning Research*, 2024.

Clément Bonet, Laetitia Chapel, Lucas Drumetz, and Nicolas Courty. Hyperbolic sliced-Wasserstein via geodesic and horospherical projections. In *Proceedings of the 2nd Annual Workshop on Topology, Algebra, and Geometry in Machine Learning*, volume 221 of *Proceedings of Machine Learning Research*, pages 334–370. PMLR, 2023.

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

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In *CVPR*, 2022.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception Score and where to find them. In *CVPR*, 2020.

Eustasio del Barrio, Juan A. Cuesta-Albertos, Carlos Matrán, and Jesús M. Rodríguez-Rodríguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936-970, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

Ishan Deshpande, Ziyu Zhang, and Alexander G. Schwing. Generative modeling using the sliced Wasserstein distance. In *CVPR*, 2018.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *NeurIPS*, 2021.

Caterina Gallegati, Monica Bianchini, Franco Scarselli, Vittorio Murino, and Barbara Toniella Corradini. APEX: Assumption-free Projection-based Embedding eXamination Metric for Image Quality Assessment. arXiv preprint 2605.07786, 2026.

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 *NeurIPS*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 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 *CVPR*, 2024.

Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up GANs for text-to-image synthesis. In *CVPR*, 2023.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Soheil Kolouri, Akif B. Tosun, John A. Ozolek, and Gustavo K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453-462, 2016.

Soheil Kolouri, Yang Zou, and Gustavo K. Rohde. Sliced Wasserstein kernels for probability distributions. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepčev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43-59, 2017.

Soheil Kolouri, Phillip E. Pope, Charles E. Martin, and Gustavo K. Rohde. Sliced Wasserstein Auto-Encoders. In *International Conference on Learning Representations*, 2019.

Soheil Kolouri, Navid Naderializadeh, Gustavo K. Rohde, and Heiko Hoffmann. Wasserstein embedding for graph learning. In *International Conference on Learning Representations*, 2021.

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

Hyeok Kyu Kwon, Jaeseung Yang, and Minwoo Chae. Evaluating image generation models via sliced Wasserstein distance. *Journal of the Korean Statistical Society*, 2026.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *NeurIPS*, 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*, 2023.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *CVPR*, 2022.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014.

Yuzhe Lu, Xinran Liu, Andrea Soltoggio, and Soheil Kolouri. SLoSH: Set locality sensitive hashing via sliced-Wasserstein embeddings. In *WACV*, 2024.

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

Navid Naderializadeh, Joseph F. Comer, Reed W. Andrews, Heiko Hoffmann, and Soheil Kolouri. Pooling by sliced-Wasserstein embedding. In *NeurIPS*, 2021.

Kimia Nadjahi, Alain Durmus, Lénaïc Chizat, Soheil Kolouri, Shahin Shahrampour, and Umut Şimşekli. Statistical and topological properties of sliced probability divergences. In *NeurIPS*, 2020.

Kimia Nadjahi, Alain Durmus, Pierre E. Jacob, Roland Badeau, and Umut Şimşekli. Fast approximation of the sliced-Wasserstein distance using concentration of random projections. In *NeurIPS*, 2021.

Khai Nguyen and Nhat Ho. Amortized projection optimization for sliced Wasserstein generative models. In *NeurIPS*, 2022.

Khai Nguyen and Nhat Ho. Sliced Wasserstein estimation with control variates. In *International Conference on Learning Representations*, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *CVPR*, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Gabriel Peyré and Marco Cuturi. *Computational Optimal Transport: With Applications to Data Science*. Foundations and Trends in Machine Learning, 11(5-6):355-607, 2019.

Moritz Piening and Robert Beinert. Slicing the Gaussian mixture Wasserstein distance. *Transactions on Machine Learning Research*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024.

Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, 2011.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In *NeurIPS*, 2018.

Axel Sauer, Katja Schwarz, and Andreas Geiger. StyleGAN-XL: Scaling StyleGAN to large diverse datasets. In *ACM SIGGRAPH Conference Proceedings*, 2022.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *ECCV*, 2024.

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

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

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *CVPR*, 2016.

Huy Tran, Yikun Bai, Abihith Kothapalli, Ashkan Shahbazi, Xinran Liu, Rocío Díaz Martin, and Soheil Kolouri. Stereographic spherical sliced Wasserstein distances. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pages 48494–48564. PMLR, 2024.

Lokesh Veeramacheneni, Moritz Wolter, Hildegard Kuehne, and Juergen Gall. Fréchet Wavelet Distance: A Domain-Agnostic Metric for Image Generation. In *International Conference on Learning Representations*, 2025.

Cédric Villani. *Optimal Transport: Old and New*. Springer, 2009.

Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254-269, 2013.

Chuang Xu and Arno Berger. Best finite constrained approximations of one-dimensional probabilities. *Journal of Approximation Theory*, 244:1–36, 2019.

A Notation and Conventions

Distributions are feature-space laws under a fixed map ψ, and hats denote empirical measures. We use n for a generated or calibration batch size, N for the empirical reference size, and M for the number of slices. Parentheses denote sorted one-dimensional samples. The reference measure and directions define the score; the coefficients below store its finite-dimensional representation at a chosen n.

Symbol	Definition
Data and projected transport	
$\mathcal{X}$, $\psi : \mathcal{X} \to \mathbb{R}^d$	Image space and fixed feature extractor.
P_{data}, p_{data}	Population image law and its feature-space law.
g_θ, p_θ	Generator and generated feature-space law.
$D_{\mathrm{ref}}, D_{\mathrm{cal}}, D_{\mathrm{test}}$	Reference, calibration, and held-out real-data splits.
$\hat{p}_{n,\theta}$, $\hat{p}_{N,\mathrm{ref}}$	Empirical generated and reference laws.
$h_i = \psi(x_i)$, $z_j = \psi(y_j)$	Generated and reference feature vectors.
$U = (u_1, \dots, u_M)$	Fixed directions $u_m \in \mathbb{S}^{d-1}$.
$u_m^\top p$, $Q_{p,m}$	Projected law and its quantile function.
W_2^2	Squared one-dimensional 2-Wasserstein distance.
$\mathrm{SW}_2^2(p, q)$	$\mathbb{E}_{u\sim\mathcal{U}(\mathbb{S}^{d-1})}[W_2^2(u^\top p, u^\top q)]$.
$d_U(p, q)$	$\big(M^{-1}\sum_m W_2^2(u_m^\top p, u_m^\top q)\big)^{1/2}$.
α	Positive score scale, fixed within a study.
Quantile coordinates and reference preparation	
$r_{m,1:N}$, g_m	Sorted reference and generated projections on slice m.
$\beta_{m,i}^{(n)}$, $c_m^{(n)}$	Reference quantile average on the ith target interval and squared projection residual.
$C_{R,U,n}$	Total scaled residual $\alpha M^{-1}\sum_m c_m^{(n)}$.
$\mathcal{C}_{R,U,n}$	Compact state $(U, \{\beta_{m,i}^{(n)}\}_{m,i}, C_{R,U,n})$ at the fixed scale α.
$R_{m,k}, S_{m,k}$	Prefix sums of $r_{m,j}$ and $r_{m,j}^2$, zero at $k = 0$.
$\mathcal{R}_{N,M}$	Optional preparation workspace $\{u_m, r_{m,1:N}, R_{m,0:N}, S_{m,0:N}\}_{m=1}^M$.
$\mathrm{QID}_{n,N,M}$	$\alpha d_U(\hat{p}_{n,\theta}, \hat{p}_{N,\mathrm{ref}})^2$.
$\Phi_{p_{\mathrm{ref}},U}$, Ψ_i	Continuous reference displacement coordinates and finite coordinates of batch i.
Gaussian approximation and calibration	
Φ, ϕ	Standard normal cdf and density.
μ_G, v_G	Generated feature mean and coordinate variances.
a_m, b_m	Projected mean and scale of the generated diagonal Gaussian.
$\bar{r}_m$, q_m, γ_m	Reference mean, second moment, and inner product with the normal quantile.
$\omega_{j,N}$	$\int_{(j-1)/N}^{j/N} \Phi^{-1}(t)\mathrm{d}t$.
$\mathrm{fQID}_{n,N,M}$	Score obtained by replacing the generated law with its diagonal Gaussian approximation.
B_ℓ, L	Real calibration batch and number of calibration batches.
s_ℓ, s_θ	Calibration and evaluated scores.
$\hat{\mu}_{0,n}, \hat{\sigma}_{0,n}, q_{0,n}(\tau)$	Estimated real-data null mean, standard deviation, and quantile at size n.
$\mathrm{cQID}_{\mathrm{ctr}}, \mathrm{cQID}_z$	Null-centered and standardized scores.
η, $\hat{p}_{\mathrm{rank}}$	Nominal test level and rank p-value.
N_{ref}, K	Reference size and number of repeated evaluations in runtime tables.

B Quantile Transport Coordinates

B.1 A common reference in quantile space

Fix $p, p_{\mathrm{ref}} \in \mathcal{P}_2(\mathbb{R}^d)$ and unit directions $U = (u_1, \dots, u_M)$. On slice m, write $Q_{p,m} = F^{-1}_{u_m^\top p}$ and $Q_{R,m} = F^{-1}_{u_m^\top p_{\mathrm{ref}}}$. If T is uniform on $(0, 1)$, then $(Q_{R,m}(T), Q_{p,m}(T))$ is an optimal coupling of the two projected laws (Villani, 2009; Peyré and Cuturi, 2019). The displacement $Q_{p,m}(t) - Q_{R,m}(t)$ therefore describes transport from the reference at each mass coordinate t. This quantile construction applies to discrete and continuous reference laws alike.

The coordinates in Eq. (7) concatenate these displacements across slices, as in sliced transport embeddings (Wang et al., 2013; Kolouri et al., 2016b). Their ambient space is $L^2(\{1,\dots,M\}\times(0,1))$, with counting measure over the slices and Lebesgue measure over t. On each slice,

$$\int_0^1 |Q_{p,m}(t)-Q_{R,m}(t)|^2\,\mathrm{d}t = W_2^2(u_m^\top p, u_m^\top p_{\mathrm{ref}}). \tag{B.1}$$

Multiplication by α/M and summation establish the norm identity in Eq. (8). For two target laws, subtraction gives

$$\Phi_{p_{\mathrm{ref}},U}(p)(m,t)-\Phi_{p_{\mathrm{ref}},U}(q)(m,t)=\sqrt{\frac{\alpha}{M}}\big(Q_{p,m}(t)-Q_{q,m}(t)\big), \tag{B.2}$$

so that

$$\|\Phi_{p_{\mathrm{ref}},U}(p)-\Phi_{p_{\mathrm{ref}},U}(q)\|_{L^2}^2=\frac{\alpha}{M}\sum_{m=1}^{M}W_2^2(u_m^\top p,u_m^\top q)=\alpha d_U(p,q)^2. \tag{B.3}$$

Thus the reference sets the origin of the coordinate system, while coordinate differences retain the transport cost between targets. The distance d_U is a pseudometric: the triangle inequality follows from the L^2 norm, and zero distance means that all the selected projected laws coincide.

B.2 Finite-dimensional coordinates

Set $p_{\mathrm{ref}}=\hat{p}_{N,\mathrm{ref}}$. For empirical target laws with n equally weighted atoms, every projected quantile belongs to the space H_n of functions that are constant on $I_k=((k-1)/n,k/n]$. The orthogonal projection of the reference quantile onto this space is

$$Q_{R,m}^{(n)}(t)=\beta_{m,k}^{(n)}\quad(t\in I_k),\qquad \beta_{m,k}^{(n)}=n\int_{I_k}Q_{R,m}(t)\mathrm{d}t. \tag{B.4}$$

Since $Q_{R,m}-Q_{R,m}^{(n)}$ is orthogonal to H_n, Pythagoras gives, for the kth sorted projection $g_{i,m,(k)}$ of batch i,

$$W_2^2(u_m^\top\hat{p}_{n,i},u_m^\top\hat{p}_{N,\mathrm{ref}})=\frac{1}{n}\sum_{k=1}^{n}\big(g_{i,m,(k)}-\beta_{m,k}^{(n)}\big)^2+c_m^{(n)}. \tag{B.5}$$

Here $c_m^{(n)}=\|Q_{R,m}-Q_{R,m}^{(n)}\|_{L^2}^2$ is the reference projection residual. The bin averages are nondecreasing, so they also define the optimal equal-weight n-point approximation of the reference on this slice (Xu and Berger, 2019, Theorem 5.5 and Remark 5.6(ii)).

The finite coordinates Ψ_i of Eq. (12) collect the scaled differences $g_{i,m,(k)}-\beta_{m,k}^{(n)}$ and one final coordinate $\sqrt{C_{R,U,n}}$, where $C_{R,U,n}=\alpha M^{-1}\sum_m c_m^{(n)}$. Summing the preceding decomposition gives

$$\|\Psi_i\|_2^2=\mathrm{QID}_{n,N,M}(\hat{p}_{n,i};\hat{p}_{N,\mathrm{ref}},U). \tag{B.6}$$

For two batches of the same size, the coefficients and the final coordinate cancel, yielding

$$\|\Psi_i-\Psi_j\|_2^2=\frac{\alpha}{Mn}\sum_{m=1}^{M}\sum_{k=1}^{n}\big(g_{i,m,(k)}-g_{j,m,(k)}\big)^2=\alpha d_U(\hat{p}_{n,i},\hat{p}_{n,j})^2. \tag{B.7}$$

For target batches of unequal sizes, take the common refinement of their quantile partitions. Project each reference quantile onto the functions that are constant on this partition and retain its squared projection residual. For an interval J, use the reference average on J and scale its displacement coordinate by $\sqrt{\alpha|J|/M}$. Adjoining the square root of the total scaled residual gives finite coordinates on this common partition. The same orthogonality and cancellation arguments preserve their reference scores and pairwise transport costs.

