MIND: Monge Inception Distance for Generative Models Evaluation

Quentin Berthet
Google DeepMind

Yu-Han Wu
Google DeepMind
LPSM, Sorbonne Université

Clément Crepy
Google DeepMind

Romuald Elie
Google DeepMind

Klaus Greff
Google DeepMind

Michaël Sander
Google DeepMind

Abstract

We propose the Monge Inception Distance (MIND), a metric for evaluating generative models that addresses key limitations of the widely adopted Fréchet Inception Distance (FID). The MIND metric leverages the sliced Wasserstein distance to compare distributions by averaging one-dimensional optimal transport distances, efficiently computed via sorting. This approach circumvents the estimation of high-dimensional means and covariance matrices, which underlie FID's poor sample complexity and vulnerability to adversarial attacks. We empirically demonstrate three primary advantages: (i) it is more sample-efficient by one order of magnitude, (ii) it is faster to compute by two orders of magnitude, (iii) it is more robust to adversarial attacks such as moment-matching. We show that MIND with 5k samples can replace the evaluation performance of FID with 50k samples, providing high correlation with this standard benchmark and superior discriminative performance. We further demonstrate that even smaller sample sizes (e.g., 1k or 2k) remain highly informative for rapid model iteration.

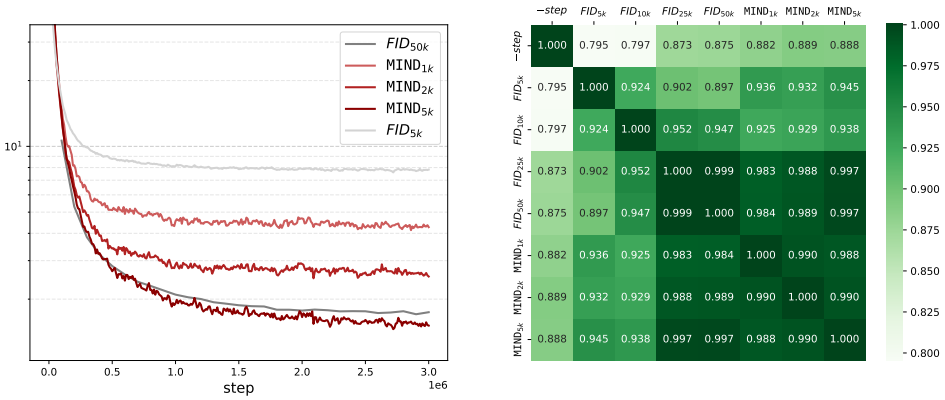


Figure 1: (**Left**) MIND metric during a diffusion model training run on ImageNet-64 (log scale), illustrating how MIND_{5k} can be used to replace FID_{50k}, with a wider dynamic range across early training steps without saturating—see Section 4.1. (**Right**) Correlation with number of training steps - better for MIND_{1k} and MIND_{5k} than FID with 50k samples.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

Generative models, especially diffusion models (Ho et al., 2020), have set new standards in high-quality data synthesis. This progress has spurred innovation across numerous fields, from creative arts to scientific simulation. However, as models grow in complexity, the metrics used to evaluate them have struggled to keep pace (Stein et al., 2023). The de facto standard, the Fréchet Inception Distance (FID) (Heusel et al., 2017) is based on a Gaussian approximation of pre-trained network embeddings, such as the Inception-v3 model (Szegedy et al., 2016).

This metric relies on estimating high-dimensional mean and covariance matrices from inception embeddings. However, this sample-heavy approach typically requires 50k samples (Chong and Forsyth, 2020), creating a significant development bottleneck. Furthermore, because FID only considers the first two moments of the distributions, it is not a true distance metric and is vulnerable to adversarial "hacking" without corresponding visual improvements (Sajjadi et al., 2018).

In this work, we introduce the Monge Inception Distance (`MIND`) addressing these limitations. Based on optimal transport theory and named in honor of Gaspard Monge, who introduced the optimal transport problem (Monge, 1781), `MIND` leverages the sliced Wasserstein distance (Rabin et al., 2011). Instead of relying on Gaussian simplification as in FID, `MIND` reduces the complexity of high-dimensional optimal transport comparison by averaging many one-dimensional projections, where the transport problem is solved exactly via a simple, parallelizable sorting operation. This captures finer distributional details without the statistical instability inherent in high-dimensional matrix estimation - see (Villani, 2008; Peyré and Cuturi, 2019) for a modern perspective on optimal transport theory and applications.

We highlight that this approach yields stable, high-quality evaluation using only 5k samples—an order of magnitude fewer than FID—while offering $100\times$ faster computation and increased robustness to adversarial moment-matching. We statistically validate the performance of `MIND` across various hypothesis testing problems at different sample sizes. While other recent works also explore the Sliced Wasserstein distance (SWD) for image evaluation (for example, Kwon et al., 2026), their experiments focus on static, offline evaluations, analyzing rank-correlation stability down to 10k samples and sensitivity to artificial class-ratio and attribute shifts. In contrast, our empirical evaluation is designed around dynamic, online training monitoring. We highlight that as few as 5k samples (and even 1k-2k for fast iteration) are sufficient for real-time monitoring and active training loop decisions, such as classifier free guidance tuning. Finally, while we present our results using Inception-v3 embeddings to facilitate a direct comparison with the current FID benchmark, the `MIND` metric is fundamentally embedding-agnostic. It is modality-independent and can be seamlessly applied to any representation space, including CLIP (Radford et al., 2021), DINO (Oquab et al., 2024; Siméoni et al., 2025), or specialized embeddings for audio and video synthesis.

Main contributions. In this work, we introduce `MIND`, a metric for improved evaluation of generative models. We demonstrate the following advantages of this metric:

- **Sample Efficiency**: We show that $\texttt{MIND}_{5k}$ provides a stable evaluation that correlates highly with FID_{50k}, enabling reliable benchmarking with $10\times$ fewer samples.
- **Computational Speed and Memory Efficiency**: Due to its reliance on 1D sorting rather than high-dimensional matrix operations, `MIND` is over $100\times$ faster to compute, and requires $10\times$ less memory, facilitating real-time evaluation during training.
- **Metric Robustness**: Since `MIND` is derived from a proper distance, we show that it is significantly more resistant to "metric hacking" via moment-matching attacks that can artificially lower FID.
- **Discriminative Power**: Our experiments show that `MIND` more reliably distinguishes between model checkpoints and identifies subtle image perturbations at low sample sizes.

2 Generative model evaluations

We consider the problem of evaluating a generative model g_θ, that generates outputs by mapping noise $Z \sim \mathcal{N}(0, I)$ to data outputs (e.g. images) $a = g_\theta(Z)$. In standard practice, these outputs are passed through a pre-trained feature extraction model ψ_w to obtain embeddings. For an Inception-v3 model (Szegedy et al., 2016), these embeddings typically reside in dimension $d = 2,048$.

The performance of the model is measured by the statistical distance between the distribution of generated embeddings, $X = \psi_w(g_\theta(Z)) \sim p_\theta$, and the distribution of real dataset embeddings, $Y = \psi_w(D) \sim p_{data}$. In practice, this distance is estimated using finite samples of size n, denoted as the empirical distributions $\hat{p}_{n,\theta}$ and $\hat{p}_{n,data}$ (see Appendix A). Any measure of statistical distance between these distributions can be used, and we consider in this work several inception distances, defined as follows for any distance or divergence Δ between distributions (see, e.g. Cover, 1999).

Definition 2.1 (General - Inception distance)**.** *Let $X = \psi_w(g_\theta(Z)) \sim p_\theta$, $Y = \psi_w(D) \sim p_{data}$. For a distribution distance function Δ, the performance of the model g_θ is given by $\Delta ID(p_\theta, p_{data})$. With a sample of size n, its empirical estimate is the plug-in value $\Delta ID(\hat{p}_{n,\theta}, \hat{p}_{n,data})$.*

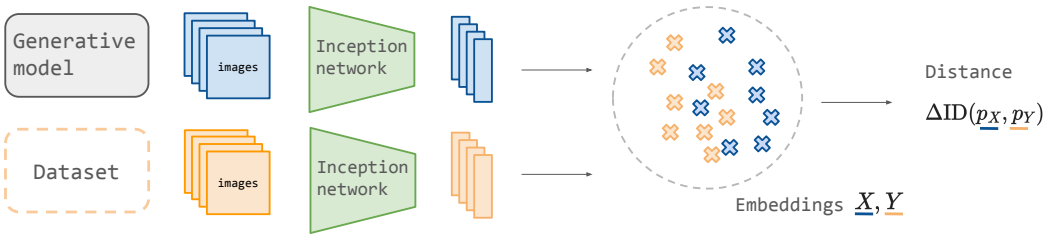


Figure 2: General pipeline for evaluating generative model sampling distance to a dataset.

2.1 Existing method: Fréchet Inception Distance - FID

FID is the most widely adopted instance of an inception distance. It measures the distance between two distributions based on their first two moments: the mean (μ) and covariance (Σ), using the squared 2-Wasserstein distance W_2^2 (see Appendix A.2)

Definition 2.2 (Fréchet Inception Distance - FID)**.** *Let $X = \psi_w(g_\theta(Z)) \sim p_\theta$ and $Y = \psi_w(D) \sim p_{data}$ and μ_X, Σ_X and μ_Y, Σ_Y be the means and covariances of p_θ and p_{data}, respectively. The FID is defined as the squared 2-Wasserstein distance between two fitted Gaussians:*

$$FID(p_\theta, p_{data}) = \|\mu_X - \mu_Y\|^2 + \mathrm{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_Y \Sigma_X)^{1/2}) \,.$$

In practice, this is computed using empirical sample means and covariances $\hat{\mu}_n$ and $\hat{\Sigma}_n$.

