Understanding Generalization in Diffusion Distillation via Probability Flow Distance

Huijie Zhang [1] **Zijian Huang** [1] **Siyi Chen** [1] **Jinfan Zhou** [2] **Zekai Zhang** [1] **Peng Wang** [3] **Qing Qu** [1]

Abstract

Diffusion distillation provides an effective approach for learning lightweight and few-steps diffusion models with efficient generation. However, evaluating their generalization remains challenging: theoretical metrics are often impractical for high-dimensional data, while no practical metrics rigorously measure generalization. In this work, we bridge this gap by introducing probability flow distance (PFD), a theoretically grounded and computationally efficient metric to measure generalization. Specifically, PFD quantifies the distance between distributions by comparing their noise-to-data mappings induced by the probability flow ODE. Using PFD under the diffusion distillation setting, we empirically uncover several key generalization behaviors, including: (1) quantitative scaling behavior from memorization to generalization, (2) epoch-wise double descent training dynamics, and (3) bias-variance decomposition. Beyond these insights, our work lays a foundation for generalization studies in diffusion distillation and bridges them with diffusion training.

1. Introduction

While diffusion models (Song et al., 2021c; Lipman et al., 2023; Liu et al., 2023; Ho et al., 2020) have revolutionized generative AI, their real-world deployment remains limited by substantial computational cost. Knowledge distillation (Hinton et al., 2015), which transfers capabilities from a high-capacity teacher to a significantly smaller student, has therefore become a standard approach for practical diffusion-model deployment. Distillation has shown impressive gains in efficiency through both lightweight student architectures (Chen et al., 2025; Kim et al., 2024; Hu et al., 2026) and shorter sampling trajectories (Yin et al., 2024b;a; Salimans & Ho, 2022; Song et al., 2023; Meng et al., 2023; Sauer et al., 2024; Xie et al., 2024). Beyond efficiency, recent study shows that distillation can also mitigate memorization (Borkar et al., 2026), a known issue in diffusion model training that raises copyright and privacy concerns (Gu et al., 2025; Carlini et al., 2023; Somepalli et al., 2023a;b). In some cases, diffusion distillation even leads to surprising improvements in generation quality (Ma et al., 2025).

However, despite the remarkable performance of distilled models, a comprehensive understanding of their generalization ability and working mechanisms remains elusive. Specifically, existing metrics for evaluating the generalizability of diffusion distillation suffer from significant shortcomings: common empirical metrics like Fréchet inception distance (FID) (Heusel et al., 2017), Inception Score (IS) (Salimans et al., 2016) focus on generation quality, but they cannot distinguish between memorization and generalization, where both can yield high-quality outputs. Neural Network Divergence (NND) (Arora et al., 2017; Gulrajani et al., 2019) proposed to measure the generalizability for generative adversarial networks (GANs) (**?**). However, it requires a large amount of data for evaluation and is not suitable for diffusion distillation. Although recent works measure generalization by evaluating the likelihood of generated samples that are copied from the training data (Zhang et al., 2024; Yoon et al., 2023), this can be misleading, as pure noise may be misclassified as generalized output.

On the other hand, other approaches aim to measure generalization by comparing the distance between the student distribution and the teacher distribution. While distributional distance such as Kullback-Leibler divergence (KL) (Chen et al., 2023e; Nie et al., 2024; Li et al., 2023), total variation (TV) (Chen et al., 2023c; Li et al., 2024b;a; Yang et al., 2024), and 2-Wasserstein distance (W_2) (Gao & Zhu, 2025; Bortoli, 2022; Chen et al., 2023a; Gao et al., 2025b) are theoretically appealing, they are often computationally expensive and thus impractical for diffusion distillation. (We defer a detailed discussion in Section 2.4.) As summarized in Table 1, existing metrics are not both accurate and efficient for evaluating diffusion distillation in practice, highlighting the need for a generalization metric that is both theoretically grounded and practically tractable.

Moreover, developing a principled evaluation framework

[1]University of Michigan, Ann Arbor, USA [2]University of Chicago, Chicago, USA [3]University of Macau, Macau, China. Correspondence to: Huijie Zhang <huijiezh@umich.edu>.

within the distillation setting is imperative. Such a framework is crucial not only for deepening our insight into the mechanisms underlying diffusion distillation but also for providing systematic guidance in designing more effective architectures, training strategies, and practical benchmarks.