C Reference Preparation and Repeated Scoring

For an empirical reference with features $z_j=\psi(y_j)$, the projected quantile has values

$$(r_{m,1},\dots,r_{m,N})=\mathrm{sort}(u_m^\top z_{1:N}),\qquad r_{m,1}\le\cdots\le r_{m,N}. \tag{C.1}$$

These values determine the coefficients of Proposition 3.1. Algorithm 1 prepares them once at the chosen batch size and scores each generated batch by a sum of squared coordinate differences plus the reference residual.

Algorithm 1 QID reference preparation and repeated scoring

Require: Reference features $z_{1:N}$, unit directions U, target batch size n, scale $\alpha > 0$

$C \leftarrow 0$
for $m = 1, \dots, M$ **do**
  $r_m \leftarrow \mathrm{sort}(u_m^\top z_{1:N})$
  Compute $\beta_{m,i}^{(n)} = n \int_{(i-1)/n}^{i/n} Q_{R,m}(t)\mathrm{d}t$ for $i = 1, \dots, n$
  $c_m^{(n)} \leftarrow \int_0^1 |Q_{R,m}(t) - Q_{R,m}^{(n)}(t)|^2 \mathrm{d}t$
  $C \leftarrow C + \alpha c_m^{(n)}/M$
Store $\mathcal{C}_{R,U,n} = (U, \{\beta_{m,i}^{(n)}\}_{m,i}, C)$ at scale α
for each generated feature batch $h_{1:n}$ **do**
  $s \leftarrow C$
  for $m = 1, \dots, M$ **do**
    $g_m \leftarrow \mathrm{sort}(u_m^\top h_{1:n})$
    $s \leftarrow s + \alpha(Mn)^{-1} \sum_{i=1}^n (g_{m,i} - \beta_{m,i}^{(n)})^2$
  Output $s = \mathrm{QID}_{n,N,M}(\hat{p}_{n,\theta}; \hat{p}_{N,\mathrm{ref}}, U)$

C.1 Interval integration

The empirical generated and reference quantiles partition $(0, 1)$ into n and N intervals. Their squared transport cost is

$$W_2^2(\hat{p}_n, \hat{q}_N) = \int_0^1 |\hat{Q}_{x,n}(t) - \hat{Q}_{r,N}(t)|^2 \mathrm{d}t. \tag{C.2}$$

Merging the breakpoints $\{i/n\}_{i=0}^n \cup \{j/N\}_{j=0}^N$ integrates this piecewise constant function. Prefix sums offer a second way to compute the interval integrals, particularly when coefficients are needed for several batch sizes. Define

$$R_{m,k} = \sum_{j \le k} r_{m,j}, \qquad S_{m,k} = \sum_{j \le k} r_{m,j}^2, \qquad R_{m,0} = S_{m,0} = 0, \tag{C.3}$$

and collect the sorted values and prefixes in the preparation workspace

$$\mathcal{R}_{N,M} = \{u_m, r_{m,1:N}, R_{m,0:N}, S_{m,0:N}\}_{m=1}^M. \tag{C.4}$$

Each target interval contains at most two partial reference bins, with complete bins between them. The complete-bin contributions follow from prefix differences, as shown in Lemma C.1.

Lemma C.1 (Reference interval integrals)**.** *Let $x_1 \le \cdots \le x_n$ and $r_1 \le \cdots \le r_N$ be sorted empirical samples on $\mathbb{R}$. For interval i, define*

$$a_i = \frac{i-1}{n}, \qquad b_i = \frac{i}{n}, \qquad s_i = \lfloor N a_i \rfloor + 1, \qquad e_i = \lceil N b_i \rceil. \tag{C.5}$$

With empty prefix ranges contributing zero, its reference first-moment integral is

$$A_i = \begin{cases} (b_i - a_i) r_{s_i}, & s_i = e_i, \\ \left(\frac{s_i}{N} - a_i\right) r_{s_i} + \frac{R_{e_i-1} - R_{s_i}}{N} + \left(b_i - \frac{e_i - 1}{N}\right) r_{e_i}, & s_i < e_i. \end{cases} \tag{C.6}$$

Its second-moment integral B_i is obtained by replacing r_j by r_j^2 and R_k by S_k. Consequently,

$$W_2^2\left(\frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \frac{1}{N}\sum_{j=1}^N \delta_{r_j}\right) = \sum_{i=1}^n \left(\frac{x_i^2}{n} - 2x_i A_i + B_i\right). \tag{C.7}$$

Proof. The generated quantile is constant at x_i on $(a_i, b_i]$. Intersecting this interval with the reference bins gives one partial bin at each end and full bins between them, or a single partial bin when $s_i = e_i$. Their lengths and reference values give $A_i = \int_{a_i}^{b_i} \hat{Q}_{r,N}(t)\mathrm{d}t$ and $B_i = \int_{a_i}^{b_i} \hat{Q}_{r,N}(t)^2 \mathrm{d}t$. Expanding the squared residual yields

$$\int_{a_i}^{b_i} \left(x_i - \hat{Q}_{r,N}(t)\right)^2 \mathrm{d}t = \frac{x_i^2}{n} - 2x_i \int_{a_i}^{b_i} \hat{Q}_{r,N}(t)\mathrm{d}t + \int_{a_i}^{b_i} \hat{Q}_{r,N}(t)^2 \mathrm{d}t. \tag{C.8}$$

Summing over the generated intervals proves the formula. □

The coefficients follow from $\beta_i^{(n)} = nA_i$ and $c^{(n)} = \sum_i (B_i - nA_i^2)$. Algorithm 2 performs this preparation on one slice. Breakpoints can be ordered by the integer products iN and jn.

Algorithm 2 Reference coefficients from prefix sums on one slice

Require: Sorted reference values $r_{1:N}$, prefixes $R_{0:N}, S_{0:N}$, target size n

$c \leftarrow 0$
for $i = 1, \dots, n$ **do**
  $a \leftarrow (i-1)/n$, $b \leftarrow i/n$, $s \leftarrow \lfloor Na \rfloor + 1$, $e \leftarrow \lceil Nb \rceil$
  if $s = e$ **then**
    $A \leftarrow (b-a) r_s$, $B \leftarrow (b-a) r_s^2$
  else
    $A \leftarrow (s/N - a) r_s + (R_{e-1} - R_s)/N + (b - (e-1)/N) r_e$
    $B \leftarrow (s/N - a) r_s^2 + (S_{e-1} - S_s)/N + (b - (e-1)/N) r_e^2$
  $\beta_i \leftarrow nA$, $c \leftarrow c + B - nA^2$
return $\beta_{1:n}, c$

With $q_m = N^{-1} \sum_j r_{m,j}^2$, the residual also equals $c_m^{(n)} = q_m - n^{-1} \sum_i (\beta_{m,i}^{(n)})^2$. Completing the square in Lemma C.1 gives

$$W_2^2(\hat{p}_{n,m}, \hat{p}_{N,m}) = \frac{1}{n} \sum_{i=1}^{n} \big(x_{m,(i)} - \beta_{m,i}^{(n)}\big)^2 + c_m^{(n)}, \qquad c_m^{(n)} \geq 0. \tag{C.9}$$

The nonnegativity follows from the orthogonal projection in Appendix B.2. At $n = N$, the coefficients are the sorted reference values and the residual vanishes, giving MIND's empirical score.

Reference projection and sorting take $O(MNd + MN \log N)$ operations. Prefix construction takes $O(MN)$ operations, after which each additional batch size requires $O(Mn)$ coefficient preparation. The prefix workspace occupies $O(MN)$ scalar storage, and the compact state occupies $O(Mn)$, in addition to the $O(Md)$ directions. Scoring a generated feature batch takes $O(Mnd + Mn \log n)$ for projection and sorting and $O(Mn)$ for the coefficient reduction. Computing the interval integrals from prefix sums also takes $O(Mn)$ operations after sorting, while QID's coefficients remove the repeated interval lookups. Appendix M compares their preparation, storage, and scoring costs.

Supplementary code. The anonymous code supplement includes NumPy and PyTorch implementations of QID and fQID, an example COCO reference, and generated features from six configurations evaluated on the same prompts. A verification script reproduces their saved scores.

D Statistical Properties

D.1 Finite-reference approximation

For the fixed directions U, the projected distance is

$$d_U(p, q) = \left(\frac{1}{M} \sum_{m=1}^{M} W_2^2(u_m^\top p, u_m^\top q) \right)^{1/2}. \tag{D.1}$$

Its triangle inequality gives, for the population reference p_{data} and empirical reference $\hat{p}_{N,\mathrm{ref}}$,

$$|d_U(p_\theta, p_{\mathrm{data}}) - d_U(p_\theta, \hat{p}_{N,\mathrm{ref}})| \leq d_U(p_{\mathrm{data}}, \hat{p}_{N,\mathrm{ref}}). \tag{D.2}$$

The right-hand side controls the reference approximation for every target p_θ. Under iid reference sampling with finite second moments, it converges to zero almost surely.

D.2 Joint consistency

Let $P = p_\theta$ and $Q = p_{\mathrm{data}}$ belong to $\mathcal{P}_2(\mathbb{R}^d)$, and form P_n, Q_N from iid sample sequences. For any bank $U_M = (u_1, \dots, u_M)$, the triangle inequality and contraction under unit projections yield

$$\begin{aligned} |d_{U_M}(P_n, Q_N) - d_{U_M}(P, Q)| &\leq d_{U_M}(P_n, P) + d_{U_M}(Q_N, Q) \\ &\leq W_2(P_n, P) + W_2(Q_N, Q). \end{aligned} \tag{D.3}$$

Empirical Wasserstein convergence sends the final bound to zero almost surely, uniformly in M (Peyré and Cuturi, 2019). Now let U_M contain the first M directions of one iid uniform sequence. Since $W_2^2(u^\top P, u^\top Q) \leq W_2^2(P, Q) < \infty$ for every unit direction, the strong law gives

$$d_{U_M}(P, Q)^2 \longrightarrow \mathrm{SW}_2^2(P, Q) \qquad \text{almost surely as } M \to \infty. \tag{D.4}$$

Combining the two limits and squaring proves $\alpha d_{U_M}(P_n, Q_N)^2 \to \alpha \mathrm{SW}_2^2(P, Q)$ almost surely as $n, N, M \to \infty$, at arbitrary relative rates.

D.3 Reference sampling and score variance

Let P_n, Q_n be empirical generated and real batches, and fix U. Assuming that

$$S(P_n, Q_n, U) = \frac{\alpha}{M} \sum_{m=1}^{M} W_2^2(u_m^\top P_n, u_m^\top Q_n) \tag{D.5}$$

is square-integrable, the law of total variance gives

$$\mathrm{Var}_{P_n, Q_n}[S(P_n, Q_n, U)] = \mathbb{E}_{Q_n}[\mathrm{Var}_{P_n}(S(P_n, Q_n, U) \mid Q_n)] + \mathrm{Var}_{Q_n}(\mathbb{E}_{P_n}[S(P_n, Q_n, U) \mid Q_n]). \tag{D.6}$$

The first term is the average conditional variance over real batches. The second is the variation of the conditional mean as the real batch changes. Once $Q_N = q_N$ is fixed, repeated evaluation has variance

$$\mathrm{Var}_{P_n}[S(P_n, q_N, U) \mid q_N, U], \tag{D.7}$$

which depends on the chosen reference. The decomposition separates the variation over generated batches from the variation introduced by resampling real data.

D.4 Paired model comparisons

Model ranking compares squared norms relative to the reference. For laws $p, q, r \in \mathcal{P}_2(\mathbb{R}^d)$, write $S_r(p) = \alpha d_U(p, r)^2$ and $\Delta_r(p, q) = S_r(p) - S_r(q)$. Sharing a reference across the two scores cancels its squared-quantile term, giving the following identities.

Proposition D.1. *Fix U and $p, q, r, s \in \mathcal{P}_2(\mathbb{R}^d)$, and let $Q_{p,m} = F^{-1}_{u_m^\top p}$. Then*

$$\Delta_r(p, q) = \frac{\alpha}{M} \sum_{m=1}^{M} \int_0^1 (Q_{p,m} - Q_{q,m})(Q_{p,m} + Q_{q,m} - 2Q_{r,m}) \mathrm{d}t, \tag{D.8}$$

$$\Delta_r(p, q) - \Delta_s(p, q) = -\frac{2\alpha}{M} \sum_{m=1}^{M} \int_0^1 (Q_{p,m} - Q_{q,m})(Q_{r,m} - Q_{s,m}) \mathrm{d}t. \tag{D.9}$$

Proof. Expand the two squared quantile residuals in $S_r(p) - S_r(q)$ and cancel $Q_{r,m}^2$. Factoring gives the first identity, and subtracting the expressions for r and s gives the second. The products are integrable because all four laws have finite second moments. □

The reference enters a model score gap through the difference between the two generated quantile functions. A real batch shared within a trial, as in MIND, already cancels the squared reference term. Keeping the same empirical reference across repeated evaluations fixes the remaining $Q_{r,m}$ factor. The second identity then quantifies how choosing another reference can change the gap.