This approach was originally motivated as a way to bypass the high computational and statistical complexity of a direct, sample-based Wasserstein distance by using a Gaussian approximation.

Drawbacks Despite its widespread use and role as de facto standard metric, there are several drawbacks in using this distance, as noted in several works (see, e.g. Karras et al., 2017; Stein et al., 2023; Jayasumana et al., 2024; Bischoff et al., 2024; Yang et al., 2026)

- Computing this distance is based on the estimate $\hat{\Sigma}_n$ of a d-by-d covariance. It is rank-deficient for $n \leq d$, creating numerical and statistical issues when estimating the second term, unless the sample size is at least of order d, which is 2048 for inception networks. For images, this means that the sample sizes usually used are 10k or 50k (Bińkowski et al., 2018; Chong and Forsyth, 2020), with a high impact on evaluation time and cost.
- FID is also not a proper distance (Jayasumana et al., 2024): two distributions can have the same mean and covariance and be very different (see, for example, Billingsley, 2017, Section 30). We show that, as a consequence, it is not robust, and this fact can be leveraged to artificially reduce the FID without visually altering the images (see Section 4.4).

We also evaluate other inception-based distances proposed as metrics, as means of comparison, such as the *Maximum mean discrepancy* (MMD), or Sinkhorn divergence. We provide full definitions and a discussion in appendix (Section A.3), and include them in comparisons.

3 Proposed method: Monge Inception Distance - MIND

We propose the following metric to overcome these challenges, based on the sliced Wasserstein distance which has several known advantages, both in terms of statistical and computational complexity. It is an average of the Wasserstein distances of the distribution of projections, over all unit directions (see, e.g. Rabin et al., 2011; Nadjahi, 2021, and Appendix A.2 in this work for details).

The sliced Wasserstein distance is a proper distance - it is equal to 0 if and only if both distributions are equal (Bonnotte, 2013, Proposition 5.1.2). We use this approach for MIND, averaging Wasserstein distances projected along finitely many unit directions.

Definition 3.1 (Monge Inception Distance)**.** *Let $X = \psi_w(g_\theta(Z)) \sim p_\theta$ and $Y = \psi_w(D) \sim p_{data}$, and $\mathcal{U}(S)$ be the uniform distribution on the unit sphere. MIND is given by averaging W_2^2 distances for projections of the distributions along unit directions, with a multiplicative scaling $\alpha = 3d$*

$$\mathit{MIND}(p_\theta, p_{data}) = \alpha \mathbb{E}_{u\sim\mathcal{U}(S)}[W_2^2(u^\intercal p_\theta, u^\intercal p_{data})]\,,$$

where $u^\intercal p_\theta$ (resp. $u^\intercal p_{data}$) is the distribution of $u^\intercal X$ when $X \sim p_\theta$ (resp. of $u^\intercal Y$ when $Y \sim p_{data}$) and d is the data dimension. For finite samples $(X_j)_{j\in[n]}$, $(Y_j)_{j\in[n]}$, random unit directions $(u_i)_{i\in[M]}$ and $\alpha = 3d$, it is given by

$$\mathit{MIND}(\hat{p}_{n,\theta}, \hat{p}_{n,data}) = \frac{\alpha}{M}\sum_{i=1}^{M} W_2^2(u_i^\intercal \hat{p}_{n,\theta}, u_i^\intercal \hat{p}_{n,data}) = \frac{\alpha}{nM}\sum_{i=1}^{M}\sum_{j=1}^{n} |\mathit{sort}(u_i^\intercal X)_j - \mathit{sort}(u_i^\intercal Y)_j|^2\,.$$

Remarks

- MIND relies on two finite sample estimates: the 1D Wasserstein distance over samples X_j and Y_j of size n, and the expectation $\mathbb{E}_{u\sim\mathcal{U}(S)}$ over M random unit directions u_i.
- Although we adopt the name of Inception Distance for consistency with established literature, the formulation of MIND does not depend on the Inception architecture. The mapping function ψ_w can represent any feature extractor; consequently, MIND serves as a general-purpose tool for evaluating distributional similarity across diverse data modalities and embedding models.

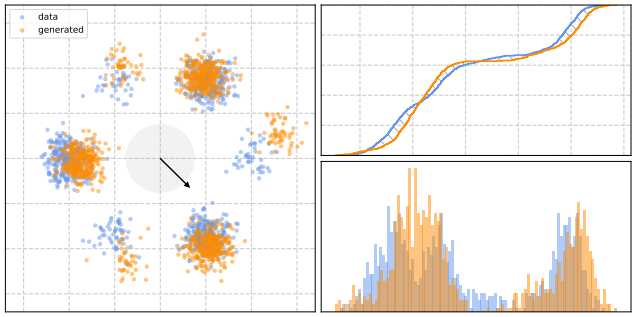


Figure 3: Computation of MIND based on the idea of Sliced Wasserstein, illustrated in 2D with a single projection. (**Left**) Two samples of synthetic embeddings (orange and blue), along with the unit sphere and a random unit direction u. (**Bottom Right**) The two histograms of distributions of the projections along u. (**Top Right**) The associated cumulative distribution functions (cdf), the hatched area is related to 1D Wasserstein distances along u: it is the W_1 distance, and used for all convex costs with pairwise sorted distances.

- This 1D formulation allows a more stable evaluation using an order of magnitude fewer samples – with n of order 5k rather than 50k (see Sections 4.2 and 4.3). Furthermore, because the sliced Wasserstein distance is a proper distance, having matching means and covariance matrices is not sufficient for MIND to be zero, making it inherently robust to moment-matching hacking (see Section 4.4).
- Leveraging the exact solution of 1D transport problem (see, e.g. Peyré and Cuturi, 2019, and Appendix A.2 in this work), the distance simplifies to pair-wise difference between sorted elements:

$$W_2^2(\hat{p}_n, \hat{q}_n) = \frac{1}{n}\sum_{j=1}^{n} |\mathrm{sort}(x)_j - \mathrm{sort}(y)_j|^2\,,$$

where sort : $\mathbb{R}^n \to \mathbb{R}^n$ is the function that maps a vector $x \in \mathbb{R}^n$ to its copy sorted in nondecreasing order, i.e. $\mathrm{sort}(x) = (x_{\sigma(1)}, \ldots, x_{\sigma(n)})^\intercal$ such that $x_{\sigma(1)} \leq \ldots \leq x_{\sigma(n)}$ (ties do not make this function ambiguous). Since the sorting operation runs in $O(n \log n)$ time, MIND avoids the need to estimate or store high-dimensional objects (e.g. $d \times d$ matrices for FID, $n \times n$ matrices for other distances). Similarly to FID, we use the square of the distance rather than taking a square root of the average. We also use a multiplicative scaling factor α—see discussion in Section 4.1.

(a) JAX

```jax
def monge_inception_distance(
  x, y, rng_seed, n_projections=100
):
  """MIND metric.

  Args:
    x: Input generated features.
    y: Ground truth features.
    rng_seed: An integer for the seed of RNG.
    n_projections: Number of projections to use.

  Returns:
    The value of the MIND metric.
  """
  num_samples, d = x.shape
  assert num_samples == y.shape[0]
  ALPHA = 3 * d
  key = jax.random.PRNGKey(rng_seed)
  u_proj = jax.random.normal(key, (n_projections, d))
  u_proj /= jnp.linalg.norm(u_proj, axis=-1, keepdims=True)

  x_proj = u_proj @ x.T
  y_proj = u_proj @ y.T
  dists = jnp.mean((
    jax.lax.top_k(x_proj, num_samples)[0]
    - jax.lax.top_k(y_proj, num_samples)[0]
  ) ** 2, axis=1)

  return ALPHA * jnp.mean(dists)
```

(b) PyTorch

```pytorch
def monge_inception_distance_torch(
  x, y, rng_seed, n_projections=100
):
  """MIND metric.

  Args:
    x: Input generated features.
    y: Ground truth features.
    rng_seed: An integer for the seed of RNG.
    n_projections: Number of projections to use.

  Returns:
    The value of the MIND metric.

  """
  num_samples, d = x.shape
  assert num_samples == y.shape[0]

  ALPHA = 3 * d
  generator = torch.Generator(device=x.device).manual_seed(
    rng_seed
  )
  u_proj = torch.randn(
    (n_projections, d),
    generator=generator,
    dtype=x.dtype,
    device=x.device
  )
  u_proj /= torch.linalg.norm(u_proj, dim=-1, keepdim=True)

  x_proj = u_proj @ x.T
  y_proj = u_proj @ y.T
  dists = torch.mean((
    torch.topk(x_proj, num_samples, dim=-1).values
    - torch.topk(y_proj, num_samples, dim=-1).values
  ) ** 2, dim=1)

  return ALPHA * torch.mean(dists)
```

Figure 4: JAX (a) and PyTorch (b) implementation of MIND.

4 Experiments

As noted above, estimating MIND on two samples of size n is both computationally and conceptually easy. It requires only trivially parallelizable projections and sorting operations. As such, it is particularly adapted to modern accelerated-oriented hardware and software. We provide in Figure 4 both a JAX (Bradbury et al., 2018) and PyTorch (Paszke et al., 2019) implementation, in the form of a short code snippet that can be directly used in an evaluation pipeline. We also provide in Section 4.6 experimental results showcasing the computation time and memory advantages of this algorithm compared to other methods.