Our Contribution. In this work, we propose a novel metric, termed the probability flow distance (PFD), that can faithfully evaluate the generalization ability of distilled diffusion models. Specifically, our PFD metric quantifies distributional differences by leveraging the backward probability flow ODE (PF-ODE) (Song et al., 2021c), which is widely used in the sampling process of diffusion models. Unlike practical metrics such as FID, which cannot distinguish between memorization and generalization, our PFD provides a theoretically grounded measure of distance between distributions, offering a more reliable assessment of generalization. Compared to theoretical metrics like the Wasserstein distance, PFD is computationally more efficient by leveraging the benign properties of PF-ODE. Moreover, by leveraging PFD, our analysis reveals several intriguing generalization phenomena that offer new insights into the learning behavior of diffusion distillation, which could potentially be informative for training diffusion models in general.

- **Scaling behavior from memorization to generalization.** Our metric quantitatively characterizes the scaling behavior of diffusion distillation in the transition from memorization to generalization. Specifically, we demonstrate that generalization follows a universal scaling behavior governed by $N/\sqrt{|\boldsymbol{\theta}|}$, where N is the training dataset size and $|\boldsymbol{\theta}|$ is the number of model parameters. In contrast, prior studies (Zhang et al., 2024; Yoon et al., 2023) have only considered the effects of model capacity or dataset size in isolation, without capturing their joint influence on generalization.

- **Understanding generalization in learning dynamics.** Our PFD metric reveals and reconciles several intriguing generalization phenomena within the learning dynamics. Specifically, under regimes of sufficient training data, we identify an "**epochwise double descent**" behavior in diffusion distillation, where the generalization error initially decreases, subsequently increases, and finally decreases again toward convergence. While similar phenomena have been observed in overparameterized supervised models, we provide the first empirical validation of this behavior in the context of diffusion distillation. Conversely, when training data is limited, we observe early learning, where models initially generalize but later transition to memorization until convergence.

- **The bias-variance trade-off of generalization errors.** Finally, we show that our PFD metric naturally introduces a bias–variance decomposition of the generalization error, extending classical statistical learning theory to diffusion models. Empirically, we observe a trade-off consistent with supervised learning: increasing model capacity reduces bias but increases variance, yielding a characteristic U-shaped generalization error curve. This finding could potentially help us identify overfitting and memorization in the distillation and training of diffusion models.

2. Introduction of Probability Flow Distance

In this section, we propose a new metric called probability flow distance (PFD), which is designed to quantify the distance between two arbitrary probability distributions. The design of PFD is motivated by the PF-ODE, which we first review in Section 2.1. We then formally define PFD in Section 2.2 and present its empirical estimation with theoretical guarantees in Section 2.3.

2.1. The Mapping Induced by PF-ODE

In general, PF-ODE (Song et al., 2021c) is a class of ordinary differential equations (ODE) that aims to reverse a forward process, where Gaussian noise is progressively added to samples drawn from an underlying distribution, denoted as $p_{\texttt{data}}$. The forward process and the PF-ODE can be described as follows:

- *Forward process.* Given a sample $\boldsymbol{x}_0 \overset{i.i.d.}{\sim} p_{\texttt{data}}(\boldsymbol{x})$, the forward process progressively corrupts it by adding Gaussian noise. This process can be characterized by the stochastic differential equation (SDE) $\mathrm{d}\boldsymbol{x}_t = f(t)\boldsymbol{x}_t\mathrm{d}t + g(t)\mathrm{d}\boldsymbol{w}_t$, , where $t \in [0, T]$ is the time index, $\{\boldsymbol{w}_t\}_{t\in[0,T]}$ is a standard Wiener process, and $f(t), g(t) : \mathbb{R}_+ \to \mathbb{R}$ are the drift and diffusion functions that control the noise schedule. In this work, we adopt the noise schedule proposed by elucidated diffusion models (EDM) (Karras et al., 2022a), where $f(t) = 0$ and $g(t) = \sqrt{2t}$. Substituting this into the SDE and integrating both sides, we obtain $\boldsymbol{x}_t = \boldsymbol{x}_0 + \int_0^t \sqrt{2\tau}\mathrm{d}\boldsymbol{w}_\tau$. For ease of exposition, we use $p_t(\boldsymbol{x}_t)$ to denote the distribution of the noisy image $\boldsymbol{x}_t$ for each $t \in [0, T]$. In particular, it is worth noting that $p_0(\boldsymbol{x}) = p_{\texttt{data}}(\boldsymbol{x})$ and $p_T(\boldsymbol{x}) \to \mathcal{N}(\mathbf{0}, T^2\boldsymbol{I})$ as $T \to +\infty$.
- *Probability flow ODE.* According to (Song et al., 2021c), the PF-ODE can transform a noise sample $\boldsymbol{x}_T$ back into a clean data sample $\boldsymbol{x}_0$. Specifically, under the EDM noise scheduler, the PF-ODE admits the following form:

$$\mathrm{d}\boldsymbol{x}_t = -t\nabla \log p_t(\boldsymbol{x}_t)\mathrm{d}t, \tag{1}$$

where $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)$ (or simply $\nabla \log p_t(\boldsymbol{x}_t)$) denotes the *score function* of the distribution $p_t(\boldsymbol{x}_t)$ at time $t \in [0, T]$. According to (Song et al., 2021c), the backward PF-ODE (1) and the forward SDE have the same distribution at each timestep t. In practice, since the score function $\log p_t(\boldsymbol{x}_t)$ is unknown, in diffusion models we approximate it using

a neural network $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ and employ a numerical solver to generate samples from Equation (1). Additional details are provided in Section A.4.

The backward PF-ODE induces a *unique* mapping $\boldsymbol{\Phi}_{p_{\texttt{data}}}$ from $\boldsymbol{x}_T$ to $\boldsymbol{x}_0$. Integrating both sides of (1) from T to 0, the mapping $\boldsymbol{\Phi}_{p_{\texttt{data}}}$ can be expressed as:

$$\boldsymbol{\Phi}_{p_{\texttt{data}}}(\boldsymbol{x}_T) \coloneqq \boldsymbol{x}_T - \int_T^0 t \nabla \log p_t(\boldsymbol{x}_t) \mathrm{d}t. \tag{2}$$

Previous work (Song et al., 2021c) shows that, when $\boldsymbol{x}_T \sim \mathcal{N}(\mathbf{0}, T^2\boldsymbol{I})$ and $T \to +\infty$, the random variable $\boldsymbol{\Phi}_{p_{\texttt{data}}}(\boldsymbol{x}_T)$ is distributed according to $p_{\texttt{data}}(\boldsymbol{x})$. Consequently, if the data distribution $p_{\texttt{data}}$ is known, the score function $\nabla \log p_t(\boldsymbol{x}_t)$ is explicitly available, and the backward PF-ODE induces a deterministic and unique mapping from the Gaussian distribution to $p_{\texttt{data}}$ (we formally prove uniqueness in the identity property of our Theorem 1).

2.2. Definition of Probability Flow Distance

Based on the above setup, we define a metric to measure the distance between any two distributions as follows.

Definition 1 (Probability flow distance (PFD))**.** *For any two given distributions p and q of the same dimension, we define their distribution distance as*

$$\texttt{PFD}\left(p, q\right) \coloneqq \left(\mathbb{E}_{\boldsymbol{x}_T}\left[\left\|\boldsymbol{\Psi} \circ \boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Psi} \circ \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T\right)\right\|_2^2\right]\right)^{1/2}. \tag{3}$$

Here, $\boldsymbol{x}_T \sim \mathcal{N}(0, T^2\boldsymbol{I})$, $\boldsymbol{\Phi}_p$ and $\boldsymbol{\Phi}_q$ denote the mappings between the noise and image spaces for distributions p and q, respectively, as defined in (2), and $\boldsymbol{\Psi}(\cdot)$ represents an image descriptor.