E The Gaussian Approximation fQID

E.1 Derivation and computation

fQID replaces the generated feature law by its diagonal Gaussian approximation. With Φ and ϕ denoting the standard normal cdf and density, the mean and coordinate variances are

$$\mu_G = \frac{1}{n} \sum_{i=1}^{n} h_i, \qquad v_G = \frac{1}{n} \sum_{i=1}^{n} (h_i - \mu_G)^{\odot 2}. \tag{E.1}$$

The corresponding projected quantile is

$$Q_{G,m}(t) \approx a_m + b_m \Phi^{-1}(t), \qquad a_m = u_m^\top \mu_G, \qquad b_m^2 = (u_m^{\odot 2})^\top v_G. \tag{E.2}$$

For the empirical reference, $Q_{R,m}(t) = r_{m,j}$ on $((j-1)/N, j/N]$. Its contribution to the Gaussian score is determined by

$$\bar{r}_m = \frac{1}{N}\sum_{j=1}^{N} r_{m,j}, \qquad q_m = \frac{1}{N}\sum_{j=1}^{N} r_{m,j}^2, \tag{E.3}$$

and its inner product with the normal quantile,

$$\gamma_m = \sum_{j=1}^{N} \omega_{j,N} r_{m,j}, \qquad \omega_{j,N} = \int_{(j-1)/N}^{j/N} \Phi^{-1}(t)\mathrm{d}t. \tag{E.4}$$

The antiderivative $\int \Phi^{-1}(t)\mathrm{d}t = -\phi(\Phi^{-1}(t))$, whose endpoint limits are zero, gives

$$\omega_{j,N} = \phi\left(\Phi^{-1}\left(\frac{j-1}{N}\right)\right) - \phi\left(\Phi^{-1}\left(\frac{j}{N}\right)\right). \tag{E.5}$$

Expanding the squared residual and using $\int_0^1 \Phi^{-1}(t)\mathrm{d}t = 0$ and $\int_0^1 (\Phi^{-1}(t))^2\mathrm{d}t = 1$ yields

$$\begin{aligned} W_{m,\mathrm{fast}}^2 &= \int_0^1 \left(a_m + b_m\Phi^{-1}(t) - Q_{R,m}(t)\right)^2\mathrm{d}t \\ &= a_m^2 + b_m^2 + q_m - 2a_m\bar{r}_m - 2b_m\gamma_m. \end{aligned} \tag{E.6}$$

Averaging over slices gives

$$\mathrm{fQID}_{n,N,M} = \frac{\alpha}{M}\sum_{m=1}^{M} W_{m,\mathrm{fast}}^2. \tag{E.7}$$

Thus preparation stores three reference scalars per slice. Each generated batch requires its mean and coordinate variances, their projections, and the preceding formula, for $O(nd + Md + M)$ operations. Under the isotropic choice $v_G = (\rho_G^2/d)\mathbf{1}$, where $\rho_G^2 = n^{-1}\sum_i \|h_i - \mu_G\|^2$, every unit direction has $b_m = \rho_G/\sqrt{d}$. This is the generated-side scale of Nadjahi et al. (2021), who approximate both centered distributions by isotropic Gaussians. fQID retains the empirical reference quantile and allows its generated scale to vary across directions.

E.2 Mean transport and approximation error

Write the empirical quantiles on one slice as $Q_G = a + \widetilde{Q}_G$ and $Q_R = \bar{r} + \widetilde{Q}_R$, with centered quantiles integrating to zero. Let $s^2 = \int \widetilde{Q}_G^2$, $\sigma_R^2 = q - \bar{r}^2$, and $c = \int \widetilde{Q}_G\widetilde{Q}_R$. The empirical cost is

$$E = (a-\bar{r})^2 + s^2 + \sigma_R^2 - 2c. \tag{E.8}$$

A Gaussian approximation using the empirical projected scale s has cost $H = (a-\bar{r})^2 + s^2 + \sigma_R^2 - 2s\gamma$. Writing F for the fQID slice cost, subtraction gives

$$F - E = \underbrace{(b-s)(b+s-2\gamma)}_{\text{omitted covariance}} + \underbrace{2(c-s\gamma)}_{\text{Gaussian quantile approximation}}. \tag{E.9}$$

The two terms measure the diagonal covariance approximation and the replacement of the empirical quantile by a Gaussian quantile at scale s. Their signs may differ, so their errors can cancel. Both scores contain the same mean transport term $A_U = \alpha M^{-1}\sum_m [u_m^\top(\mu_G - \mu_R)]^2$, whose expectation under uniform directions is $\alpha\|\mu_G - \mu_R\|^2/d$. The centered-feature experiments in Appendix L remove this mean term before comparing the remaining components.

E.3 Projection onto Gaussian quantiles

A Gaussian-reference control using the same generated mean a and scale b has cost $G = (a-\bar{r})^2 + (b-\sigma_R)^2$. Cauchy–Schwarz gives $F - G = 2b(\sigma_R - \gamma) \geq 0$. Completing the square also gives

$$F = (a-\bar{r})^2 + (b-\gamma)^2 + \rho_R^2, \qquad \rho_R^2 = \sigma_R^2 - \gamma^2 \geq 0. \tag{E.10}$$

The function $\bar{r} + \gamma\Phi^{-1}$ is the orthogonal projection of Q_R onto $\mathrm{span}\{1, \Phi^{-1}\}$. Since both quantiles are nondecreasing, $\gamma \geq 0$, so the projection defines the W_2-closest Gaussian $\mathcal{N}(\bar{r}, \gamma^2)$ to the reference slice, with squared approximation error ρ_R^2 (del Barrio et al., 1999). This projection parallels the equal-bin construction for QID, but here the generated quantile is also replaced by a Gaussian. The resulting score is determined by the generated mean and coordinate variances.

For example, take $R = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_1$ in one dimension and $\alpha = 1$. Its summaries are $\bar{r} = 0$, $\sigma_R = 1$, and $\gamma = \sqrt{2/\pi}$. The population fast score of R against itself is $2 - 2\sqrt{2/\pi} \approx 0.4042$. The Gaussian $P_* = \mathcal{N}(0, 2/\pi)$ has the smaller fast score $1 - 2/\pi \approx 0.3634$, whereas $W_2^2(R,R) = 0$ and $W_2^2(P_*, R) = 1 - 2/\pi$. Thus the Gaussian approximation changes self-comparison and can change rankings.

F CALIBRATION

F.1 REAL-DATA SCORE ADJUSTMENT

Calibration compares a generated score with the variation obtained from real batches of the same size. We construct the reference from D_{ref}, draw L calibration batches B_ℓ from D_{cal}, and reserve D_{test} for evaluation. The calibration batches follow the generated-batch sampling rule. Their scores $s_\ell = \mathrm{QID}_{n,N,M}(\hat{p}_{B_\ell}; \hat{p}_{N,\mathrm{ref}}, U)$ determine the null mean $\hat{\mu}_{0,n}$, standard deviation $\hat{\sigma}_{0,n}$, and quantile $q_{0,n}(\tau)$. The centered and standardized scores are

$$\begin{aligned} \mathrm{cQID}_{\mathrm{ctr}\,n,N,M} &= \mathrm{QID}_{n,N,M}(\hat{p}_{n,\theta}; \hat{p}_{N,\mathrm{ref}}, U) - \hat{\mu}_{0,n}, \\ \mathrm{cQID}_{z\,n,N,M} &= \frac{\mathrm{QID}_{n,N,M}(\hat{p}_{n,\theta}; \hat{p}_{N,\mathrm{ref}}, U) - \hat{\mu}_{0,n}}{\hat{\sigma}_{0,n} + \varepsilon}. \end{aligned} \tag{F.1}$$

Null-mean subtraction and standardization are affine transformations of the score. For fixed $\varepsilon > 0$, reference, calibration data, and n, both preserve rankings. Subtraction leaves score variance unchanged, and standardization multiplies it by $(\hat{\sigma}_{0,n} + \varepsilon)^{-2}$.

F.2 RANK TEST

Conditional on the empirical reference, feature map, and projection bank, suppose that the null evaluation score s_θ and calibration scores $s_1, \ldots, s_L$ are jointly exchangeable. Then

$$\hat{p}_{\mathrm{rank}} = \frac{1 + \sum_{\ell=1}^{L} \mathbf{1}\{s_\ell \geq s_\theta\}}{L+1} \tag{F.2}$$

satisfies

$$\Pr(\hat{p}_{\mathrm{rank}} \leq \eta) \leq \frac{\lfloor (L+1)\eta \rfloor}{L+1} \leq \eta \tag{F.3}$$

under this conditioning (Lei et al., 2018). Counting ties with $\geq$ makes the rule conservative. Equivalently, the test rejects scores strictly above the $\lceil (L+1)(1-\eta) \rceil$th calibration order statistic, with an infinite threshold if the index exceeds L. Applying this rule to fQID uses fast scores throughout and its own calibration distribution.

F.3 CALIBRATION TRANSFER ON CIFAR-10

To study transfer of a fitted threshold between real-data pools, we use the first 40,000 CIFAR-10 training images for the reference, the remaining 10,000 for calibration, and the official test set for evaluation. For each $n \in \{64, 128, 256, 512, 1000\}$, we draw $L = 400$ calibration batches independently and score the 512 paired test batches of Appendix K. Sampling is without replacement within a batch, with overlap allowed across batches. At $\eta = 0.05$, the threshold is the 381st ordered calibration score. The 95% Clopper–Pearson intervals describe rejection probability over independent draws from the fixed test pool, conditional on the fitted threshold. The rank guarantee instead averages over jointly exchangeable calibration and test scores.

Table 4: False-positive rates (%) at nominal level 0.05, with 95% Clopper–Pearson intervals over 512 held-out CIFAR-10 batches. Each method and batch size uses 400 calibration batches. All methods receive the same torchvision Inception features, real reference and test-image subsets.

n	QID	fQID	Fréchet	RBF-MMD U
64	4.10 (2.56–6.20)	3.91 (2.40–5.97)	7.42 (5.31–10.05)	3.52 (2.10–5.50)
128	8.20 (5.98–10.93)	6.05 (4.15–8.48)	9.38 (6.99–12.24)	3.52 (2.10–5.50)
256	5.47 (3.66–7.81)	5.27 (3.50–7.58)	6.05 (4.15–8.48)	2.93 (1.65–4.79)
512	3.71 (2.25–5.73)	4.30 (2.71–6.43)	9.38 (6.99–12.24)	2.15 (1.08–3.81)
1000	6.25 (4.31–8.71)	6.05 (4.15–8.48)	13.48 (10.64–16.74)	0.98 (0.32–2.26)

Table 4 reports the rejection rates obtained with 400 calibration batches for all four methods. At $n = 128$, the fitted QID threshold rejects 8.20% of clean test batches (95% CI 5.98–10.93%), above the nominal 5%. To examine variation across fitted thresholds, we repeat calibration with twenty independent sets of 400 batches per size and evaluate every threshold on the same 512 test batches. At $n = 128$, these QID thresholds reject between 26 and 39 clean batches. Pooling the twenty sets gives 8000 calibration scores and $22, 33, 28, 25, 22$ clean rejections for QID and $21, 30, 28, 23, 31$ for fQID at $n = 64, 128, 256, 512, 1000$, respectively. The pooled QID rejection rate at $n = 128$ is 6.45%, with a 95% Clopper–Pearson interval of $[4.48, 8.93]\%$.

On the occluded batches, QID, fQID, and RBF-MMD detect every nonzero severity in all 512 trials at every batch size, with 95% Clopper–Pearson intervals of $[99.28, 100]\%$. The Fréchet comparator detects $508/512$ batches for the smallest square at $n = 64$, with interval $[98.01, 99.79]\%$, and every batch in the other settings. These are the same batches used to evaluate the ordering of corruption severity in Table 15.

G Experimental Protocol and ImageNet Model Ranking

G.1 Reference policies and ranking measures

We extend the small-sample metric comparisons of Berthet et al. (2026); Jayasumana et al. (2024) to rankings across generator configurations. To measure sampling reliability, we compare each metric with its own large-sample ordering. For the reference comparisons, we keep one full-reference QID ordering as the target to study the effects of reuse and size, and use that ordering for fQID to measure how closely the approximation preserves QID's decisions. On ImageNet and COCO, we establish the large-sample ordering and draw evaluation batches from disjoint image pools. In the CIFAR-10 sampler experiments and the COCO reference-size and approximation studies, we draw evaluation batches from the pools used to define the large-sample rankings.