4.1 Hyperparameter choices

As noted in Definition 3.1, we scale MIND by a factor $\alpha > 0$ so that its order of magnitude matches that of FID, facilitating comparison and adoption. Throughout our experiments, we denote by $\texttt{MIND}_n$ (or $\texttt{MIND}_n^\alpha$ when emphasizing α, with $\texttt{MIND}_n^1$ for the unscaled metric $\alpha = 1$) the empirical estimate on n samples. The linear scaling $\alpha \propto d$ has a natural geometric justification: projecting onto $u \sim \mathcal{U}(S)$ in $\mathbb{R}^d$ satisfies $\mathbb{E}[uu^\intercal] = \frac{1}{d} I_d$, contracting squared Euclidean norms and Gaussian W_2^2 distances by $1/d$ (see Proposition B.1 in Appendix B.2). Empirically, on ImageNet-64 ($d = 2,048$), $\alpha = 3d \approx 6,000$ closely aligns $\texttt{MIND}_{5k}$ with FID_{50k} later in training, while preserving a wider dynamic range early in training and higher correlation with training steps (Figure 1, right, and Figure 5, left). Evaluating $\texttt{MIND}_n^1$ and FID on truncated Inception-v3 features of varying dimension d (Figure 5, right) confirms this $\propto 1/d$ scaling across dimensions.

4.2 MIND analysis

We illustrate the behavior of the MIND metric by analyzing its dependency on n (the number of samples from the distributions) and M (the number of random projections). Fixing a number $k > 0$ (varies in different settings), we evaluate its ability to correctly order k different distributions $p^1, \dots p^k$ relative to the true data distribution. Specifically, we calculate the probability of error in correctly ranking the sequence of distances $\texttt{MIND}(\hat{p}_n^1, \hat{p}_{n,\text{data}}), \dots, \texttt{MIND}(\hat{p}_n^k, \hat{p}_{n,\text{data}})$. This measures the ability of the metric to distinguish different images from the elements of the dataset. Our observations indicate that $n = 5,000$ samples is sufficient to reliably distinguish these distributions (see Figure 6 and Section 4.3).

We also plot the dependency of the estimated metric on M, the number of uniformly chosen random projections, in Figure 6 (right) and Figure 13a in Appendix C.1. This dependency is easier to analyze, since the Monte-Carlo estimate is obtained by averaging unbiased terms to compute the expected Wasserstein distance. We observe that choosing $M \in [100,\ 1000]$ is sufficient.

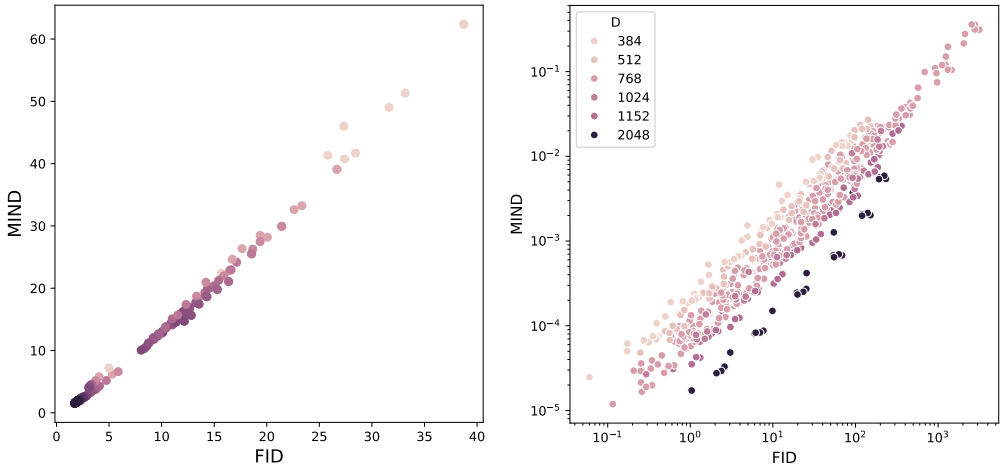


Figure 5: $\texttt{MIND}_{5k}$ and FID_{50k} (and other sample sizes) for a model at different steps of training on ImageNet-64. **Left**: All MIND metrics are rescaled by a factor $\alpha \approx 6,000$ chosen to optimize proximity with FID, colors indicating the step at which the metric is evaluated. **Right**: Unscaled MIND ($\texttt{MIND}^1_n$, $\alpha = 1$) for various embedding dimensions d, colors indicating the dimension in which the metric is evaluated.

4.3 Metric Comparison

Running evaluations during training of a diffusion model, we observe that instead of using FID_{50k} (commonly used post-training because of the cost and time associated with the high sample size), we can use $\texttt{MIND}_{5k}$ (we evaluate their precisions more quantitatively in the rest of this section). As visible in Figure 5, these two metrics are highly correlated, especially later during training.

In order to compare different metrics in a principled fashion, we evaluate how useful they are to distinguish distributions. This provides natural *tasks* for which these metrics can be evaluated as *methods*, through the lens of statistical hypothesis testing. This can also be understood as a comparison in distribution of the metrics (for random samples), rather than a single value. Given sample size n and metric Δ, we state our three statistical hypothesis testings in the following.

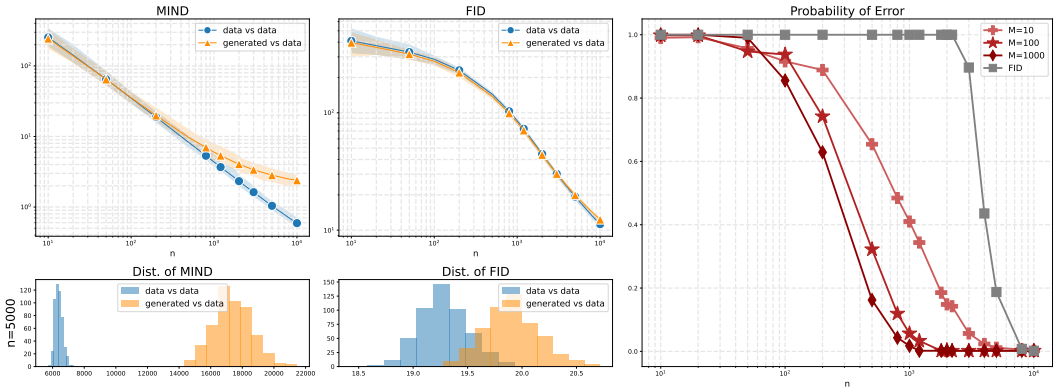


Figure 6: (**Top Left and Middle**) Behavior of the MIND and FID metric in n, to distinguish true images from the dataset (base - in blue) from generated images (model - in orange). (**Bottom Left and Middle**) Histogram of the trials for $n = 5,000$ - A bigger gap is better. (**Right**) Probability of error defined in (1) for three values of $M \in \{10, 100, 1000\}$ and for FID.

4.3.1 Generated vs. True Data

To test whether a metric can distinguish generated samples from real data, we compare p_θ (for a pre-trained model g_θ) and p_{data} by estimating $\Delta(\hat{p}_{n,\theta}, \hat{p}_{n,\text{data}})$. Our diffusion model utilizes a U-Net backbone (Ronneberger et al., 2015; Nichol and Dhariwal, 2021) trained on ImageNet-64. Experiments relying on real data use a fixed set of 100,000 original ImageNet-64 images. Conversely, for evaluations involving data generated from models, we generate a dedicated fixed set of 50,000 samples from each model under evaluation. We also ablate our study and conduct the same experiment using BigGAN (Brock et al., 2019).

We compare the values of the metric under two settings, where $\hat{p}_{n,\theta}$ is a distribution of n samples generated from a trained model, and $\hat{p}_{n,\text{data}}$ and $\hat{p}'_{n,\text{data}}$ are two independent samples of size n from the data. Denoting by $\mathbf{P}(\cdot)$ the empirical probability across 512 independent subsampling trials, the probability of error is defined as:

$$\mathbf{P}\big(\Delta(\hat{p}_{n,\text{data}}, \hat{p}'_{n,\text{data}}) \geq \Delta(\hat{p}_{n,\text{data}}, \hat{p}_{n,\theta})\big)\,. \tag{1}$$

This measures the ability of the metric to distinguish generated images from elements of the dataset. The results are given in Figure 6 and 7. We remark that MIND separates the two distributions with zero error at $n \geq 5,000$ for the diffusion model (Figure 6, left) and $n \geq 500$ for BigGAN (Figure 7, left), whereas FID fails even at $n = 10,000$ on the diffusion model and requires $> 4\times$ as many samples ($n > 2,000$) on BigGAN (Figure 6 and 7, middle and right).

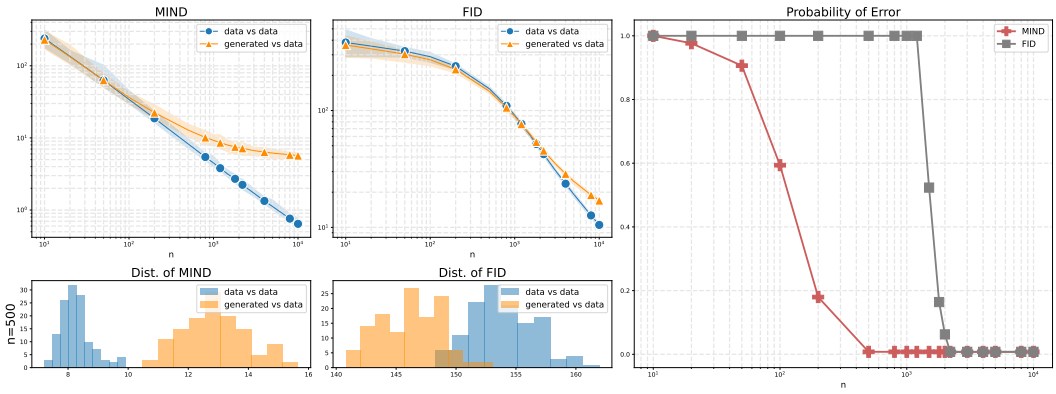


Figure 7: Same experiment as in Figure 6 conducted with GAN.