Intuitively, when the image descriptor $\boldsymbol{\Psi}(\cdot)$ is an *identity mapping*, PFD measures the distance between two distributions p and q by comparing their respective noise-to-image mappings $\boldsymbol{\Phi}_p(\cdot)$ and $\boldsymbol{\Phi}_q(\cdot)$ starting from the same Gaussian noise input $\boldsymbol{x}_T$. Small values of PFD indicate that the two distributions generate similar data from identical noise and, therefore, are close to each other.

However, in practice, measuring distributional distance alone does *not* fully capture the perceptual quality of natural images. To incorporate perceptual quality, we employ an image descriptor $\boldsymbol{\Psi}(\cdot)$, which is typically implemented using a pre-trained neural network such as DINOv2 (Oquab et al., 2024) or Self-Supervised Copy Detection Descriptor (SSCD) (Pizzi et al., 2022). Measuring distances in such feature spaces is a common practice in existing metrics for generative models (Heusel et al., 2017; Salimans et al., 2016; Stein et al., 2024), as it tends to align better with human perception (Stein et al., 2024). As shown by our ablation studies in Section E.2, the PFD metric with an image descriptor effectively captures differences in perceptual quality while models are generalizing, where a smaller quantity indicates better perceptual quality. Therefore, for all experiments in Sections 3 and 4, we use an image descriptor (e.g., SSCD) for $\boldsymbol{\Psi}(\cdot)$.

Nonetheless, for simplicity and analytical tractability, we assume the image descriptor $\boldsymbol{\Psi}(\cdot)$ to be an identity mapping in the following theoretical analysis. Specifically, under Definition 1, we show that PFD satisfies the axioms of a metric (Definition 2.15 in (Rudin, 2021)).

Theorem 1. *For any two distributions p and q, the* PFD *satisfies the following properties:*

- *(Positivity)* $\texttt{PFD}(p, q) > 0$ *for any* $p \neq q$.
- *(Identity Property)* $\texttt{PFD}(p, q) = 0$ *if and only if* $p = q$.
- *(Symmetry)* $\texttt{PFD}(p, q) = \texttt{PFD}(q, p)$.
- *(Triangle Inequality)* $\texttt{PFD}(p, q) \leq \texttt{PFD}(p, p') + \texttt{PFD}(p', q)$ *for all* p'.

We defer the proof to Section B. Note that Theorem 1 establishes the theoretical validity of PFD as a metric for measuring the distance between any two probability distributions.

2.3. Empirical Estimation of PFD

In practice, the expectation in (3) is intractable due to the complexity of the underlying distributions. Thus, we approximate the PFD using finite samples:

$$\hat{\texttt{PFD}}(p, q) = \left(\frac{1}{M}\sum_{i=1}^{M}\left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T^{(i)}\right) - \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T^{(i)}\right)\right\|_2^2\right)^{1/2}. \tag{4}$$

Here, $\hat{\texttt{PFD}}(p, q)$ is the empirical version of $\texttt{PFD}(p, q)$ computed over M independent samples $\{\boldsymbol{x}_T^{(i)}\}_{i=1}^M \overset{i.i.d.}{\sim} \mathcal{N}(0, T^2\boldsymbol{I})$ with $T \to \infty$.

Specifically, our finite-sample approximation relies on two key assumptions: (i) the score functions are smooth at all timesteps, and (ii) the score functions of two distributions remain uniformly close within a bounded region of the input space, which can be described as follows.

Assumption 1. *Let p and q be two distributions with the same dimension, where we assume:*

(i) (Lipschitz score functions) There exists a constant $L > 0$ such that for all $\boldsymbol{x}_1, \boldsymbol{x}_2$ and $t \in [0, T]$, it holds that

$$\left\|\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_1) - \nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_2)\right\|_2 \leq L\left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|_2, \tag{5}$$

and similarly for q_t.