QID uses the full real reference pool, and QID ($N = n$) keeps one n-image subset fixed across models and trials. In the ImageNet and COCO comparisons, we evaluate MIND with an n-image reference shared across models and redrawn in each trial, following the ranking experiments of Berthet et al. (2026). Our MIND (independent references) control instead draws a separate real batch for each model. These reference choices use the same empirical sliced-Wasserstein formula on their respective samples. fQID uses QID's full reference and replaces the generated empirical quantiles by Gaussian quantiles. Reference subsets are nested across batch sizes, and methods receive the same generated images and projection directions within each comparison.

Let a_i denote model i's large-sample score and s_{it} its evaluation score in trial t. Pairwise agreement averages over all unordered model pairs, assigning credit 1 to the same order, 0 to opposite strict orders, and $1/2$ to a tie in only one ranking. A tie in both rankings receives credit 1. Without ties, agreement equals $(1 + \tau)/2$, where τ is Kendall's rank correlation. Spearman's ρ uses average ranks for ties. With $A = \arg\min_i a_i$ and $S_t = \arg\min_i s_{it}$, model-selection error is $\mathbf{1}\{S_t \not\subseteq A\}$. Thus a correct selection requires every minimizer of the evaluation score to be a minimizer of the large-sample score.

For the ImageNet and COCO ranking curves, we run $T = 512$ trials per batch size. We construct pointwise 95% intervals for agreement and paired method differences from 5000 bootstrap resamples of complete trials, keeping all models and methods together to preserve dependence among model pairs. Selection-error intervals use Wilson's method, and paired selection-error differences use the trial bootstrap. With zero errors in 512 trials, the Wilson upper endpoint is 0.745%. These intervals describe evaluation-batch variation conditional on the image pools, ranking split, real reference and projection bank. We examine variation in the split, reference subset and directions separately below, together with uncertainty in the large-sample ranking itself.

G.2 ImageNet data and feature representations

We use the released ImageNet-256 samples of Stein et al. (2023) from eleven configurations: ADM, guided ADM and guided and upsampled ADM (Dhariwal and Nichol, 2021), BigGAN (Brock et al., 2019), DiT-XL/2 at classifier-free guidance scales 1.5 and 4.0 (Peebles and Xie, 2023), GigaGAN (Kang et al., 2023), LDM (Rombach et al., 2022), MaskGIT (Chang et al., 2022), RQ-Transformer (Lee et al., 2022) and StyleGAN-XL (Sauer et al., 2022). Each configuration contributes $50{,}000$ generated images, compared with $100{,}000$ real ImageNet training images. We split each generated pool into two disjoint halves of $25{,}000$, use one for the large-sample ranking and the other for evaluation, then exchange their roles in the second fold. Independent permutations of each model's evaluation pool supply nested batches at $n \in \{64, 128, 256, 512, 1000, 2000, 5000, 10000\}$. Sampling is without replacement within a batch, and images may recur across trials. All methods and encoders receive the same generated images in each trial.

Image features are 2048-dimensional Inception-v3 (Szegedy et al., 2016) pool features from the torch-fidelity-compatible implementation and 1024-dimensional CLS features from DINOv2 ViT-L/14 without registers (Oquab et al., 2024). Inception uses its standard preprocessing. DINOv2 uses a 224×224 bicubic resize and ImageNet channel normalization. Both representations retain their original feature norms, with $\alpha = 1$ for squared projected distances. For each encoder, we sample three banks of $M = 100$ normalized Gaussian directions and five fixed-reference subsets. The main comparison uses bank 0, fold 0 and subset seed 0.

Fréchet distance and the RBF and polynomial MMD U-statistics use the full real reference. Fréchet covariance estimates use divisor $n - 1$, while fQID's coordinate variances use divisor n, the moment of the empirical distribution. The RBF kernel is $\exp(-\|x - y\|^2/\tau)$, with τ the median off-diagonal squared distance in a fixed 2048-image reference sample. The polynomial kernel is $(x^\top y/d + 1)^3$. We retain signed U-statistic estimates, including negative values. Fréchet distance and polynomial MMD-U correspond to FID and KID in Inception space; the DINOv2 versions use the same formulas in that representation.

G.3 Recovery of large-sample rankings

Figure 3 compares the sampling reliability of five metrics over all eight batch sizes. Each curve measures recovery of that metric's ranking on the separate $25{,}000$-image pools. At $n = 512$, QID recovers the Inception

Table 5: ImageNet ranking reliability at $n = 512$, relative to each metric's own ranking from disjoint large samples. Confidence intervals across sample sizes are shown in Figure 3.

Method	Ranking agreement (%)	Selection error (%)
Inception		
QID	93.5	31.8
fQID	93.8	30.1
FID	68.7	100.0
RBF MMD-U	94.9	44.3
KID	94.7	5.5
DINOv2		
QID	97.2	1.4
fQID	97.3	1.2
Fréchet distance	91.5	44.5
RBF MMD-U	97.8	0.0
Polynomial MMD-U	97.7	0.0

ranking more reliably than FID, while the two kernel discrepancies have higher pairwise agreement than QID in both representations. Inception KID and the DINOv2 kernels also have lower model-selection error. Both DINOv2 kernels select their large-sample minimizer in all 512 trials.

To study reference sampling, we hold the large-sample QID ranking fixed and vary the reference used for evaluation (Figure 7). At $n = 512$, the full reference has higher pairwise agreement than the small fixed reference and both MIND controls. Relative to MIND (independent references), the gain is 4.49 percentage points in Inception (95% CI 4.07 to 4.90) and 0.87 in DINOv2 (95% CI 0.68 to 1.07). The paired intervals also favor the full reference over MIND with a shared reference and QID ($N = n$). Model selection depends on the gaps among the leading configurations rather than all 55 pairs: here the small fixed reference lowers selection error by 7.23 percentage points in Inception (95% CI 3.71 to 10.74) and 1.17 in DINOv2 (95% CI 0.39 to 2.15).

The Gaussian approximation preserves the common ranking closely. The paired QID–fQID agreement difference at $n = 512$ is -0.263 percentage points in Inception (95% CI -0.444 to -0.075) and -0.057 in DINOv2 (95% CI -0.138 to 0.025). Thus fQID has slightly higher agreement in Inception, with a smaller difference in DINOv2 whose interval includes zero. Appendix L relates this behavior to the information retained by the Gaussian approximation.

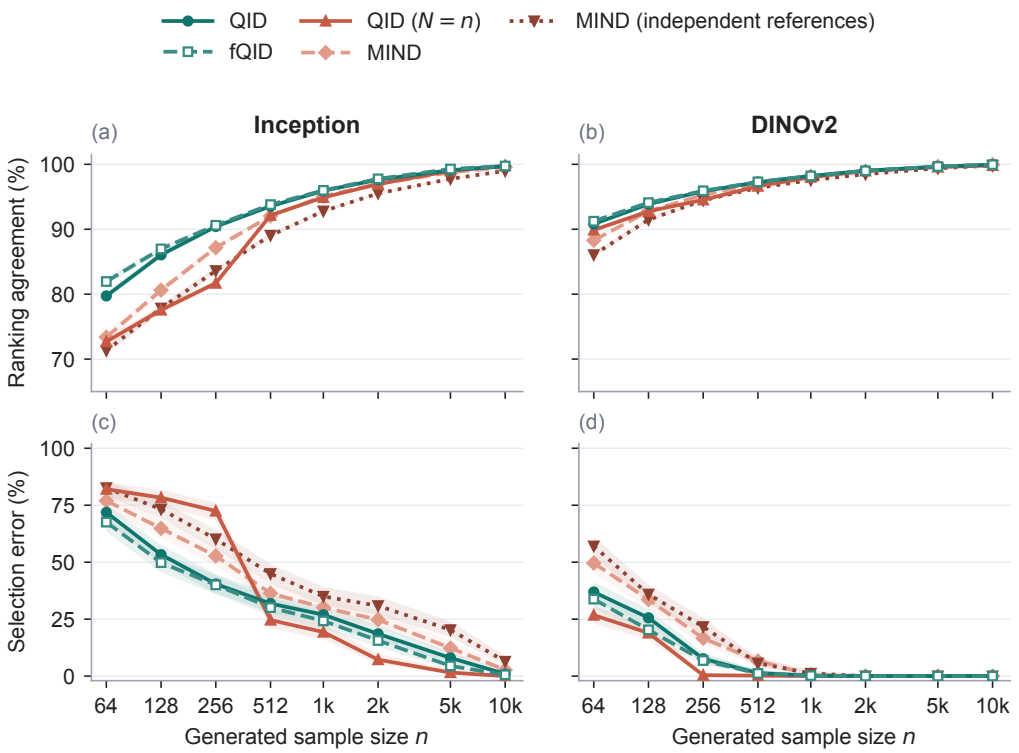

Figure 7: ImageNet ranking under different reference choices. All curves use the full-reference QID large-sample ranking. MIND shares one real batch within a trial; the independent-reference control draws one per model. Upper panels show pairwise agreement and lower panels show model-selection error. Bands are pointwise 95% intervals. QID ($N = n$) uses seed 0.

We assess the large-sample ordering with 256 bootstrap resamples of each model's ranking pool, pairing methods within each model. Centered maximum-standardized intervals cover the 55 score gaps approximately simultaneously. For the main bank and fold, $3/55$ QID gaps remain unresolved in Inception and $1/55$ in DINOv2. We retain a configuration as a possible minimizer unless an interval places its score above another's. This leaves DiT-XL/2 at CFG 1.5 and StyleGAN-XL in Inception, and DiT-XL/2 at CFG 4.0 in DINOv2. The ranking curves retain every model pair and use the observed large-sample minimizer set A.

H Model Ranking on COCO

We extend model ranking to text-to-image generation using CLIP features. With SD-Turbo and SDXL-Turbo (Sauer et al., 2024), we generate images for all 25,014 COCO val2017 captions at one, two and four sampling steps. These distilled models use the latent diffusion architectures of Rombach et al. (2022); Podell et al. (2024). We split the 5000 source images into two groups of 2500, keeping every caption of an image together. Fold 0 uses 12,505 generated images per configuration for the large-sample ranking and the remaining 12,509 for evaluation; fold 1 exchanges their roles. All six configurations share the captions, OpenAI CLIP ViT-B/32 features (Radford et al., 2021), $M = 100$ projection directions and a 118,287-image real reference, with squared scores scaled by $\alpha = 1536$.

We draw 512 caption subsets at each $n \in \{64, 128, 256, 512, 1000, 2000, 5000\}$, pairing subsets across configurations and methods and nesting them across sizes. Sampling is without replacement within a batch, and n counts generated images, including those whose captions share a source image. The reference policies and ranking statistics follow Appendix G.1, with all 15 configuration pairs compared against the full-reference QID ranking. Five fixed-reference subsets assess reference choice; seed 0 supplies the small-reference curve in Figure 8.

To assess the large-sample ordering, we bootstrap the ranking pools 2000 times by source-image group, retaining all captions within a sampled group. Approximate simultaneous intervals resolve all 15 score gaps for both QID and fQID in both folds. Both metrics give the same ordering, with SD-Turbo one-step below every other configuration in these intervals. QID and MIND have similar agreement at $n = 512$: their paired differences are

0.08 percentage points in fold 0 (95% CI −0.08 to 0.23) and 0.21 in fold 1 (95% CI 0.03 to 0.39). Appendix J examines reference-subset variation and the effect of increasing reference size.

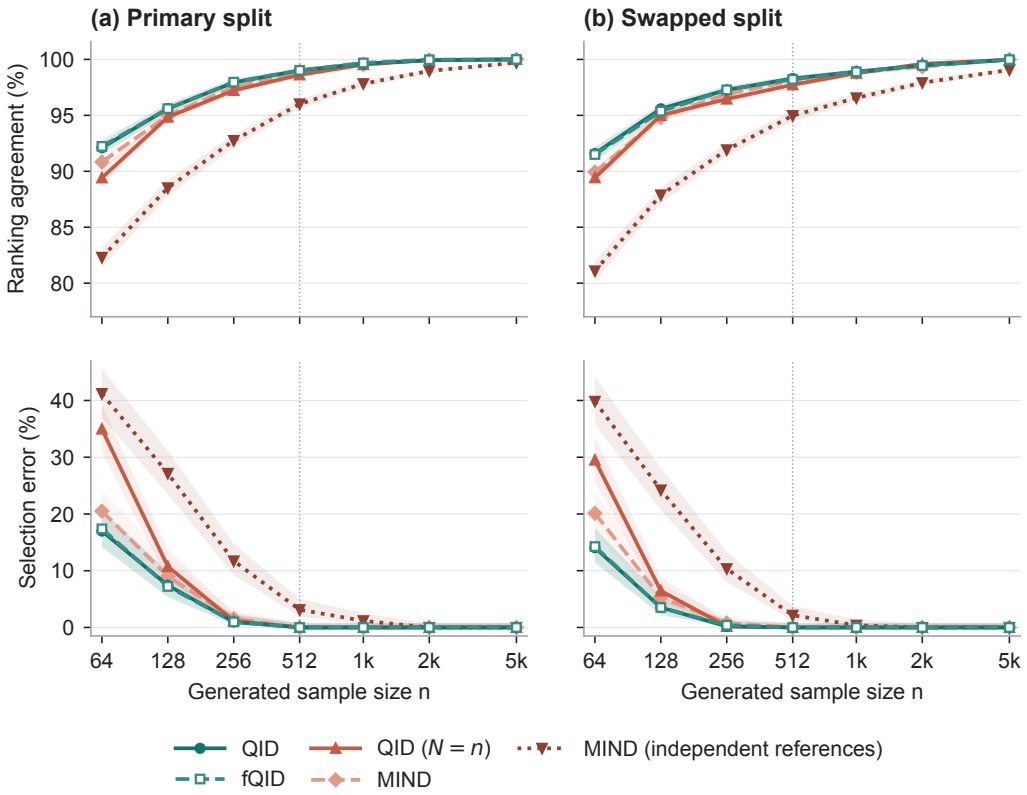

Figure 8: Model ranking on COCO under reference reuse and resampling. Rows show agreement over all 15 configuration pairs and model-selection error relative to the disjoint full-reference QID ranking. Columns show the two folds. Curves use 512 trials per batch size with pointwise 95% intervals; the dotted line marks $n = 512$.