4.3.2 Monotonicity

We compare MIND with other metrics using a diffusion model g_θ trained on the ImageNet-64 dataset (Deng et al., 2009), from which we selected five models, $g_{\theta_1}, \ldots, g_{\theta_5}$, corresponding to five distinct training checkpoints and generate 50k images with each of them. The probability of error in ranking these checkpoints correctly is:

$$1 - \mathbf{P}\big(\Delta(\hat{p}_{n,\text{data}}, \hat{p}_{n,\theta_1}) \geq \ldots \geq \Delta(\hat{p}_{n,\text{data}}, \hat{p}_{n,\theta_k})\big)\,.$$

For sample sizes n from 10 to 10k (512 independent trials per metric), MIND, MMD, and sliced FID (σFID) reach an error probability below 0.5% by $n = 1,000$ (0% at $n \geq 2,000$), whereas FID requires $n \geq 5,000$ (Figure 8).

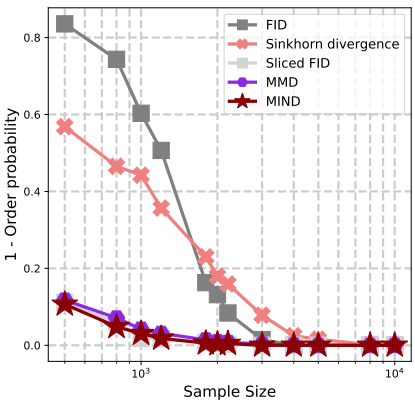


Figure 8: Sample complexity measured by the probability of error.

4.3.3 Perturbations

We consider three types of image perturbations with strictly increasing severity levels $\varepsilon_1 < \cdots < \varepsilon_k$ (512 independent trials each). We measure the performance of each metric using

$$1 - \mathbf{P}\big(\Delta(\hat{p}_{n,\text{data}}, \hat{p}_{n,\text{data},\varepsilon_1}) \leq \ldots \leq \Delta(\hat{p}_{n,\text{data}}, \hat{p}_{n,\text{data},\varepsilon_k})\big)\,.$$

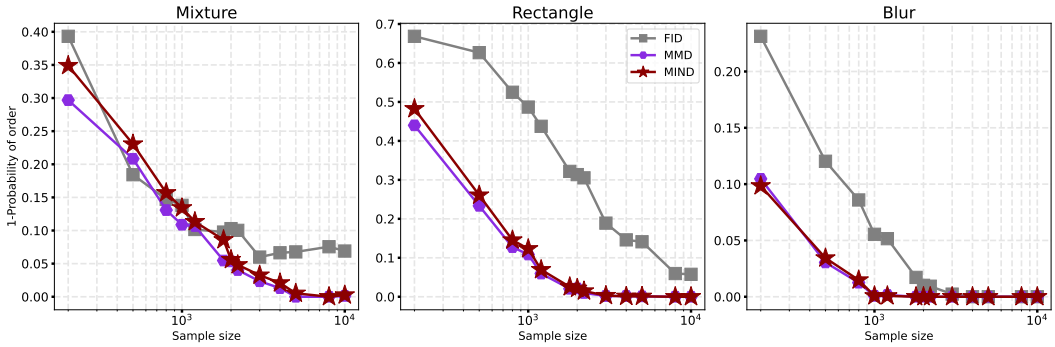


Figure 9: Sample complexity measured by the probability of detecting a small perturbation.

which is the probability of failing to order all perturbation levels. The results are summarized in Figure 9. We highlight that, for $n \geq 5,000$, MIND achieves the same level of performance as MMD while FID is worse on all tasks.

Gaussian blur. For $\varepsilon \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, we apply a Gaussian filter of standard deviation ε to ImageNet-64.

Rectangle. For $\varepsilon \in \{0.05, 0.1, 0.15, 0.2\}$, we place 5 random 10×10 squares of opacity ε in each ImageNet-64 image.

Mixture of datasets. For $\varepsilon \in \{1\%, 3\%, 5\%, 7\%, 10\%\}$, we mix $(1 - \varepsilon)$ of ImageNet-64 with ε of CelebA (Liu et al., 2015).

4.3.4 Guidance Strength Selection

We show that MIND reliably identifies the optimal classifier-free guidance (CFG, Ho and Salimans, 2022) strength while requiring substantially fewer generated samples. Recall that a diffusion model (s) steers conditional sampling at guidance strength ω by extrapolating between conditional and unconditional score estimates,

$$\tilde{s} = (1 + \omega)s(z, c) - \omega s(z)\,.$$

We note that the optimal strength is typically selected with metric-based methods, which demands a large amount of generated samples. In contrast, evaluating across a sweep of guidance strengths on ImageNet, we observe that MIND accurately recovers the optimal guidance strength using only a fraction of the samples required by FID (see Figure 10).

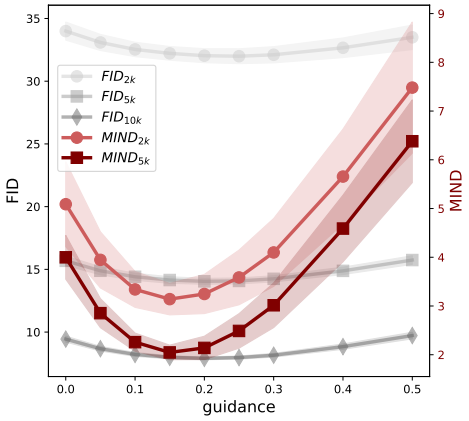


Figure 10: We vary the CFG strength and recover the same optimal strength for MIND at a smaller sample size as for FID.

4.3.5 Early Stopping

Beyond tracking training steps, evaluation metrics are also used to detect when reverse-diffusion sampling should be stopped early (Jayasumana et al., 2024; Wu et al., 2025; Hurault et al., 2025). Varying the stopping step t out of 250 on ImageNet, Figure 11 shows that, as a function of denoising steps, MIND at a 2k sample size provides the same distinguishing ability as its 10k counterpart, with overlapping confidence intervals throughout the trajectory. On the other hand, FID cannot: small-sample bias shifts and flattens FID_{2k} relative to FID_{10k} in the final denoising regime.

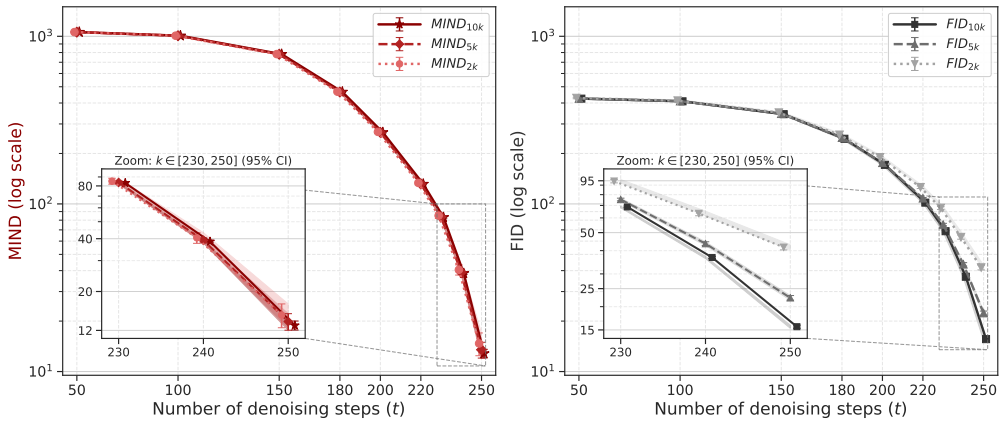


Figure 11: MIND and FID (log scale, mean $\pm$ 95% bootstrap CI) on ImageNet-64 when stopping a 250-step diffusion sampler early after t denoising steps. MIND at $n = 2$k matches its 10k trajectory across all t (inset: $t \in [230, 250]$), whereas FID suffers from strong upward bias at small n.

4.4 Robustness to Metric Hacking With Moment Matching

As mentioned above, one of the weaknesses of FID is that it is not a proper distance. Indeed, since this metric is only a function of the means and covariances of the considered distributions, if p and q share the same first and second moments, then $\text{FID}(p, q) = 0$. This is not the case for proper metrics deriving from distances such as MIND. We leverage this fact to create an artificial distribution of samples that have a desired mean and covariance. This construction is based on the following property whose proof is in Appendix B.1.

Proposition 4.1. *Let p be a target distribution over $\mathbb{R}^d$ with mean μ and covariance $\Sigma = USU^\intercal = \sum_{i=1}^r \lambda_i u_i u_i^\intercal$. For $i \in [r]$ and $\alpha = \sqrt{\boldsymbol{Tr}(\Sigma)}$, define the $2r$ vectors $v_i^{(+)} = \mu + \alpha u_i$ and $v_i^{(-)} = \mu - \alpha u_i$, and weights $\pi_i^{(+)} = \pi_i^{(-)} = \pi_i := \lambda_i / (2\boldsymbol{Tr}(\Sigma))$ (which are nonnegative and sum to 1). Let δ_v denote the Dirac delta at $v \in \mathbb{R}^d$, and let $\hat{q} = \sum_{i=1}^r \pi_i^{(+)} \delta_{v_i^{(+)}} + \sum_{i=1}^r \pi_i^{(-)} \delta_{v_i^{(-)}}$ be the discrete distribution on $\{v_i^{(+)}, v_i^{(-)}\}_{i=1}^r$. Then*

$$\mathbb{E}_{\hat{q}}[v] = \mu\,, \quad \mathbb{E}_{\hat{q}}[(v-\mu)(v-\mu)^\intercal] = \Sigma\,, \quad \textit{FID}(\hat{q}, p) = 0\,.$$

4.5 Moment Matching Procedure

Proposition 4.1 can be leveraged to perform *metric hacking* with moment matching: For a batch of n images a^0 with embedding distribution $\hat{q}^0$ we construct $a = a^0 + \varepsilon$ whose embeddings have distribution $\hat{q}$ such that the metric $\Delta(\hat{q}, \hat{p}_{n,\text{data}})$ is much smaller than $\Delta(\hat{q}^0, \hat{p}_{n,\text{data}})$ (for some data distribution p_{data}) by optimizing the moments of these embeddings, using the target vectors given by Proposition 4.1 with no visually discernible alteration (see Figure 14 in Appendix C.4).