(ii) (Uniform Closeness) For all $t \in [0, T]$, there exists a constant $\epsilon > 0$ such that

$$\left\|\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x})\right\|_2 \leq \epsilon. \tag{6}$$

The Lipschitz continuity of the score function is a common assumption widely adopted in the theoretical analysis of score functions in diffusion models (Block et al., 2020; Lee et al., 2022; Chen et al., 2023d;a; Zhu et al., 2023; Chen et al., 2023b). More recently, this property has been rigorously established under the assumption that the data distribution is a mixture of Gaussians (Liang et al., 2024). The uniform closeness assumption holds when p and q both follow Assumption 1 (*i*) and have support on compact domains, which is often the case for image distributions. Under Assumption 1, the concentration of the empirical estimate $\hat{\texttt{PFD}}(p, q)$ to $\texttt{PFD}(p, q)$ can be characterized as follows.

Theorem 2. *Let p and q be two distributions satisfying Assumption 1, and let $\hat{\texttt{PFD}}(p, q)$ denote the empirical estimate of $\texttt{PFD}(p, q)$ using M independent samples, as defined in (4). Then, for any $\gamma > 0$, this empirical estimate satisfies the following bound:*

$$\left|\hat{\texttt{PFD}}(p, q) - \texttt{PFD}(p, q)\right| \leq \gamma \text{ whenever } M \geq \frac{\kappa^4(L, \epsilon)}{2\gamma^4} \log \frac{2}{\eta}, \quad (7)$$

with probability at least $1 - \eta$. Here, $\kappa(L, \epsilon) := \exp\left(\frac{LT_\xi^2}{2}\right)\xi + \frac{\epsilon}{L}\left(\exp\left(\frac{LT_\xi^2}{2}\right) - 1\right)$ is a constant, with a numerical constant $\xi > 0$ and a finite timestep T_ξ depending only on ξ.

We defer the proof to Section B. Given the score functions of both distributions are smooth and uniformly close, our result in Theorem 2 guarantees that $\texttt{PFD}(p, q)$ can be approximated to arbitrary precision by its empirical estimate $\hat{\texttt{PFD}}(p, q)$ with high probability, given a finite number of samples.

2.4. Advantages of PFD over Existing Metrics.

As summarized in Table 1, we conclude this section by highlighting the advantages of the proposed PFD over commonly used theoretical metrics for measuring distributional distances, including both density-based and sample-based methods. A comparison with practical evaluation metrics is presented at the end of Section 3.

Sampling efficiency. We compare the sampling efficiency of PFD, FID, and W_2 on Gaussian distributions, as shown in Figure 6a, using the experimental settings described in Section C.1. With the same number of estimated samples ($M = 4096$), PFD attains a relative error of approximately 4×10^{-3}, whereas FID and W_2 achieve only about 2×10^{-2}. Consequently, to reach the same level of relative error, PFD requires substantially fewer samples.

Computational efficiency. Even with the same number of estimated samples M, PFD requires less computation time than other methods:

- **Comparison with sample-based distances.** Sample-based distances such as the Wasserstein distance and Maximum Mean Discrepancy (MMD) incur $O(M^2)$ computational complexity. In contrast, PFD requires only $O(M)$ computation, making it substantially more efficient.
- **Comparison with density-based distances.** Density-based distances, such as KL divergence, total variation distance, and Jensen–Shannon divergence, require approximating probability densities through computationally intensive methods like the Skilling-Hutchinson trace estimator (Song et al., 2021c; Skilling, 1989; Hutchinson, 1989), which become prohibitively expensive in high-dimensional settings for evaluating diffusion models. In contrast, PFD directly estimates the distributional distance using the score function inherently learned by diffusion models. Furthermore, density-based metrics are theoretically ill-suited for image data, as probability densities remain undefined outside the image manifold (Loaiza-Ganem et al., 2024).

Table 1. **Metrics comparison.**

	Efficient	Accurate
Theoretical distances		
Density-based (KL, TV)	✗	✓
Sample-based (W_2, MMD)	✗	✓
Practical metrics		
FID, IS, NND, *etc.*	✓	✗
PFD (**Ours**)	✓	✓

3. Quantifying Generalization Errors via PFD

In this section, we leverage the PFD metric in Section 2 to rigorously define and evaluate the generalization error of diffusion models under distillation settings. Specifically, this metric enables us to distinguish between memorization and generalization behaviors, as well as analyze the transition from memorization to generation (MtoG).