I Small-Sample Ranking on CIFAR-10

We examine sampler settings and controlled image corruptions on CIFAR-10 using torchvision Inception-v3 features. With a pretrained DDPM (Ho et al., 2020), we generate 5000 images at each of five DDIM (Song et al., 2021) step counts, 5, 10, 20, 50, 100. To compare more closely spaced distributions, we form nine nearby mixtures of the DDIM-100 and DDIM-20 feature pools. We use each method's own full-pool ranking, drawing separate real references to define the MIND and QID ($N = n$) targets. Within each evaluation trial, we share one fresh real batch across the samplers or mixtures for MIND. For QID ($N = n$), we draw a separate reference subset at each batch size and keep it fixed across trials. We sample evaluation batches from the same generated pools used for the rankings, and extend the ranking comparison to sixteen image conditions.

I.1 Sampler ranking and nearby mixtures

Reusing the real reference reduces score variation across the tested sample sizes in both tasks (Tables 6 and 7). The five-sampler ranking is recovered reliably from small batches: at $n = 256$, all three reference choices correctly order at least 99.4% of model pairs, and at $n = 1000$ they recover every pair in all 100 trials. The nearby mixtures have smaller score gaps and are harder to rank. QID has higher pairwise agreement than MIND at every sample size in Table 7, reaching 64.4% versus 63.1% at $n = 64$ and approximately 96% for both methods at $n = 2000$.

Table 6: Ranking five DDIM samplers from a pretrained DDPM-CIFAR-10 model over 100 trials per sample size. Agreement uses each method's ranking on 5000 images per sampler. Mean SD averages score standard deviations over the five samplers.

	Pairwise agreement ↑			Mean SD ↓		
n	MIND	QID ($N = n$)	QID	MIND	QID ($N = n$)	QID
64	0.958	0.960	0.972	11.661	8.286	7.988
128	0.981	0.978	0.983	8.200	5.544	5.888
256	0.994	0.997	0.995	5.421	4.314	3.946
512	0.998	0.999	0.999	3.918	2.740	2.770
1000	1.000	1.000	1.000	2.658	1.887	1.899
2000	1.000	1.000	1.000	1.671	1.118	1.123

Table 7: Ranking nine mixtures of the DDIM-100 and DDIM-20 generated-feature pools over 100 trials per sample size. Agreement uses each method's ranking on 5000 images per mixture. Mean SD averages score standard deviations over the nine mixtures.

	Pairwise agreement ↑			Mean SD ↓		
n	MIND	QID ($N = n$)	QID	MIND	QID ($N = n$)	QID
64	0.631	0.624	0.644	8.089	4.394	4.938
128	0.707	0.726	0.720	5.015	4.059	3.375
256	0.784	0.772	0.794	3.307	2.657	2.287
512	0.853	0.865	0.858	2.216	1.474	1.486
1000	0.913	0.911	0.914	1.531	1.065	1.043
2000	0.959	0.959	0.960	0.933	0.618	0.607

I.2 Corruption rankings across metrics

The metrics give similar large-sample orderings: FID and QID agree on the five sampler rankings and on 98.3% of corruption pairs (Table 8). For the corruption targets, FID and QID use all 10,000 images per condition and the 40,000-image real reference. KID and MIND use the same sampled 5000 images per condition and a shared 5000-image real subset. We then draw 40 smaller batches per setting, pairing the evaluated images across metrics. KID and MIND share a fresh n-image real batch within each trial, while FID and QID retain the full real reference.

With 256 images, QID recovers 98.3% of its large-sample pairwise ordering, compared with 96.8% for FID and 97.7% for KID (Table 9). Agreement increases with sample size, with FID reaching the highest value at $n = 1000$. KID has the highest agreement with the known corruption order at all three sizes and correctly orders every adjacent severity pair at $n = 1000$. The score estimates also illustrate the distinction between sampling variability and estimation error discussed by Binkowski et al. (2018); Chong and Forsyth (2020): at $n = 256$, FID has a mean standard deviation of 1.89 but an RMSE of 70.11 relative to its large-sample scores.

Table 8: Pairwise agreement between large-sample metric rankings. Sample counts give the evaluated images per condition for the first and second metric in each pair; reference sizes follow the text.

Dataset	Variants	Samples	Metric pair	Agreement
CIFAR-10 corruptions	16	$10k/10k$	FID vs. QID	0.983
CIFAR-10 corruptions	16	$5k/10k$	KID vs. QID	0.967
CIFAR-10 corruptions	16	$5k/10k$	MIND vs. QID	0.992
DDPM-CIFAR-10 samplers	5	$5k/5k$	FID vs. QID	1.000

Table 9: Score estimation and ranking on sixteen CIFAR-10 corruption conditions over 40 repeated subsamples per setting. RMSE and pairwise agreement use each metric's large-sample scores, while severity agreement uses the known order of adjacent corruption levels. RMSE pools conditions and repetitions, and mean SD averages within-condition standard deviations, both in each metric's native units.

Metric	n	RMSE ↓	Mean SD ↓	Pairwise agreement ↑	Severity agreement ↑
FID	256	70.11	1.89	0.968	0.945
FID	512	42.85	1.27	0.989	0.972
FID	1000	23.14	0.90	0.997	0.995
KID	256	4.57×10^{-3}	3.76×10^{-3}	0.977	0.963
KID	512	3.15×10^{-3}	2.61×10^{-3}	0.984	0.990
KID	1000	2.33×10^{-3}	1.80×10^{-3}	0.985	1.000
MIND	256	14.35	6.41	0.969	0.930
MIND	512	7.53	4.24	0.982	0.968
MIND	1000	4.71	3.10	0.986	0.990
QID	256	8.69	4.51	0.983	0.955
QID	512	4.88	2.93	0.989	0.978
QID	1000	3.14	2.12	0.993	0.997

fQID also recovers its own corruption and sampler rankings with small batches (Table 10). Its pairwise agreement reaches 99.3% on corruptions at $n = 1000$ and 99.8% on the five samplers at $n = 512$. The close-mixture task requires larger batches, reaching 96.0% agreement at $n = 2000$.

Table 10: Small-batch fQID ranking agreement. Targets use all 10,000 images per corruption condition or 5000 generated images per sampler or mixture.

Task	n	Pairwise agreement ↑
CIFAR-10 corruptions	256	0.982
CIFAR-10 corruptions	512	0.989
CIFAR-10 corruptions	1000	0.993
CIFAR-10 corruptions	2000	0.997
CIFAR-10 corruptions	5000	0.9995
DDPM samplers	64	0.968
DDPM samplers	128	0.985
DDPM samplers	256	0.992
DDPM samplers	512	0.998
DDPM samplers	1000	1.000
DDPM mixtures	64	0.674
DDPM mixtures	256	0.794
DDPM mixtures	512	0.857
DDPM mixtures	1000	0.916
DDPM mixtures	2000	0.960

J Reference Reuse, Reference Size and Score Variation

We examine reference reuse and reference size separately because they affect model comparisons in different ways. Fixing a reference removes variation caused by drawing new real images at every evaluation, while increasing its size changes the reference distribution being compared with each model. We study model-score gaps across reference subsets, projection directions and data splits. Their variation includes the covariance induced when models share a real-reference draw and is therefore directly relevant to ranking.

J.1 Model-score gaps on ImageNet

Write $\Delta_{ijt} = s_{it} - s_{jt}$ for a trial's score gap, $\delta_{ij} = a_i - a_j$ for the corresponding large-sample gap, and $\bar{\Delta}_{ij} = T^{-1}\sum_t \Delta_{ijt}$. Its mean squared estimation error decomposes as

$$\frac{1}{T}\sum_{t=1}^{T}(\Delta_{ijt} - \delta_{ij})^2 = (\bar{\Delta}_{ij} - \delta_{ij})^2 + \frac{1}{T}\sum_{t=1}^{T}(\Delta_{ijt} - \bar{\Delta}_{ij})^2. \tag{J.1}$$

The first term is the squared error of the mean gap and the second is trial variance, with divisor T. Figure 9 averages these terms over all 55 ImageNet model pairs. At $n = 512$, the full and small fixed references have nearly the same gap variance in both feature spaces, below that of the fresh-reference controls. The full reference lowers the squared error of the mean gap. The size comparison varies with the evaluation batch: at $n = 2000$ in DINOv2, the small-reference mean squared gap error is approximately 0.922 times the full-reference value.

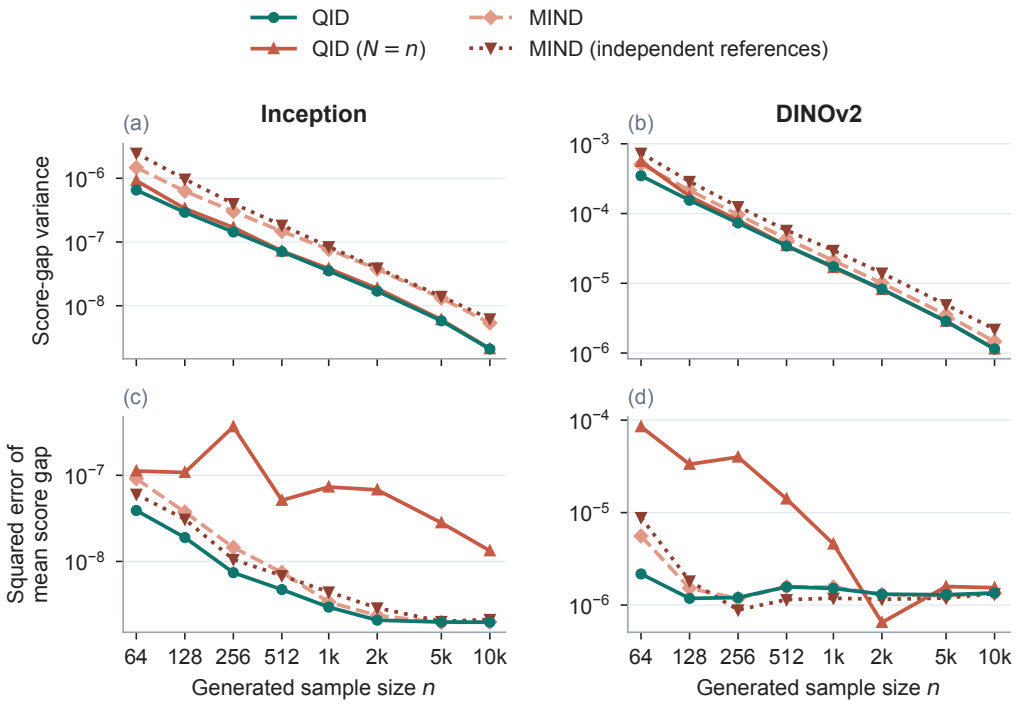

Figure 9: Error in ImageNet model-score gaps under reference reuse and resampling. Mean squared error is decomposed into the squared error of the mean score gap and trial variance, then averaged over all 55 pairs. Variance uses divisor $T = 512$; each encoder retains its squared-score-gap units.

J.2 Projection directions, ranking splits and fixed subsets

We repeat the ImageNet comparison across three projection banks in both ranking folds and across five fixed-reference subsets with bank 0 and fold 0. Each setting uses its corresponding large-sample ranking. Figures 10 and 11 report paired differences at $n = 512$, separating the effect of the directions and image split from the choice of a small fixed reference.