Metric	ratio
FID	11.2%
μFID	2.6%
σFID	4.2%
MMD	12.2%
MIND	31.1%

Table 1: Robustness of several metrics under moment matching

For $n = 2r$, a batch $a^0 \in \mathbb{R}^{2r \times [\text{dims}]}$ of $2r$ images of shape [dims] (e.g. $[512, 512, 3]$), and target $(v_1, \ldots, v_{2r}) = (v_1^{(+)}, \ldots, v_r^{(+)}, v_1^{(-)}, \ldots, v_r^{(-)})$ from Proposition 4.1, we minimize

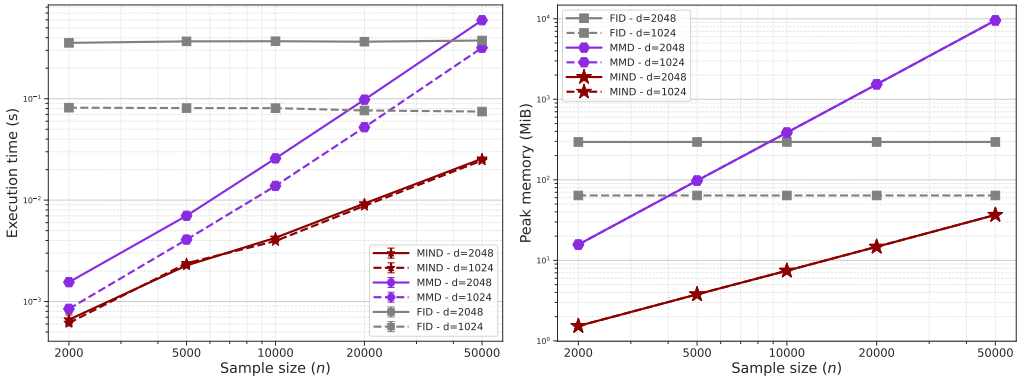
$$\min_{a \in \mathbb{R}^{2r \times [\text{dims}]}} \ell(a) = \min_{a \in \mathbb{R}^{2r \times [\text{dims}]}} \sum_{i=1}^{2r} \|\psi_w(a_i) - v_i\|^2\,.$$

We initialize a at a^0 ($2r$ copies of the same image). Since $\ell(a)$ separates across images a_i, each a_i is optimized independently in parallel via gradient methods. If $\ell(a) = 0$, by Proposition 4.1 the weighted empirical distribution $\hat{q}$ of $\psi_w(a_i)$ with weights π_i satisfies $\text{FID}(\hat{q}, p_{\text{data}}) = 0$, and optimizing $\ell(a)$ reduces empirical FID significantly.

Using a full-rank batch $r = 2048$ ($n = 2r = 4096$), $M = 1000$ for MIND, and 50k samples for p_{data}, Table 1 reports ratio $= \Delta(\hat{q}, \hat{p}_{n,\text{data}})/\Delta(\hat{q}^0, \hat{p}_{n,\text{data}})$. Baseline metrics retain only $\sim 11\%$ of their initial value for FID/MMD (and $< 5\%$ for μFID/σFID), whereas MIND retains 31.1%.

4.6 Computation Time and Peak Memory Comparison

We compare the running time of *computing* the different metrics on TPUv4, given two sets of embeddings of size n (excluding sample generation time, which is linear in n). In Figure 12a, computing MIND at its recommended size ($n = 5$k) is $\approx 150\times$ faster than FID_{50k} (over 2 orders of magnitude, and also $> 100\times$ faster at equal n), on top of the $10\times$ reduction in sampling time.



(a) Execution time

(b) Peak Memory

Figure 12: Walltime and peak memory comparison for MIND, MMD, and FID

We also compare peak memory during metric computation on TPUv4 (excluding input data memory; see Appendix C.2 for total memory). In Figure 12b, computing MIND at $n = 5$k requires over an order of magnitude less memory than MMD or FID at $n = 5$k, and nearly two orders of magnitude less than at $n = 50$k (curves for MMD and MIND overlap across d).

Conclusion

In this work, we introduced MIND, a metric for evaluating generative models that addresses statistical and computational limitations of FID. Our empirical results demonstrate that MIND is faster to compute and achieves stable evaluations with sample sizes as low as 2k, compared to the 50k typically required for FID. Furthermore, as a proper distance metric, MIND exhibits better robustness to moment-matching adversarial attacks than other metrics, while being affected by it. As a purely statistical metric, MIND measures the distributional distance to a reference dataset (such as the training data). It is not designed to evaluate other qualitative aspects of the generated images, such as visual aesthetics or text legibility. We believe this metric provides a rigorous, efficient, and reliable standard for assessing the quality of modern generative models.

References

Billingsley, P. (2017). *Probability and measure*. John Wiley & Sons.

Bischoff, S., Darcher, A., Deistler, M., Gao, R., Gerken, F., Gloeckler, M., Haxel, L., Kapoor, J., Lappalainen, J. K., Macke, J. H., et al. (2024). A practical guide to sample-based statistical distances for evaluating generative models in science. *arXiv:2403.12636*.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. In *The Sixth International Conference on Learning Representations*.

Bonnotte, N. (2013). *Unidimensional and evolution methods for optimal transportation*. PhD thesis, Université Paris Sud-Paris XI; Scuola normale superiore (Pise, Italie).

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. (2018). JAX: composable transformations of Python+ NumPy programs.

Brock, A., Donahue, J., and Simonyan, K. (2019). Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*.

Chong, M. J. and Forsyth, D. (2020). Effectively unbiased FID and inception score and where to find them. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6070–6079.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, pages 248–255, Los Alamitos, CA, USA. IEEE Computer Society.

Dudley, R. M. (1969). The speed of mean Glivenko-Cantelli convergence. *Annals of Mathematical Statistics*, 40:40–50.

Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. PMLR.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

Ho, J. and Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv:2207.12598*.

Hurault, S., Terris, M., Moreau, T., and Peyré, G. (2025). From score matching to diffusion: A fine-grained error analysis in the Gaussian setting. *arXiv:2503.11615*.

Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., and Kumar, S. (2024). Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315.

Kantorovich, L. V. (1942). On the translocation of masses. *Doklady Akademii Nauk SSSR*, 37(7-8):227–229.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2017). Progressive growing of GANs for improved quality, stability, and variation. *arXiv:1710.10196*.

Kwon, H. K., Yang, J., and Chae, M. (2026). Evaluating image generation models via sliced Wasserstein distance. *Journal of the Korean Statistical Society*, 55(2):730–750.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704.

Nadjahi, K. (2021). *Sliced-Wasserstein distance for large-scale machine learning : theory, methodology and extensions*. Theses, Institut Polytechnique de Paris.

Nadjahi, K., Durmus, A., Chizat, L., Kolouri, S., Shahrampour, S., and Simsekli, U. (2020). Statistical and topological properties of sliced probability divergences. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20802–20812. Curran Associates, Inc.

Nichol, A. Q. and Dhariwal, P. (2021). Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., and Bojanowski, P. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. (2018). Assessing generative models via precision and recall. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training GANs. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al. (2025). DINOv3. *arXiv:2508.10104*.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. (2023). Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 3732–3784. Curran Associates, Inc.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826. IEEE.

Villani, C. (2008). *Optimal transport: old and new*, volume 338. Springer.

Wu, Y.-H., Berthet, Q., Biau, G., Boyer, C., Elie, R., and Marion, P. (2025). Optimal stopping in latent diffusion models. *arXiv:2510.08409*.

Yang, J., Geng, Z., Ju, X., Tian, Y., and Wang, Y. (2026). Representation Fréchet loss for visual generation. *arXiv:2604.28190*.

A Definitions

A.1 Empirical measures

Definition A.1. *For a sample $Y_1, \dots, Y_n$ of size n from some data distribution p_{data}, we denote by $\hat{p}_{n,data}$ the* empirical distribution *of the Y_is, defined by*

$$\hat{p}_{n,data} = \frac{1}{n} \sum_j \delta_{Y_j} .$$

Similarly, for $X_1, \dots, X_n$ from p_θ we denote by $\hat{p}_{n,\theta}$ the empirical distribution of the X_is

$$\hat{p}_{n,\theta} = \frac{1}{n} \sum_j \delta_{X_j} .$$

A.2 Optimal transport

Definition A.2. *The optimal transport problem for Euclidean cost, also called the 2-Wasserstein distance is defined for two probability distributions $p, q \in \mathcal{P}_2(\mathbb{R}^d)$ with finite second moments as*

$$\begin{aligned} W_2^2(p,q) &= \min_{T : T_\# p = q} \mathbb{E}_{X \sim p}[\|X - T(X)\|^2] \\ &= \min_{\pi \in \Pi(p,q)} \mathbb{E}_\pi[\|X - Y\|^2] , \end{aligned} \tag{2}$$

The first formulation is known as the Monge *formulation (Monge, 1781), and the second one as the* Kantorovich *formulation (Kantorovich, 1942), with the equivalence holding when p, q are absolutely continuous, or discrete uniform samples of the same finite size.*

Note that this distance can be approximated with sample access to p and q by plugging directly these samples empirical measures $\hat{p}_n$ of the X_j and $\hat{q}_n$ of the Y_j. However, this approach suffers from two issues: It suffers from a curse of dimensionality, and the convergence of the estimate $W_2^2(\hat{p}_n, \hat{q}_n)$ to $W_2^2(p,q)$ is slow, in $n^{-1/d}$ (Dudley, 1969), it is slow to compute in general, with a worst case super cubic cost of n^3 for the Hungarian algorithm, and methods based on the Sinkhorn algorithm in $n^2 \log(n)$. The latter is motivated by an entropic-regularized formulation (Cuturi, 2013)

$$W_{2,\varepsilon}^2 = \min_{\pi \in \Pi(p,q)} \mathbb{E}_\pi[\|X - Y\|^2] - \varepsilon H(\pi) .$$

An interesting exception is the one-dimensional case: when $d = 1$, the solution of (2) is given by

$$W_2^2(\hat{p}_n, \hat{q}_n) = \frac{1}{n} \sum_{j=1}^n |\text{sort}(x)_j - \text{sort}(y)_j|^2$$

where $\text{sort} : \mathbb{R}^n \to \mathbb{R}^n$ is the function that maps a vector $x \in \mathbb{R}^n$ to its copy sorted in nondecreasing order $\text{sort}(x) = (x_{\sigma(1)}, \dots, x_{\sigma(n)})^\intercal$ such that $x_{\sigma(1)} \leq \dots \leq x_{\sigma(n)}$ (ties do not make this function ambiguous). It can therefore be computed in time of order $n \log n$. This can be leveraged for $d > 1$ by considering the average Wasserstein distance over uniformly random unit directions. This is called the *sliced Wasserstein distance*.