This MtoG transition has been explored in recent studies for training diffusion models (Yoon et al., 2023; Zhang et al., 2024; Kadkhodaie et al., 2024; Bonnaire et al., 2025), which highlight two learning regimes of diffusion models depending on dataset size and model capacity: (*i*) **Memorization regime:** Large models trained on small datasets memorize the empirical distribution $p_{\texttt{emp}}(\boldsymbol{x})$ of the training data, yielding poor generalization and no novel samples. (*ii*) **Generalization regime:** For fixed model capacity, as the number of training samples increases, the model transitions into generalization, approximating the true data distribution $p_{\texttt{data}}(\boldsymbol{x})$ and generating new samples.

However, existing approaches (Yoon et al., 2023; Zhang et al., 2024; Alaa et al., 2022) quantify generalization solely by measuring the distance from a generated sample to its nearest neighbor in the training dataset. While effective for

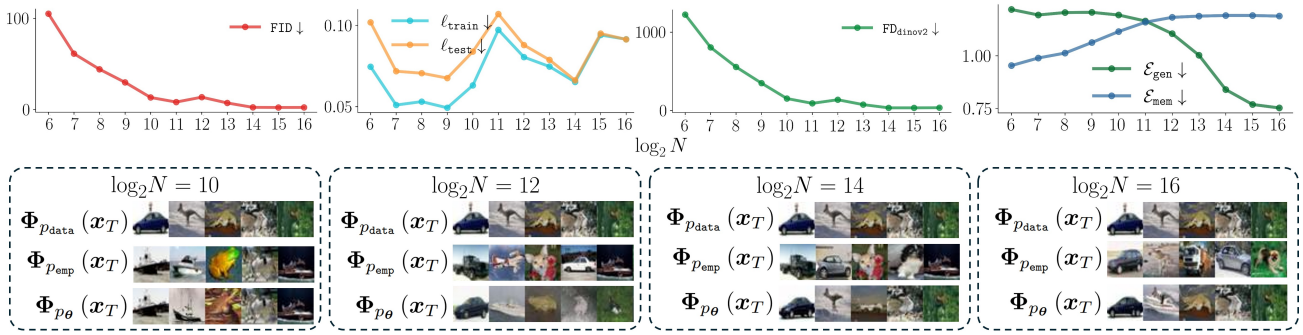


Figure 1. **Comparison of practical metrics on the MtoG transition.** The top figure plots multiple evaluation metrics as functions of $\log_2 N$. The bottom figure visualizes the generation when $N = 2^{10}, 2^{12}, 2^{14}, 2^{16}$, sampled from the $p_{\texttt{data}}$ (top row), the $p_{\texttt{emp}}$ (middle row), and $p_{\boldsymbol{\theta}}$ (bottom row). The same column shared the same initial noise $\boldsymbol{x}_T$.

identifying memorization, these sample-level measures are inadequate for capturing true generalization because they do not quantify distributional distance. Simply showing that an output is distinct from training data is insufficient; for instance, such metrics could erroneously classify a sample of pure noise as a valid generated sample. To address these limitations, we leverage the PFD metric to assess generalization at the distributional level. Specifically, we quantify how closely the distribution learned via diffusion models, $p_{\boldsymbol{\theta}}$, approximates the underlying distribution $p_{\texttt{data}}(\boldsymbol{x})$ and how closely it aligns with the empirical distribution $p_{\texttt{emp}}(\boldsymbol{x})$. Based on this approach, we formally define generalization and memorization errors as follows.

Definition 2 (Generalization and Memorization Errors)**.** *Consider a diffusion model $\boldsymbol{s}_{\boldsymbol{\theta}}$ trained on a finite dataset $\mathcal{D} = \{\boldsymbol{y}^{(i)}\}_{i=1}^N$, where each sample $\boldsymbol{y}^{(i)}$ is drawn i.i.d. from the underlying distribution $p_{\texttt{data}}(\boldsymbol{x})$. Denote the learned distribution induced by a diffusion model $\boldsymbol{s}_{\boldsymbol{\theta}}$ as $p_{\boldsymbol{\theta}}(\boldsymbol{x})$. Using the* PFD *metric, we can formally define the generalization and memorization errors as follows:*

$$\mathcal{E}_{\texttt{gen}}(\boldsymbol{\theta}) \coloneqq \texttt{PFD}\left(p_{\boldsymbol{\theta}}, p_{\texttt{data}}\right), \quad \mathcal{E}_{\texttt{mem}}(\boldsymbol{\theta}) \coloneqq \texttt{PFD}\left(p_{\boldsymbol{\theta}}, p_{\texttt{emp}}\right), \quad (8)$$

where the empirical distribution $p_{\texttt{emp}}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^N \delta(\boldsymbol{x} - \boldsymbol{y}^{(i)})$, with $\delta(\cdot)$ denoting the Dirac delta function.