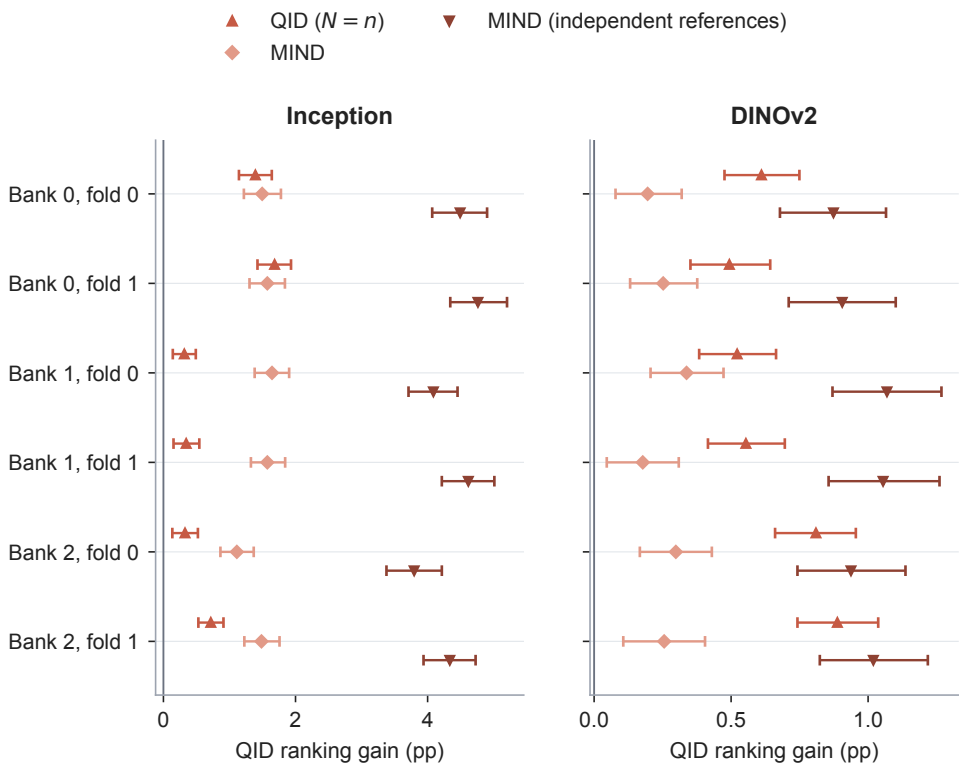

Figure 10: ImageNet ranking agreement across three projection banks and two data splits at $n = 512$. Points show QID minus MIND, its independent-reference control or QID ($N = n$), in percentage points, with paired pointwise 95% intervals. Positive differences favor the full reference. QID ($N = n$) uses seed 0.

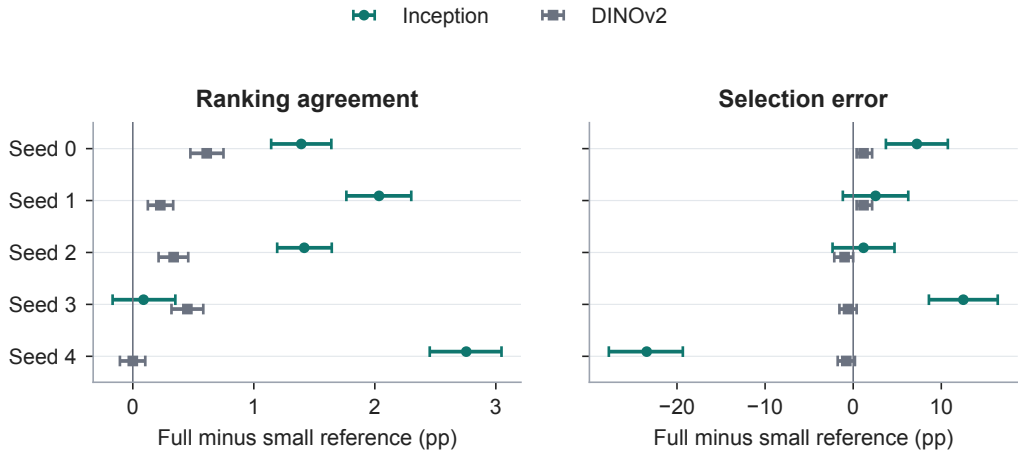

Figure 11: ImageNet ranking and selection across five small fixed references at $n = 512$, using bank 0 and fold 0. Differences are QID minus QID ($N = n$), with paired pointwise 95% intervals. Positive agreement and negative selection-error differences favor the full reference.

We repeat the five-reference comparison on the COCO splits (Figure 12). Seed 0, used in Figure 8, gives full-reference agreement gains of 0.39 and 0.56 percentage points in the two folds. Seeds 1 and 3 give higher agreement to the small reference in both folds. We therefore average over nested reference choices in the following study of reference size.

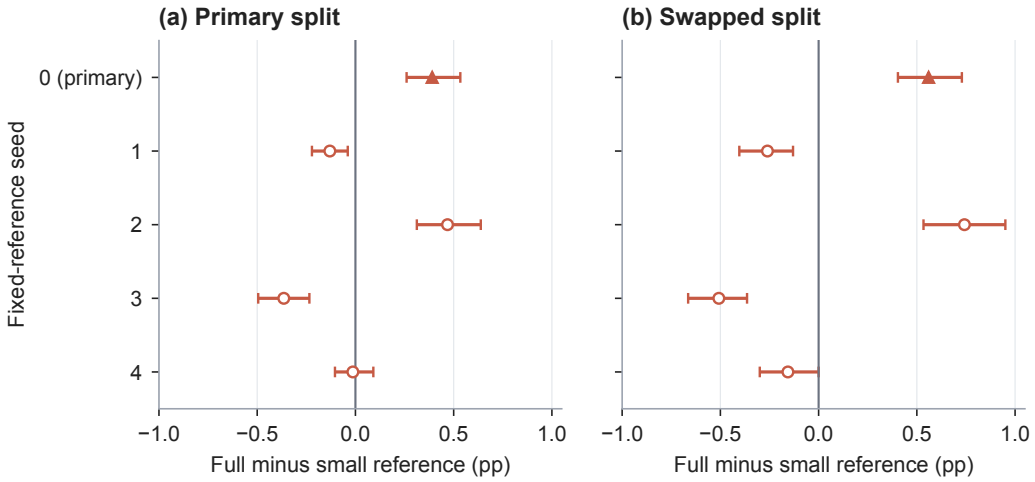

Figure 12: COCO ranking agreement across five small fixed references at $n = 512$. Points show QID minus QID ($N = n$) over all 15 pairs, in percentage points, with paired pointwise 95% intervals. Both ranking folds are shown; seed 0 supplies the small-reference curve in Figure 8.

J.3 Nested reference sizes on COCO

For each of ten reference choices, we score the same COCO batches against nested real-image subsets of sizes n, 5000 and 50,000, and against all 118,287 real images. This study uses all 25,014 generated outputs to define the large-sample QID ranking and to draw evaluation batches. Score-gap RMSE averages squared errors over the 15 model pairs, trials and reference choices before taking the square root. Confidence intervals resample trials after averaging over the ten choices, with one full-reference value per trial.

Larger references lower gap RMSE at every reported batch size (Table 11). At $n = 128$, the full reference raises mean pairwise agreement over the ten fixed-n choices by 0.0079 (paired 95% CI 0.0059 to 0.0099). Its 23 model-selection errors compare with a mean of 30.5 and a range of 13–48 for small references. Thus the larger reference improves average gap estimation and agreement, while individual subsets can produce better selection decisions.

Table 11: COCO model-score-gap RMSE ($\times 10^{-3}$) over 512 paired trials and ten nested reference choices. Selection-error counts at $n = 128$ are means, with ranges across reference choices in parentheses.

Reference size	$n = 128$	$n = 512$	$n = 2000$	Selection errors at $n = 128$
n	12.087	5.921	2.860	30.5 (13–48)
5000	11.106	5.429	2.751	25.7 (21–32)
50,000	11.105	5.386	2.647	23.7 (23–25)
118,287	11.103	5.384	2.642	23

J.4 Score variation on CIFAR-10

The CIFAR-10 reference comparison holds the full-reference QID scores fixed as the target for all three policies. Across sixteen conditions and 100 repeated subsamples at $n = 512$, reference reuse lowers mean score standard deviation from 4.37 to 3.06 at either reference size (Table 12). The full reference has lower RMSE and better severity agreement, while the small reference has slightly higher pairwise agreement. Averaging over conditions, empirical variance with divisor 100 and the squared error of the mean score contribute $14.01 + 33.00$ for the small reference and $14.37 + 11.40$ for the full reference. The lower RMSE therefore comes from mean scores that are closer to the large-sample target.

Table 12: Reference sampling on sixteen CIFAR-10 conditions at $n = 512$, $M = 100$, over 100 subsamples. RMSE and pairwise agreement use the full-reference QID scores. Mean SD averages sample standard deviations over conditions; severity agreement averages adjacent corruption levels.

Metric	Mean SD ↓	RMSE↓	Severity agreement↑	Pairwise agreement↑
MIND	4.37	8.30	0.965	0.980
QID ($N = n$)	3.06	6.86	0.973	**0.987**
QID	3.06	**5.08**	**0.984**	0.985

K Sensitivity to Distribution Changes

K.1 Distributions with the same mean and covariance

We first isolate information beyond the first two moments, following the feature-distribution tests of Berthet et al. (2026); Jayasumana et al. (2024). COCO train2017 images are resized bicubically to 512×512 and encoded with torch-fidelity Inception-v3-compat. Let μ and $\Sigma = \sum_{j=1}^{d} \lambda_j v_j v_j^\top$ be the mean and unbiased covariance of FID's 50,000-image reference, with orthonormal eigenvectors v_j. The $2d$ equally weighted vectors

$$x_{j,\pm} = \mu \pm \sqrt{(2d-1)\lambda_j/2}\, v_j$$

have sample mean μ and unbiased sample covariance Σ. With $d = 2048$, this gives 4096 feature vectors with a full-rank sample covariance, preserving the statistics used by FID.

Table 13 compares this distribution with real-image batches. QID uses the same 50,000-image reference as FID, $M = 1000$ directions and $\alpha = 6144$. Its scores are 190.6468 for the construction and 0.6233 for a real batch of 4096 images. Using all 118,287 reference images gives 190.58 and 0.60 with the same directions and real batch. MIND averages construction scores over 16 fresh real-reference batches of size 4096, and its real-data entry uses one real–real comparison. The fQID real-data entry averages two batches.

FID is zero up to a numerical residual of -1.17×10^{-12} because the constructed features have the reference mean and covariance. The Gaussian controls are FID's mean term and sliced FID, which averages Gaussian Wasserstein costs between projected features (Berthet et al., 2026). The polynomial and RBF discrepancies use V-statistics. Each metric retains its native units in the table.

Table 13: Scores on real data and features with the reference sample mean and covariance, in each metric's native units. QID and FID use the same 50,000-image reference. The MIND entry is mean $\pm$ SD over 16 reference draws; fQID's real-data entry averages two batches. The construction and sampling are given in Appendix K.1.

Metric	Real-image features	Constructed features
QID	0.62	190.65
MIND	1.13	192.08 ± 0.58
fQID	4.75	4.62
FID	8.48	≈ 0
RBF MMD (V)	2.89×10^{-4}	0.0377
Poly MMD (V)	2.08×10^{-4}	1.13×10^{-4}
Sliced FID	0.22	≈ 0
FID (mean term)	0.05	≈ 0

K.2 Changes in class proportions

We vary CIFAR-10 class proportions using the construction of Kwon et al. (2026) and OpenAI CLIP ViT-B/32 features without L2 normalization. For twenty ordered class pairs, each of ten repetitions partitions the 6000 images per class into 2500 reference images and 3500 evaluation images. A balanced $N = 5000$ reference is compared with independent batches at $n \in \{1000, 2000, 3000, 4000, 5000\}$ and first-class proportions $p \in \{0.500, 0.505, \ldots, 0.550\}$. Reference and evaluation images are disjoint, with sampling without replacement within each batch. Each repetition uses one reference and a new projection bank. We evaluate QID with all 10,000 directions and with the first 100, and fQID with the first 100.

For each pair, we correlate the ten positive proportion shifts with their mean distances over ten repetitions, then average Spearman correlations over the twenty pairs. Distances use $\alpha = 1$, with square roots taken after averaging squared slice or kernel contributions. The CLIP Fréchet distance uses covariance divisor $n - 1$.

RBF-MMD uses $\sigma = 10$, with root V-statistic, squared V-statistic and signed U-statistic variants. QID at both direction counts and fQID have higher correlation than CLIP Fréchet at every size in Table 14; QID with 10,000 directions and the RBF scores reach correlations near one with the smallest tested batch.

Table 14: Spearman correlation between increasing class-proportion shifts and mean distances on CIFAR-10. Entries average twenty ordered class pairs, with ten repetitions per distance. Higher values indicate closer agreement with the shift order.

Method	$n = 1000$	2000	3000	4000	5000
QID ($M = 10{,}000$)	0.9927	0.9964	1.0000	1.0000	1.0000
QID ($M = 100$)	0.9788	0.9952	0.9994	0.9994	0.9994
fQID ($M = 100$)	0.9794	0.9891	0.9988	0.9988	0.9994
CLIP Fréchet, root	0.9527	0.9800	0.9879	0.9952	0.9988
RBF-MMD, root V	0.9927	0.9982	1.0000	1.0000	1.0000
RBF-MMD, squared V	0.9921	0.9982	0.9994	1.0000	1.0000
RBF-MMD, signed U	0.9915	0.9982	0.9994	1.0000	1.0000

K.3 Increasing corruption and restricting image diversity

Following Berthet et al. (2026), we test whether scores increase with blur and occlusion severity. We apply three Gaussian blur levels with $\sigma \in \{0.8, 1.2, 1.6\}$ and kernel widths $5, 7, 9$, and three black-square occlusions with side lengths $6, 8, 11$ pixels. To keep the masks nested, we draw one random top-left location per image that is valid for the largest square. We use the same CIFAR-10 images across corruption levels and methods, with 512 trials at each $n \in \{64, 128, 256, 512, 1000\}$. An ordering error occurs when a score decreases between either pair of adjacent positive severity levels.