Definition A.3. *The sliced Wasserstein distance (Rabin et al., 2011; Nadjahi, 2021) is defined as the average of the Wasserstein distances over 1-d projections along $u \sim \mathcal{U}(S)$ a uniformly random unit direction*

$$SW_2^2(p,q) = \mathbb{E}_{u \sim \mathcal{U}(S)}[W_2^2(u^\intercal p, u^\intercal q)]$$

where $u^\intercal p$ (resp. $u^\intercal q$) denotes the distribution of $u^\intercal X$ when $X \sim p$ (resp. $u^\intercal Y$ when $Y \sim q$).

The sliced Wasserstein distance is still a distance between distributions. It can also be easily estimated from samples, given empirical measures $\hat{p}_n$ and $\hat{q}_n$ and M i.i.d. unit vectors $u_1, \dots u_M$

$$\begin{aligned} \hat{SW}_{2,M}^2(\hat{p}_n, \hat{q}_n) &= \frac{1}{M} \sum_{i=1}^M W_2^2(u_i^\intercal \hat{p}_n, u_i^\intercal \hat{q}_n) \\ &= \frac{1}{Mn} \sum_{i=1}^M \sum_{j=1}^n |\text{sort}(u_i^\intercal X)_j - \text{sort}(u_i^\intercal Y)_j|^2 , \end{aligned}$$

One of the advantages of this approach is the relaxed computational load: computing Wasserstein distances in 1D only requires to sort all the elements in the sample, which can be done in order of $n \log n$ time, and is highly parallelizable, allowing to perform M of these operations with little to no overhead, and M projections from dimension d to 1, for n points each time.

In particular, under mild assumptions, the sample complexity of estimating the sliced Wasserstein distance does not depend on the dimension of the problem (Nadjahi et al., 2020), in contrast to the standard Wasserstein distance for which the sample complexity grows exponentially with the dimension.

We finally note that as in Rabin et al. (2011), we consider the average of the *squared* 1-D Wasserstein distances, and would do so for other ℓ_p, $p \neq 2$ norm costs.

A.3 Metric comparison

Remarks about FID

- The trace term in the FID formula can also be written in the following symmetric form; both expressions are equal:

$$W_2^2(\mathcal{N}(\mu_X, \Sigma_X), \mathcal{N}(\mu_Y, \Sigma_Y)) = \|\mu_X - \mu_Y\|^2 + \mathrm{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_Y^{1/2}\Sigma_X\Sigma_Y^{1/2})^{1/2})\,.$$

- In practice, the expectations are obtained based on a finite sample $X_1, \dots, X_n$ from a generative model and $Y_1, \dots, Y_n$ from a dataset, and we actually compute the plug-in estimate

$$\mathrm{FID}(\hat{p}_{n,\theta}, \hat{p}_{n,\mathrm{data}}) = \|\hat{\mu}_X - \hat{\mu}_Y\|^2 + \mathrm{tr}(\hat{\Sigma}_X + \hat{\Sigma}_Y - 2(\hat{\Sigma}_Y\hat{\Sigma}_X)^{1/2})\,.$$

- This distance is motivated by the *Wasserstein distance* W_2^2 (see, e.g. Villani, 2008, and Appendix A.2 in this work), obtained by solving an optimal transport problem, with a square Euclidean distance cost. For FID, this distance is applied to two fitted Gaussian distributions rather than to the sample distributions $\hat{p}_{n,\theta}$ and $\hat{p}_{n,\mathrm{data}}$.
- Using this method, rather than a sample-based estimate of $W_2^2(\hat{p}_{n,\theta}, \hat{p}_{n,\mathrm{data}})$, allows to overcome the two main obstacles when using the Wasserstein distance between two distributions based on sample access: *statistical* and *computational complexity*. Computing the FID only requires to estimate the mean and covariance matrices, and to perform a conceptually simple, closed-form computation.

Definition A.4 (mean FID)**.** *Let $X = \psi_w(g_\theta(Z)) \sim p_\theta$ and $Y = \psi_w(D) \sim p_{data}$,*

$$\mu_X = \mathbb{E}_{p_\theta}[X]\,, \quad \mu_Y = \mathbb{E}_{p_{data}}[Y]\,.$$

The mean FID (that we denote by μFID) is defined as

$$\mu FID(p_\theta, p_{data}) = \|\mu_X - \mu_Y\|^2\,.$$

Remarks

- Much like for FID, it is also very easy to estimate the mean FID from a finite sample with the plug-in empirical measures $\mu\mathrm{FID}(\hat{p}_{n,\theta}, \hat{p}_{n,\mathrm{data}}) = \|\hat{\mu}_X - \hat{\mu}_Y\|^2$.
- We show in Section 4.3 that the sample complexity of μFID is much lower than that of FID - this probably stems from the fact that only a vector of size d must be evaluated rather than a d-by-d matrix.
- We show in Section 4.4 that it is even less robust than FID.

Definition A.5 (Sliced FID)**.** *Let $X = \psi_w(g_\theta(Z)) \sim p_\theta$ and $Y = \psi_w(D) \sim p_{data}$, the sliced FID (that we denote by σFID) is defined as*

$$\sigma FID(p_\theta, p_{data}) = \mathbb{E}_{u \sim \mathcal{U}(S)}[FID(u^\intercal p_\theta, u^\intercal p_{data})]\,.$$

For finite samples $(X_j)_{j\in[n]}$, $(Y_j)_{j\in[n]}$ *and* $(u_i)_{i\in[M]}$ *it can be estimated by*

$$\begin{aligned}\sigma\mathit{FID}(\hat{p}_{n,\theta}, \hat{p}_{n,\mathit{data}}) &= \frac{1}{M}\sum_{i=1}^{M} \mathit{FID}(u_i^\intercal \hat{p}_n, u_i^\intercal \hat{q}_n) \\ &= \frac{1}{M}\sum_{i=1}^{M} \left\{(u_i^\intercal \hat{\mu}_{n,X} - u_i^\intercal \hat{\mu}_{n,Y})^2 \right. \\ &\qquad \left. + (\hat{\sigma}_{n,u_i^\intercal X} - \hat{\sigma}_{n,u_i^\intercal Y})^2\right\},\end{aligned}$$

where $\hat{\mu}_{n,X}, \hat{\mu}_{n,Y} \in \mathbb{R}^d$ *denote the empirical means and* $\hat{\sigma}_{n,u_i^\intercal X}, \hat{\sigma}_{n,u_i^\intercal Y} \in \mathbb{R}_+$ *denote the empirical standard deviations of the projected samples* $(u_i^\intercal X_j)_{j=1}^n$ *and* $(u_i^\intercal Y_j)_{j=1}^n$.

Remarks

- While very easy to estimate from samples, we also show that it suffers from the same robustness issues as the FID.

Definition A.6 (Sinkhorn divergence (Genevay et al., 2018))**.** *For two distributions, we denote by* $W_\varepsilon(p, q)$ *the value of the entropic-regularized optimal transport problem between* p *and* q *(see, e.g. Peyré and Cuturi, 2019, and Appendix A.2 in this work). Let* $X = \psi_w(g_\theta(Z)) \sim p_\theta$ *and* $Y = \psi_w(D) \sim p_{\mathit{data}}$*, the Sinkhorn Divergence Inception Distance (SDID) is defined by*

$$\mathit{SDID}_\varepsilon(p_\theta, p_{\mathit{data}}) = W_\varepsilon(p_\theta, p_{\mathit{data}}) - \frac{1}{2}W_\varepsilon(p_\theta, p_\theta) - \frac{1}{2}W_\varepsilon(p_{\mathit{data}}, p_{\mathit{data}})\,.$$

Remarks

- For finite samples, the empirical measures $\hat{p}_{n,\theta}, \hat{p}_{n,\text{data}}$ can be split in $\hat{p}_{1,n,\theta}, \hat{p}_{2,n,\theta}, \hat{p}_{1,n,\text{data}}, \hat{p}_{2,n,\text{data}}$ and the divergence can be estimated by

$$\begin{aligned}\text{SDID}_\varepsilon(\hat{p}_{n,\theta}, \hat{p}_{n,\text{data}}) &= W_\varepsilon(\hat{p}_{1,n,\theta}, \hat{p}_{1,n,\text{data}}) \\ &\quad - \frac{1}{2}W_\varepsilon(\hat{p}_{1,n,\theta}, \hat{p}_{2,n,\theta}) \\ &\quad - \frac{1}{2}W_\varepsilon(\hat{p}_{1,n,\text{data}}, \hat{p}_{2,n,\text{data}})\,.\end{aligned}$$

  SDID can be computed by computing each entropic regularized optimal transport problem with a fast GPU-friendly alternate projection method, called Sinkhorn's algorithm (Cuturi, 2013).