Here, given access to $p_{\texttt{emp}}(\boldsymbol{x})$, the memorization error $\mathcal{E}_{\texttt{mem}}(\boldsymbol{\theta})$ can be exactly computed (see Section D). We further show that $\mathcal{E}_{\texttt{mem}}(\boldsymbol{\theta})$ coincides with metrics introduced in (Yoon et al., 2023; Zhang et al., 2024).

Evaluation protocol for generalization. Under the diffusion distillation setting (see Section 4), we study generalization behaviors of a student model distilled from a teacher model. Specifically, a pretrained diffusion model $\boldsymbol{s}_{\boldsymbol{\theta}_{\texttt{t}}}(\boldsymbol{x}_t)$ with parameters $\boldsymbol{\theta}_{\texttt{t}}$ serves as the teacher and induces a distribution $p_{\boldsymbol{\theta}_{\texttt{t}}}$, which we treat as the underlying data distribution, i.e., $p_{\texttt{data}} = p_{\boldsymbol{\theta}_{\texttt{t}}}$. We employ the most straightforward distillation strategy by directly training a student model $\boldsymbol{s}_{\boldsymbol{\theta}}$ on samples drawn from $p_{\boldsymbol{\theta}_{\texttt{t}}}$. Generalization behaviors of the student model is then evaluated by comparing the student-induced distribution $p_{\boldsymbol{\theta}}$ with $p_{\boldsymbol{\theta}_{\texttt{t}}}$ using the generalization error metrics defined in Definition 2.

In our experiments for the rest of the paper, both the teacher and student models adopt the U-Net architecture (Ronneberger et al., 2015). The teacher model $\boldsymbol{s}_{\boldsymbol{\theta}_{\texttt{t}}}$ is trained on the CIFAR-10 dataset (Krizhevsky et al., 2009) with a fixed model architecture (UNet-10 introduced in Section C.2). The student model $\boldsymbol{s}_{\boldsymbol{\theta}}$ is trained on samples generated by the teacher, with the number of distillation samples varying from $N = 2^6$ to $N = 2^{16}$, using the same training hyperparameters but different model sizes. For evaluating the generalization error in (8), we compute the PFD between the teacher and student models using $M = 10^4$ samples drawn from shared initial noise, as defined in (4). Similarly, for the memorization error, we compute the PFD between the student model and the empirical distribution of the training data. Additional details for the evaluation protocol and ablation studies are provided in Section C.3 and Section E.

Comparison with practical metrics for evaluating generalization. Before we use the proposed metrics $\mathcal{E}_{\texttt{gen}}$ and $\mathcal{E}_{\texttt{mem}}$ for revealing the generalization properties of diffusion models in Section 4, we conclude this section by demonstrating their advantages over commonly used practical metrics, such as FID, $\text{FD}_{\texttt{DINOv2}}$ (Stein et al., 2024), training and testing loss $\ell_{\texttt{train}}, \ell_{\texttt{test}}$ (see Equation (10)). We defer a more comprehensive comparison with other metrics such as IS, NND, KID (Bińkowski et al., 2018), CMMD (Jayasumana et al., 2024), Precision, and Recall (Kynkäänniemi et al., 2019) to Section C.4.