This experiment uses torchvision Inception features with ImageNet weights and the first 40,000 training images as the real pool. We evaluate MIND with n real images shared across corruption levels in each trial, while QID ($N = n$) holds one subset fixed. We evaluate five banks of 100 directions at $\alpha = 6144$. Fréchet distance uses the full-reference moments, and RBF-MMD uses the signed U-statistic against all 40,000 real images. Its kernel is $\exp(-\|x - y\|^2/h)$, where h is the median positive squared distance in a 2048-image calibration subset.

All methods correctly order blur in every trial with the bank used in Table 15. Occlusion separates the methods at small batch sizes. At $n = 64$, the paired QID-minus-MIND error-rate difference is -3.71 percentage points (95% CI -6.05 to -1.17). Across the five banks, the corresponding error-rate ranges are 2.34–13.09% for full-reference QID, 6.25–16.80% for MIND and 1.37–19.92% for small-reference QID. The two fixed-reference choices change order across banks. RBF-MMD correctly orders every occlusion trial at every tested size and bank.

Table 15: Occlusion-ordering errors out of 512 trials, for one projection bank. An error is a score decrease between adjacent levels of the three nested occlusions.

Method	$n = 64$	128	256	512	1000
QID	67	17	2	0	0
fQID	62	17	3	0	0
MIND	86	30	5	0	0
QID ($N = n$)	102	50	2	0	0
Fréchet (torchvision), full reference	106	8	0	0	0
RBF-MMD U, full reference	0	0	0	0	0

Restricting the available images tests sensitivity to reduced diversity. We draw $n = 1000$ images with replacement from subsets containing 512, 256, 128 or 64 CIFAR-10 images, and compare scores with paired clean batches over 50 repetitions. Both MIND and QID increase as the available subset shrinks (Table 16), with similar changes over the tested range.

Table 16: Score increases after restricting available CIFAR-10 images. Batches of $n = 1000$ are drawn with replacement. Entries are mean increases relative to paired clean batches over 50 repetitions, with pointwise 95% paired bootstrap intervals.

Available images	MIND	QID
512	3.30 (3.13–3.48)	3.43 (3.32–3.55)
256	6.75 (6.52–6.98)	6.96 (6.74–7.19)
128	13.02 (12.68–13.35)	13.22 (12.89–13.53)
64	25.32 (24.60–26.11)	25.72 (25.04–26.45)

L Gaussian Approximation and Projection Sampling

L.1 Covariance and quantile contributions

We use the class mixtures to separate the two approximations in fQID: using diagonal covariance to obtain projected variances and replacing the generated projected distributions by Gaussians. We evaluate twenty ordered class pairs, eleven proportions and ten joint repetitions at $n = 1000, 5000$, giving 4400 paired batches with $N = 5000$, $M = 100$ and $\alpha = 1$. To measure the contribution of each approximation, we compare fQID with the Gaussian with measured projected variances H, the Gaussian-reference cost G with fQID's diagonal scales, FastSW and the projected mean contribution A_U defined in Appendix E. All methods receive the same CLIP features, reference images and directions. We measure distance error relative to QID before and after subtracting each distribution's empirical mean.

Measured projected variances reduce mean absolute error from 0.02536 and 0.02983 for diagonal fQID to 0.00422 and 0.00764 at $n = 1000$ and 5000 (Table 17). In the squared-score decomposition of Eq. (E.9), each error term is averaged over directions before squaring. Before centering, covariance omission accounts for 72.6–78.3% of mean squared error, Gaussian quantile approximation for 1.4–2.3%, and the cross term for 20.3–25.2%. The Gaussian-reference control also has lower error than fQID on these distributions. Thus the diagonal variance approximation is the larger source of error in this experiment.

Table 17: Mean absolute distance error relative to QID on paired class-mixture batches. Each column averages 2200 batches. Mean QID distances at $n = 1000, 5000$ are $0.01501, 0.01011$ before centering and $0.01182, 0.00717$ after centering. The mean-only distance is zero after centering.

	Before centering		After centering	
Method	$n = 1000$	$n = 5000$	$n = 1000$	$n = 5000$
Diagonal fQID	0.02536	0.02983	0.02744	0.03208
Projected-variance Gaussian	0.00422	0.00764	0.00498	0.00899
Gaussian reference	0.02258	0.02703	0.02458	0.02922
FastSW	0.00577	0.00310	0.01107	0.00671
Projected mean only	0.00589	0.00317	0.01182	0.00717

For each sample size, we compare orderings across 200 proportion curves, with 55 pairs per curve and 11,000 comparisons in total. Before centering, the projected mean contribution alone disagrees with QID on 5.27% and 1.16% of pairs. Centering removes that contribution, leaving projected shape to determine the order. fQID then disagrees with QID on 49.25% and 53.99% of pairs, with ranges of 47.36–52.09% and 50.00–58.09% across the ten joint repetitions. Using measured projected variances reduces this disagreement (Table 18).

Table 18: Pairwise ranking disagreement with QID on class mixtures, in percent. All scored proportion pairs have distinct values except the centered mean-only control, which ties every pair.

	Before centering		After centering	
Method	$n = 1000$	$n = 5000$	$n = 1000$	$n = 5000$
Diagonal fQID	7.71	2.27	49.25	53.99
Projected-variance Gaussian	3.85	0.75	28.01	29.18
Gaussian reference	7.52	2.19	49.25	53.44
FastSW	8.14	1.86	44.79	37.46
Projected mean only	5.27	1.16	All ties	All ties

L.2 Approximation fidelity on COCO model comparisons

To measure how closely the Gaussian approximation preserves model comparisons, we score the same COCO batches with QID and fQID, using 100 projection directions and 512 trials per size. The batches come from the full generated pools that define the common model ranking. We compare the methods through pairwise ranking disagreement and the RMSE of their score gaps over all 15 model pairs, and assess selection errors against the common full-pool minimizer. Agreement increases with the generated batch size (Table 19). At $n = 512$, pairwise disagreement is 0.35% (95% CI 0.23 to 0.48), and the methods select the same best configuration in every trial. Both make zero selection errors from $n = 512$ onward.

Table 19: Agreement of fQID and QID on paired COCO batches. Pair disagreement has pointwise 95% trial-bootstrap intervals; winner disagreement counts different trial selections. Gap RMSE compares the 15 model-score gaps. Selection errors use the common full-pool ranking.

				Selection errors	
n	Pair disagreement (%)	Winner disagreement	Gap RMSE ($\times 10^{-3}$)	QID	fQID
64	2.30 (1.99–2.63)	38/512	3.174	80	80
128	1.08 (0.87–1.30)	8/512	1.976	23	27
256	0.68 (0.51–0.86)	4/512	1.321	5	7
512	0.35 (0.23–0.48)	0/512	0.958	0	0
1000	0.22 (0.13–0.34)	0/512	0.779	0	0
2000	0.05 (0.01–0.10)	0/512	0.682	0	0
5000	0.01 (0.00–0.04)	0/512	0.612	0	0

L.3 Approximation error across dimensions

We use the independent-coordinate Gamma construction of Nadjahi et al. (2021) to vary feature dimension. At $n = N = 10{,}000$ and $d \in \{10, 50, 100, 500, 1000\}$, reference coordinate shapes are uniform on $[1, 5)$ with scale 2, and generated coordinate shapes are uniform on $[5, 10)$ with scale 3. Shape vectors remain fixed within each dimension for 100 independent dataset pairs. Each pair receives independent banks of $L = 100, 1000, 5000$ directions and an independent 20,000-direction empirical SW target. We evaluate the pairs before and after empirical centering. FastSW uses its analytic mean-and-trace expression, with moment divisor n and $\alpha = 1$.

Let S_L be empirical squared SW on L directions, F_L be squared fQID on the same bank and S_* the independent target. The signed distance error decomposes as $(\sqrt{F_L} - \sqrt{S_L}) + (\sqrt{S_L} - \sqrt{S_*})$, separating Gaussian approximation from projection sampling. Table 20 averages absolute errors over the 100 dataset pairs. Before centering, projection error dominates fQID's total error and FastSW is more accurate at every dimension and direction budget. After centering, fQID has lower total error at $L = 5000$ for $d \geq 50$, while FastSW has lower error at $d = 10$. The target's median Monte Carlo standard error is 0.074–0.084 before centering and 0.00055–0.00147 after centering.

Table 20: Mean absolute distance error on Gamma distributions before and after empirical centering. Approximation error compares fQID with SW on the same bank; Monte Carlo error compares that SW estimate with the independent 20,000-direction target. The fQID and FastSW columns give total error relative to that target.

		Before centering				After centering	
d	L	Approx.	MC	fQID	FastSW	fQID	FastSW
10	100	0.00450	0.84778	0.84794	0.06563	0.01920	0.01317
10	1000	0.00439	0.24581	0.24604	0.06563	0.01710	0.01317
10	5000	0.00441	0.13904	0.13904	0.06563	0.01625	0.01317
50	100	0.00117	0.93564	0.93589	0.05844	0.01678	0.00670
50	1000	0.00036	0.24870	0.24862	0.05844	0.00515	0.00670
50	5000	0.00018	0.13878	0.13875	0.05844	0.00276	0.00670
100	100	0.00107	0.99175	0.99205	0.06783	0.01352	0.00411
100	1000	0.00041	0.30911	0.30906	0.06783	0.00362	0.00411
100	5000	0.00022	0.14798	0.14803	0.06783	0.00175	0.00411
500	100	0.00121	1.00825	1.00848	0.07188	0.00552	0.00195
500	1000	0.00044	0.29204	0.29213	0.07188	0.00195	0.00195
500	5000	0.00022	0.15595	0.15592	0.07188	0.00103	0.00195
1000	100	0.00122	0.93460	0.93476	0.06856	0.00370	0.00174
1000	1000	0.00045	0.30113	0.30115	0.06856	0.00171	0.00174
1000	5000	0.00022	0.15536	0.15540	0.06856	0.00097	0.00174

L.4 Number of projection directions

To measure how projection sampling affects model ranking, we vary the number of directions on the full COCO generated pools in two separate studies. We use L2-renormalized CLIP features for QID and the stored near-unit features for fQID, with independently drawn banks. For each score, we define a 10,000-direction ranking by averaging ten banks of 1000 directions, then evaluate additional banks against that ranking. We use ten additional banks at $M = 1000$ and twenty at each smaller size, except for nineteen fresh fQID banks at $M = 100$, which exclude the original bank. Table 21 shows the convergence of each score in a separate panel.

Pair agreement increases with the direction budget in both studies. Exchanging the banks used to define and evaluate each large-sample ranking preserves its order. Pointwise paired 95% intervals across the ten target banks exclude zero for all 15 model-score gaps. Both 10,000-direction rankings select SD-Turbo one-step and reverse the SDXL-Turbo two-step/four-step order relative to the original 100-direction bank.

Table 21: Separate COCO projection studies, each using its own 10,000-direction ranking and full generated pools. Pair agreement averages all 15 pairs within each bank; winner agreement counts banks selecting the large-sample minimizer.

(a) QID

M	Pair agreement (%)	Winner agreement
50	86.33	17/20
100	93.00	19/20
250	96.33	20/20
500	97.67	20/20
1000	98.67	10/10

(b) fQID

M	Pair agreement (%)	Winner agreement
50	88.67	19/20
100	92.63	16/19
250	94.33	20/20
500	96.67	20/20
1000	99.33	10/10

M Scoring Time and Reference Storage

We measure the cost of preparing a real reference and evaluating subsequent generated batches. Timings start from extracted image features, following the computational comparisons of Berthet et al. (2026); Jayasumana et al. (2024). To isolate the cost of the reference representation, we compute QID in three ways using the same images, directions and scale. Sorted projections permit paired or grouped squared differences when the sample sizes are equal or divisible. Prefix sums provide the reference integral over each generated quantile interval, while the interval averages and residual in Eq. (C.9) prepare these contributions for the chosen batch size. All three compute the same empirical transport score. We compare their preparation, retained storage and repeated scoring costs, alongside the Gaussian approximations and other metrics.

M.1 GPU SCORING ON IMAGENET FEATURES

We use an Apple M5 Max GPU with up to 100,000 real ImageNet features, generated batches drawn from 25,000 ADM outputs and 100 common projection directions. Each method starts from the same Inception or DINOv2 feature arrays and returns a scalar score; timings include computation and device transfers. The Gaussian approximations are diagonal fQID, the Gaussian with measured projected variances H, and FastSW's mean-and-trace summary of both distributions. The polynomial and RBF kernel discrepancies use full-reference U-statistics. Fréchet distance uses full covariance, computed on the GPU with decomposition on the CPU, and both stages are included in its timing.

We time reference preparation from raw features separately from scoring, preparing quantities independent of n once and reusing them across batch sizes. For each setting, we use three generated batches, three timed repetitions per batch and two warmup calls. Within each process run, we take the median over repetitions and then batches. We report medians over five process runs (Table 22), with interquartile ranges in Figure 6 and Table 23, and compute speed ratios from paired per-run timings.

Table 22: Reference preparation, storage and GPU scoring on ImageNet features with $N = 100{,}000$. Preparation and storage use $n = 1000$; quantities independent of batch size are reused. Storage includes projection directions where applicable. Times are medians over five runs, including device transfers and the CPU covariance decomposition for Fréchet distance.