- In practice, to overcome a curse of dimensionality, we have found it better to estimate the correction term from two independent samples $\hat{p}_{1,n}, \hat{p}_{2,n}$. This concretely doubles the required sample size. We have found this metric to be much more robust than FID, and to require a smaller sample size, but of a similar order (ignoring this doubling) - see Section 4.3.

Definition A.7 (Maximum mean discrepancy - MMD (Gretton et al., 2012; Jayasumana et al., 2024))**.** *Let* $X = \psi_w(g_\theta(Z)) \sim p_\theta$ *and* $Y = \psi_w(D) \sim p_{\mathit{data}}$*, and the kernel function* $k_\sigma(x, y) = \exp(-\|x - y\|^2/\sigma)$*, for* $\sigma > 0$*, the MMD is defined as*

$$\begin{aligned}\mathit{MMD}(p_\theta, p_{\mathit{data}}) &= \mathbb{E}_{X,X'\sim p_\theta}[k_\sigma(X, X')] - 2\mathbb{E}_{X\sim p_\theta, Y\sim p_{\mathit{data}}}[k_\sigma(X, Y)] \\ &\quad + \mathbb{E}_{Y,Y'\sim p_{\mathit{data}}}[k_\sigma(Y, Y')]\,.\end{aligned}$$

Remarks

- Since it is defined as a two-sample mean, the MMD can also be estimated quickly from empirical distributions $\hat{p}_{n,\theta}$ and $\hat{p}_{n,\text{data}}$.
- One of the drawbacks of this metric is the need to select a kernel bandwidth hyperparameter $\sigma > 0$ (typically chosen via the median pairwise distance heuristic on the reference dataset).
- Another drawback is the computational and memory cost, as an $n \times n$ kernel matrix must be formed in $O(n^2 d)$ time and $O(n^2)$ memory.
- The Sinkhorn divergence and MMD are related: when $\varepsilon \to +\infty$, we have that

$$\text{SDID}_\varepsilon \to \frac{1}{2}\text{MMD}_{-\|\cdot\|^2}$$

where the kernel function k is given by the negative squared Euclidean distance (rather than a Gaussian kernel).

- The use of this metric, with another embedding network, is recommended in (Jayasumana et al., 2024).

B Proofs

B.1 Proof of Proposition 4.1

Proof. Recall that the target distribution p has mean μ and covariance Σ. We construct the discrete distribution $\hat{q} = \sum_{i=1}^{r} \pi_i(\delta_{v_i^{(+)}} + \delta_{v_i^{(-)}})$ using the $2r$ vectors $v_i^{(+)} = \mu + \alpha u_i$ and $v_i^{(-)} = \mu - \alpha u_i$, where each vector is assigned probability $\pi_i^{(+)} = \pi_i^{(-)} = \pi_i = \lambda_i/(2\operatorname{tr}(\Sigma))$, and $\alpha = \sqrt{\operatorname{tr}(\Sigma)}$.

1. Mean of $\hat{q}$ By the definition of expectation for a discrete distribution:

$$\begin{aligned}\mathbb{E}_{\hat{q}}[v] &= \sum_{i=1}^{r} \pi_i v_i^{(+)} + \sum_{i=1}^{r} \pi_i v_i^{(-)} \\ &= \sum_{i=1}^{r} \pi_i(\mu + \alpha u_i) + \sum_{i=1}^{r} \pi_i(\mu - \alpha u_i) \\ &= \sum_{i=1}^{r} \pi_i(2\mu) = \mu \sum_{i=1}^{r} 2\pi_i.\end{aligned}$$

Since $2\pi_i = \lambda_i/\operatorname{tr}(\Sigma)$ and $\sum_{i=1}^{r} \lambda_i = \operatorname{tr}(\Sigma)$, it follows that $\sum 2\pi_i = 1$, hence $\mathbb{E}_{\hat{q}}[v] = \mu$.

2. Covariance of $\hat{q}$ The covariance of $\hat{q}$ is given by $\mathbb{E}_{\hat{q}}[(v-\mu)(v-\mu)^\intercal]$:

$$\begin{aligned}\operatorname{Cov}_{\hat{q}}(v) &= \sum_{i=1}^{r} \pi_i(v_i^{(+)} - \mu)(v_i^{(+)} - \mu)^\intercal + \sum_{i=1}^{r} \pi_i(v_i^{(-)} - \mu)(v_i^{(-)} - \mu)^\intercal \\ &= \sum_{i=1}^{r} \pi_i(\alpha u_i)(\alpha u_i)^\intercal + \sum_{i=1}^{r} \pi_i(-\alpha u_i)(-\alpha u_i)^\intercal \\ &= \sum_{i=1}^{r} 2\pi_i \alpha^2 u_i u_i^\intercal.\end{aligned}$$

Substituting $\alpha^2 = \operatorname{tr}(\Sigma)$ and $2\pi_i = \lambda_i/\operatorname{tr}(\Sigma)$:

$$\begin{aligned}\operatorname{Cov}_{\hat{q}}(v) &= \sum_{i=1}^{r} \left(\frac{\lambda_i}{\operatorname{tr}(\Sigma)}\right) \operatorname{tr}(\Sigma) u_i u_i^\intercal \\ &= \sum_{i=1}^{r} \lambda_i u_i u_i^\intercal = \Sigma.\end{aligned}$$

3. FID Value The FID between two distributions is defined as the 2-Wasserstein distance between their associated Gaussians (Heusel et al., 2017). Consequently, FID is strictly a function of the first two moments. Since $\mathbb{E}_{\hat{q}}[v] = \mu_p$ and $\operatorname{Cov}_{\hat{q}}(v) = \Sigma_p$, the means and covariances match exactly:

$$\begin{aligned}\operatorname{FID}(\hat{q}, p) &= \|\mu - \mu\|^2 + \operatorname{tr}(\Sigma + \Sigma - 2(\Sigma\Sigma)^{1/2}) \\ &= 0 + \operatorname{tr}(2\Sigma - 2\Sigma) = 0.\end{aligned}$$

This concludes the proof. □

B.2 Scaling factor for `MIND`

We formalize here the relationship between the unscaled sliced 2-Wasserstein distance (`MIND`[1], with $\alpha = 1$) and FID in dimension d, justifying the linear scaling $\alpha \propto d$ discussed in Section 4.1.

Proposition B.1 (Dimensional scaling of MIND vs. FID)**.** *Let $p = \mathcal{N}(\mu_X, \Sigma_X)$ and $q = \mathcal{N}(\mu_Y, \Sigma_Y)$ be two Gaussian distributions on $\mathbb{R}^d$, and let $SW_2^2(p,q) = \mathbb{E}_{u\sim\mathcal{U}(S)}[W_2^2(u^\intercal p, u^\intercal q)]$ denote the unscaled MIND distance ($\alpha = 1$). Then:*

$$SW_2^2(p,q) = \frac{1}{d}\|\mu_X - \mu_Y\|^2 + \mathbb{E}_{u\sim\mathcal{U}(S)}\left[\left(\sqrt{u^\intercal\Sigma_X u} - \sqrt{u^\intercal\Sigma_Y u}\right)^2\right]. \tag{3}$$

In particular:

1. ***(Isotropic covariances)** If $\Sigma_X = \sigma_X^2 I_d$ and $\Sigma_Y = \sigma_Y^2 I_d$, then $SW_2^2(p,q) = \frac{1}{d}FID(p,q)$ holds with equality.*

2. ***(General covariances—upper bound)** By Cauchy–Schwarz and $\Sigma_Y^{1/2}\Sigma_X\Sigma_Y^{1/2} \succeq 0$, for any $\Sigma_X, \Sigma_Y \succeq 0$,*

$$\mathbb{E}_{u\sim\mathcal{U}(S)}\left[\left(\sqrt{u^\intercal\Sigma_X u} - \sqrt{u^\intercal\Sigma_Y u}\right)^2\right] \le \frac{1}{d}\operatorname{tr}(\Sigma_X + \Sigma_Y).$$

Moreover, when Σ_X and Σ_Y commute and share condition number bounded by κ, or in the high-dimensional regime where the quadratic forms $u^\intercal\Sigma_X u$ and $u^\intercal\Sigma_Y u$ concentrate around $\frac{1}{d}\operatorname{tr}(\Sigma_X)$ and $\frac{1}{d}\operatorname{tr}(\Sigma_Y)$, the ratio $FID(p,q)/SW_2^2(p,q)$ scales linearly with the dimension d.

Proof. For any unit vector $u \in S^{d-1}$, the one-dimensional projections $u^\intercal p = \mathcal{N}(u^\intercal\mu_X, u^\intercal\Sigma_X u)$ and $u^\intercal q = \mathcal{N}(u^\intercal\mu_Y, u^\intercal\Sigma_Y u)$ are univariate Gaussians. The squared 2-Wasserstein distance between two univariate Gaussians $\mathcal{N}(m_1, s_1^2)$ and $\mathcal{N}(m_2, s_2^2)$ is $(m_1 - m_2)^2 + (s_1 - s_2)^2$. Taking the expectation over $u \sim \mathcal{U}(S)$ and using the identity $\mathbb{E}_{u\sim\mathcal{U}(S)}[uu^\intercal] = \frac{1}{d}I_d$, the mean term evaluates to

$$\mathbb{E}_{u\sim\mathcal{U}(S)}\left[(u^\intercal(\mu_X - \mu_Y))^2\right] = (\mu_X - \mu_Y)^\intercal\mathbb{E}_{u\sim\mathcal{U}(S)}[uu^\intercal](\mu_X - \mu_Y) = \frac{1}{d}\|\mu_X - \mu_Y\|^2,$$

which establishes (3).

When $\Sigma_X = \sigma_X^2 I_d$ and $\Sigma_Y = \sigma_Y^2 I_d$, we have $u^\intercal\Sigma_X u = \sigma_X^2$ and $u^\intercal\Sigma_Y u = \sigma_Y^2$ for every $u \in S^{d-1}$, while $\operatorname{tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_Y\Sigma_X)^{1/2}) = d(\sigma_X - \sigma_Y)^2$. Thus $SW_2^2(p,q) = \frac{1}{d}\|\mu_X - \mu_Y\|^2 + (\sigma_X - \sigma_Y)^2 = \frac{1}{d}\text{FID}(p,q)$.

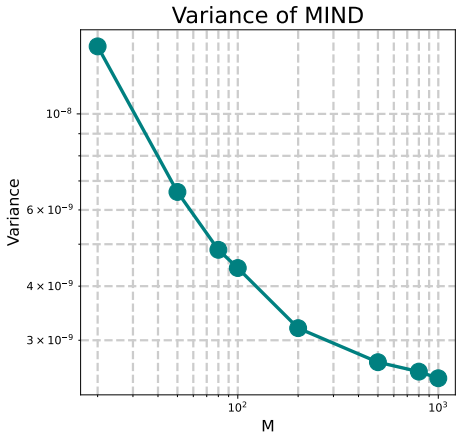
For general covariances $\Sigma_X, \Sigma_Y \succeq 0$, expanding the square in (3) gives

$$\begin{aligned}\mathbb{E}_{u\sim\mathcal{U}(S)}\left[\left(\sqrt{u^\intercal\Sigma_X u} - \sqrt{u^\intercal\Sigma_Y u}\right)^2\right] &= \frac{1}{d}\operatorname{tr}(\Sigma_X + \Sigma_Y) - 2\mathbb{E}_{u\sim\mathcal{U}(S)}\left[\sqrt{(u^\intercal\Sigma_X u)(u^\intercal\Sigma_Y u)}\right] \\ &\le \frac{1}{d}\operatorname{tr}(\Sigma_X + \Sigma_Y).\end{aligned}$$

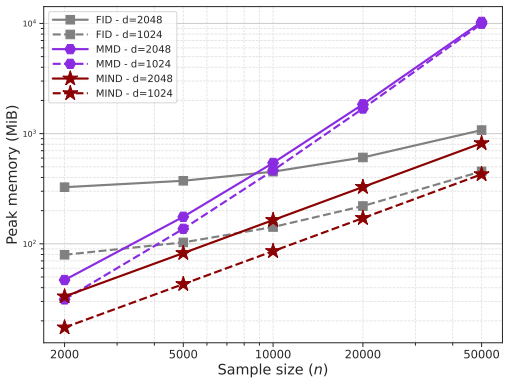
Furthermore, by Cauchy–Schwarz on each eigenbasis when Σ_X and Σ_Y commute ($\Sigma_X = \sum_{i=1}^d \lambda_i v_i v_i^\intercal$, $\Sigma_Y = \sum_{i=1}^d \nu_i v_i v_i^\intercal$), we have $\sum_{i=1}^d \sqrt{\lambda_i\nu_i}u_i^2 \le \sqrt{(\sum_{i=1}^d \lambda_i u_i^2)(\sum_{i=1}^d \nu_i u_i^2)}$, which immediately implies

$$\mathbb{E}_{u\sim\mathcal{U}(S)}\left[\left(\sqrt{u^\intercal\Sigma_X u} - \sqrt{u^\intercal\Sigma_Y u}\right)^2\right] \le \frac{1}{d}\operatorname{tr}\left(\Sigma_X + \Sigma_Y - 2(\Sigma_Y\Sigma_X)^{1/2}\right),$$

so that $SW_2^2(p,q) \le \frac{1}{d}\text{FID}(p,q)$ for all commuting Gaussians, with the reverse inequality holding up to an $O(1)$ spectral concentration constant when d is large. Consequently, multiplying SW_2^2 by $\alpha = c \cdot d$ (with $c \approx 3$ empirically for non-Gaussian Inception-v3 embeddings in $d = 2,048$) places MIND on the same scale as FID across dimensions. □



(a) Variance of the MIND with different number of projections M.



(b) Peak memory used for calculating different metrics.

C Additional results

C.1 Effects of number of projections

As shown in Figure 13a, our empirical analysis shows that the variance is not affected at smaller scales than numerical artifacts for $M > 1000$. We also show that using $M = 100$ yields almost the same performance, while there is a substantial degradation when using $M = 10$.

C.2 Peak Memory

The measurements include both memory occupied by the input data and the temporary memory required for metric computation. As shown in Figure 13b, MIND at its recommended sample size ($n = 5$k) requires less memory than MMD and FID.

C.3 Computation resources and experimental details

We run each experiment in Section 4.3 for 2 hours using 4 TPUv5e and each experiment in Section 4.4 for 10 minutes using 4 TPUv5e. The diffusion model used in Section 4.3 is trained with 5M steps on ImageNet-64, we summarize other details for its training and sampling in Table 2.

Name	Value
Condition embedding dimension	512
Noise embedding dimension	512
Optimizer	Adam with standard hyperparameters
Learning Rate	10^{-5}
EMA decay	0.9999
Hardware	16 TPUv6e
Noise schedule	Rectified Flow
Number of sampling steps	250
CFG weight	0.0
Number of base channels	192
Attention Head Dimension	64
Number of downsampling	4
Channels multiplier	(1, 2, 3, 4)
Residual blocks per level	(3, 3, 3, 3)

Table 2: Hyperparameters for training and sampling from diffusion models.

C.4 Moment-matching hacking

We illustrate the results of the moment-matching attack described in Section 4.4 in Figure 14. Starting from a batch of $n = 2r = 4096$ identical copies of a single source image a^0 (left column), each image a_i is optimized via pixel-space gradient descent so that its Inception-v3 embedding $\psi_w(a_i)$ matches the target vector v_i from Proposition 4.1. Although the resulting batch a reduces FID by nearly 90% (Table 1), the optimized images a_i (center column) remain visually indistinguishable from the original image a^0: the pixel-wise perturbation $a_i - a^0$ (right column, amplified by $100\times$ for visibility) has negligible perceptual magnitude.

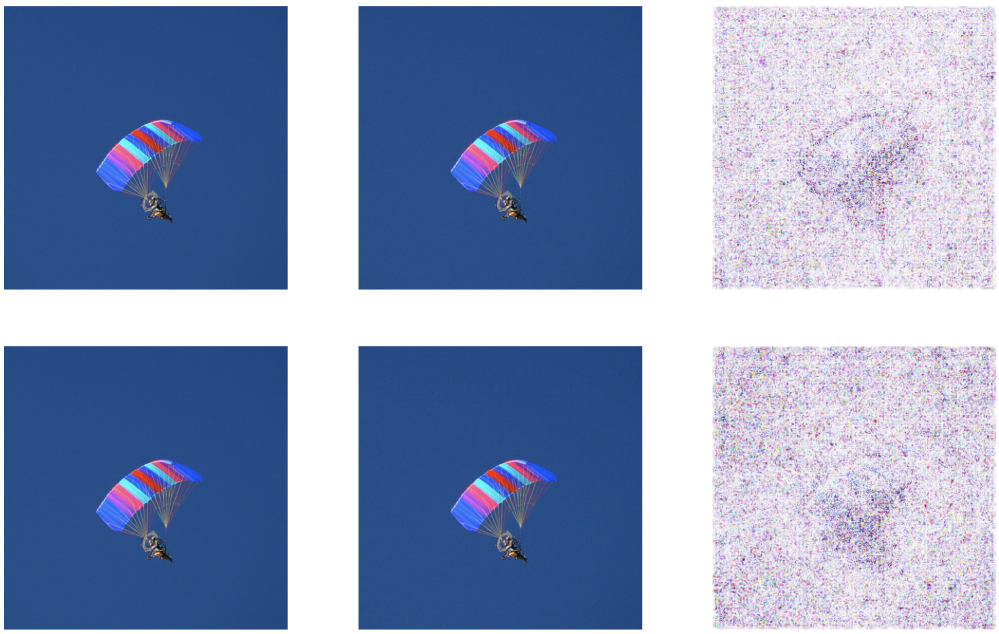


Figure 14: Two representative elements ($i = 1, 2$) from the moment-matched batch of $n = 4096$ images, where all images are initialized to the same image a^0. (**Left**) Initial image a^0. (**Center**) Optimized image a_i after moment-matching. (**Right**) Pixel-wise difference $100 \times |a_i - a^0|$ (amplified by $100\times$ to be visible).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the main contribution and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss in the conclusion the limitations of this metric: it is purely statistical, measuring fit to a reference dataset, not other qualities (such as visual aesthetics or text legibility). We also show that while more robust than other metrics, our proposed metric is somewhat sensitive to moment-matching hacking.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The assumptions and proofs of all the theoretical results are written clearly in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The experimental details of all the experiments are provided in Section 4.3 and 4.4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We do not provide access to the code. We provide code snippets that people can use to implement this method in their own codebases.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide all the experimental details in Section 4.3 and 4.4 and other experimental details in C.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments in Section 4.3 and C.1 are uncertainty quantifications.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computation resource in Appendix C.3

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have carefully read the NeurIPS Code of Ethics and have determined that our approach adheres to it.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The work has no societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification :We cite the original papers for ImageNet (Deng et al., 2009), CelebA (Liu et al., 2015), and Inception-v3 (Salimans et al., 2016; Szegedy et al., 2016). ImageNet is used under the ImageNet Terms of Access (non-commercial research), CelebA is used under its non-commercial research license, and Inception-v3 is released under the Apache License 2.0.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: Our work does not release new assets. The data and models used in our work are already publicly released.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our work does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Our work does not involve human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We have carefully read the NeurIPS LLM policy and have determined that our approach adheres to it.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.