	Reference, $n = 1000$		Scoring time (ms)		
Method	Setup (s)	Storage (MiB)	$n = 128$	$n = 1000$	$n = 5000$
Inception					
Reference representations					
Sorted projections	0.0492	38.9	2.13	1.27	2.84
Prefix sums	0.0504	77.1	0.542	0.846	2.66
QID	0.0471	1.17	0.525	0.798	2.51
Gaussian approximations					
fQID	0.136	1.57	0.465	0.726	2.26
Projected-Gaussian H	0.132	0.79	0.523	0.779	2.26
FastSW	0.088	0.00782	0.406	0.604	1.82
Kernel discrepancies					
Polynomial MMD-U	1.8	781	6.08	35.4	175
RBF MMD-U	2.27	781	11.3	46.7	223
Fréchet distance					
Fréchet distance	0.409	16	9.63	83.4	583
DINOv2					
Reference representations					
Sorted projections	0.0277	38.5	2.11	1.07	1.98
Prefix sums	0.0296	76.7	0.521	0.712	1.74
QID	0.0278	0.776	0.502	0.649	1.65
Gaussian approximations					
fQID	0.106	0.786	0.475	0.588	1.35
Projected-Gaussian H	0.104	0.395	0.507	0.632	1.43
FastSW	0.0634	0.00391	0.623	0.538	1.04
Kernel discrepancies					
Polynomial MMD-U	0.999	391	3.81	20	96
RBF MMD-U	1.42	391	7.01	28.9	138
Fréchet distance					
Fréchet distance	0.131	4	5.68	49.1	152

At $N = 100{,}000$, QID's interval averages give a 1.14–4.14× speedup over computation from sorted projections and 1.03–1.10× over using prefix sums. The comparison across coprime, divisible and equal sample sizes in Table 23 shows why the gain depends on the batch size. At unequal sizes, the prepared coefficients avoid revisiting reference quantiles; at $N = n = 10{,}000$, they coincide with the sorted reference, giving paired differences and comparable cost. Approximating the generated quantiles gives fQID a further 1.05–1.20× speedup over QID at $N = 100{,}000$.

Table 23: GPU computation times for three QID reference representations at equal, divisible and coprime sample sizes, using a prepared reference and 100 directions. Entries are median milliseconds over five process runs, with interquartile ranges in parentheses.

Features	N	n	Sorted projections	Prefix sums	QID
Inception	10,000	128	0.757 (0.724–0.852)	0.604 (0.560–0.633)	0.583 (0.559–0.633)
	10,000	1,000	0.822 (0.807–0.824)	0.839 (0.836–0.850)	0.790 (0.789–0.869)
	10,000	5,000	2.550 (2.533–2.560)	2.623 (2.605–2.649)	2.520 (2.518–2.523)
	10,000	10,000	4.404 (4.383–4.409)	4.608 (4.585–4.611)	4.417 (4.408–4.434)
	100,000	128	2.129 (2.128–2.173)	0.542 (0.517–0.552)	0.525 (0.523–0.553)
	100,000	997	2.456 (2.452–2.464)	0.860 (0.857–0.863)	0.808 (0.804–0.821)
	100,000	1,000	1.269 (1.259–1.271)	0.846 (0.845–0.887)	0.798 (0.797–0.810)
	100,000	5,000	2.844 (2.833–2.877)	2.656 (2.631–2.674)	2.514 (2.503–2.529)
DINOv2	10,000	128	0.706 (0.594–0.757)	0.503 (0.502–0.520)	0.476 (0.470–0.501)
	10,000	1,000	0.659 (0.654–0.676)	0.682 (0.680–0.693)	0.651 (0.649–0.665)
	10,000	5,000	1.645 (1.614–1.676)	1.754 (1.737–1.780)	1.625 (1.605–1.643)
	10,000	10,000	2.699 (2.657–2.709)	2.873 (2.865–2.893)	2.689 (2.664–2.702)
	100,000	128	2.108 (2.089–2.128)	0.521 (0.521–0.524)	0.502 (0.497–0.504)
	100,000	997	2.285 (2.267–2.301)	0.693 (0.692–0.743)	0.652 (0.650–0.652)
	100,000	1,000	1.073 (1.059–1.080)	0.712 (0.704–0.713)	0.649 (0.648–0.655)
	100,000	5,000	1.983 (1.978–2.008)	1.741 (1.741–1.760)	1.645 (1.619–1.651)

A CPU comparison uses $N = 10{,}000$, identical batches and arithmetic precision, and the first three process runs. GPU QID is 1.72–3.32× faster at $n = 1000, 5000, 10{,}000$, while CPU execution is about twice as fast at $n = 128$. The crossover reflects device-execution costs at small batch sizes.

M.2 Reference preprocessing and repeated CPU scoring

The CLIP experiment varies $N \in \{10{,}000, 50{,}000, 100{,}000\}$ and $n \in \{128, 997, 1000, 5000\}$, using a common reference, $M = 100$ directions, generated batches and $\alpha = 1536$. Table 24 reports medians over six runs on the same CPU, separating sorting and integration from computation that also includes projection. QID's interval averages give the lowest computation time from features at every setting. At $N = 100{,}000$ and $n = 128$, this takes 0.181 ms, compared with 0.486 ms using prefix sums and 35.786 ms using sorted projections. Across the grid, the speedup over using prefix sums is 1.36–2.69× including projection and sorting.

Table 24: CPU time to compute QID using three reference representations. Computation from features includes projection, sorting and integration. Entries are median milliseconds over six runs.

		Sorting and integration			Scoring from features		
N	n	Sorted projections	Prefix sums	QID	Sorted projections	Prefix sums	QID
10000	128	3.807	0.284	0.121	3.745	0.337	0.168
10000	997	5.313	2.475	1.294	5.484	2.745	1.577
10000	1000	1.843	2.336	1.311	2.142	2.684	1.603
10000	5000	8.521	11.328	7.959	10.203	13.055	9.572
50000	128	18.314	0.394	0.126	17.335	0.445	0.176
50000	997	19.247	3.432	1.301	19.937	3.728	1.597
50000	1000	3.923	3.221	1.313	4.219	3.574	1.601
50000	5000	10.555	15.758	7.888	12.230	17.437	9.576
100000	128	35.752	0.435	0.132	35.786	0.486	0.181
100000	997	36.517	3.518	1.312	37.668	3.863	1.615
100000	1000	6.118	3.449	1.315	6.430	3.740	1.612
100000	5000	12.783	19.675	7.904	14.487	21.323	9.556

At $N = 100{,}000$, reference sorting takes 200.3 ms and computing its integrals takes 23.0 ms. Preparing coefficients for $n = 128$ takes a further 0.051 ms. The stored coefficients range from 0.098 MiB at $n = 128$ to 3.815 MiB at $n = 5000$. At $n = 1000$, they occupy 0.764 MiB, compared with 76.294 MiB for sorted reference projections. These sizes count reference arrays, with the common projection bank stored separately. Retaining sorted projections permits preparation for additional batch sizes.

We estimate the total time for repeated evaluations as $T_{\text{setup}} + KT_{\text{score}}$ from measured preparation and scoring times, using $K = 1, 10, 100, 1000$ and reporting medians across six runs (Table 25). Each representation includes its required reference preparation, and QID reuses the integrals when preparing coefficients for multiple batch sizes.

Table 25: Estimated reference-preparation and scoring time for K evaluations at $N = 100{,}000$, in milliseconds. Totals use $T_{\text{setup}} + KT_{\text{score}}$ from measured timings, including reference sorting and preparation for each representation.

n	Representation	$K = 1$	$K = 10$	$K = 100$	$K = 1000$
128	Sorted projections	241	565	3786	35993
	Prefix sums	241	246	289	728
	QID	223	225	242	405
997	Sorted projections	243	582	3974	37875
	Prefix sums	245	280	627	4104
	QID	225	240	385	1838
1000	Sorted projections	207	265	846	6633
	Prefix sums	245	278	614	3980
	QID	225	240	385	1836
5000	Sorted projections	215	345	1649	14687
	Prefix sums	262	454	2373	21564
	QID	237	323	1183	9783

M.3 Cost of Gaussian approximations

We compare QID, diagonal fQID and the projected-variance Gaussian H on CPU with prepared references (Table 26). The CLIP experiment uses $N = 50{,}000$, $M = 100$ and SDXL-Turbo one-step outputs, while the Inception experiment uses disjoint COCO reference and generated sets with $M = 100$ and 1000. Medians and interquartile ranges summarize four runs. Replacing generated quantile sorting by moment computation gives larger savings as the batch or direction bank grows: the QID/fQID speedup increases from $5.29\times$ to $12.53\times$ between $n = 128$ and $n = 2000$ in CLIP space, and from $4.01\times$ to $20.08\times$ between $M = 100$ and $M = 1000$ at $n = 512$ in Inception space.

Table 26: CPU scoring time with prepared references of $N = 50{,}000$, in milliseconds. Entries are medians over four runs, with interquartile ranges in parentheses. Gaussian H uses measured projected variances.

Features	M	n	QID	fQID	Gaussian H	QID/fQID
CLIP	100	128	0.166 (0.165–0.170)	0.031 (0.030–0.033)	0.053 (0.052–0.054)	$5.29\times$
CLIP	100	512	0.752 (0.749–0.755)	0.084 (0.083–0.084)	0.162 (0.161–0.162)	$9.00\times$
CLIP	100	2000	3.513 (3.459–3.552)	0.280 (0.279–0.291)	0.603 (0.598–0.609)	$12.53\times$
Inception	100	512	1.169 (1.165–1.177)	0.292 (0.290–0.297)	0.589 (0.582–0.595)	$4.01\times$
Inception	1000	512	11.324 (11.255–11.362)	0.564 (0.556–0.577)	5.100 (5.019–5.144)	$20.08\times$

We also time the Gamma experiment in Appendix L.3, recomputing reference statistics at every evaluation. Across three datasets per dimension, we include transfers and reference computation and exclude sample generation and construction of the 20,000-direction target. fQID is 1.27–$1.35\times$ faster than empirical SW at $L = 1000$ and 1.37–$1.48\times$ faster at $L = 5000$, while empirical SW is faster at $L = 100$ for $d = 500, 1000$. Reference reuse and Gaussian approximation therefore affect different parts of the evaluation cost.

N Additional Related Work

Evaluation Metrics for Generative Models. FID compares Gaussian approximations in embedding space using a reusable reference mean and covariance (Heusel et al., 2017). Its reported values depend on finite-sample bias, preprocessing, implementation, ImageNet-class structure, and the choice of Inception features (Chong and Forsyth, 2020; Parmar et al., 2022; Kynkäänniemi et al., 2023; Stein et al., 2023). Fréchet Wavelet Distance (FWD) applies this moment comparison to wavelet-packet coefficients (Veeramacheneni et al., 2025). KID and other MMD metrics compare feature distributions through kernels (Binkowski et al., 2018; Gretton et al., 2012; Bischoff et al., 2024), while precision–recall metrics characterize sample quality and distribution coverage (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). CMMD combines CLIP embeddings with an RBF-kernel MMD (Jayasumana et al., 2024), while the recent projected-transport metrics MIND, APEX, and CLIP-SWD replace Gaussian moment comparison with averaged one-dimensional Wasserstein distances in feature space (Berthet et al., 2026; Gallegati et al., 2026; Kwon et al., 2026).

Sliced Optimal Transport. Sliced optimal transport reduces high-dimensional transport comparisons to one-dimensional projections, where monotone rearrangement has a closed-form empirical solution (Rabin et al., 2011; Peyré and Cuturi, 2019). The same construction has supported generative modeling losses and evaluation metrics, including sliced Wasserstein generative modeling, sliced Wasserstein autoencoders, and the multiscale patch SWD metric in ProgressiveGAN (Deshpande et al., 2018; Kolouri et al., 2019; Karras et al., 2018). Subsequent work studies statistical properties and fast approximations of sliced probability divergences (Nadjahi et al., 2020; 2021), optimized or learned slicing directions (Nguyen and Ho, 2022), and non-Euclidean variants (Tran et al., 2024; Bonet et al., 2023). Control-variate estimators use projected Gaussian moments to reduce angular Monte Carlo error while retaining empirical slice costs (Nguyen and Ho, 2024).

Transport-Based Embeddings. Linear optimal transport represents distributions by transport coordinates around a fixed reference (Wang et al., 2013; Kolouri et al., 2016a; 2017). Sliced-Wasserstein kernels use quantile-space geometry (Kolouri et al., 2016b), while WEGL and SLoSH apply Wasserstein embeddings to graphs and hashing (Kolouri et al., 2021; Lu et al., 2024). PSWE learns common-reference transport coordinates for set pooling, interpolating variable-size empirical quantiles on a reference grid (Naderializadeh et al., 2021). FourierSWE integrates projected quantiles against sampled cosine functions to obtain finite embeddings whose squared distances estimate sliced-Wasserstein costs (Amir and Dym, 2025). Mixture-Wasserstein and projected mixture distances use structured mixture representations (Delon and Desolneux, 2020; Piening and Beinert, 2025), and WaM applies this approach to generative evaluation by fitting Gaussian mixtures in feature space and reusing the reference parameters (Luzi et al., 2023).