As shown in Figure 1, we compare several metrics for capturing the MtoG transition under distillation. Among them, only the proposed metric $\mathcal{E}_{\texttt{gen}}$ consistently tracks this transition as the number of distillation samples increases. This conclusion is supported by the qualitative results in the bottom row of Figure 1: when $N \geq 2^{10}$, increasing N makes the distilled model's generations (bottom row) visually closer to the teacher distribution (top row), indicating

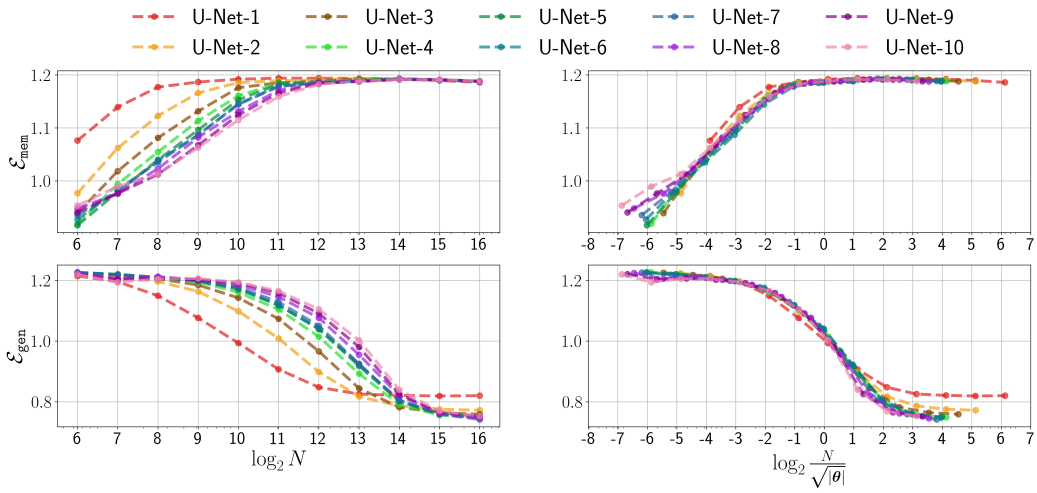


Figure 2. **Scaling behavior in the MtoG transition.** $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ plotted against Left: $\log_2(N)$ for a range of U-Net architectures (U-Net-1 to U-Net-10). Right: the same metrics plotted against $\log_2(N/\sqrt{|\boldsymbol{\theta}|})$, where $|\boldsymbol{\theta}|$ is the number of model parameters.

improved generalization. Among all metrics, only PFD captures this decreasing trend.

From a theoretical perspective, FID, $\texttt{FD}_{\texttt{DINOv2}}$ relies on a Gaussian assumption for feature distributions, while $\ell_{\texttt{train}}$ and $\ell_{\texttt{test}}$ provide only upper bounds on the negative log-likelihood of the learned distribution $p_{\boldsymbol{\theta}}$ (Song et al., 2021b); as a result, none of these metrics can accurately capture generalization. In contrast, PFD is both theoretically well-founded and empirically validated as a reliable metric for measuring generalization.

4. Findings of Key Generalization Behaviors

Based on the evaluation protocol in Section 3, this section reveals several key generalization behaviors in diffusion distillation: (i) MtoG scaling behaviors with model capacity and distillation dataset size (Section 4.1), (ii) double descent in learning dynamics (Section 4.2), and (iii) the bias-variance trade-off of generalization error (Section 4.3).

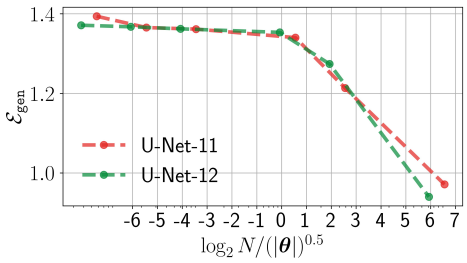


Figure 3. **Scaling behavior over ImageNet dataset.** U-Net-11, U-Net-12 contains 124.2M and 295.9M parameters. Distilling data size N ranging from 2^6 to 2^{20}.

4.1. Scaling Behaviors of Generalization

We investigate the scaling behavior of the distilled models with respect to both model capacity $|\boldsymbol{\theta}|$ and distillation data size N, using the metrics $\mathcal{E}_{\texttt{gen}}$ and $\mathcal{E}_{\texttt{mem}}$. We evaluate ten U-Net architectures over CIFAR-10 dataset, with model sizes ranging from 0.9M to 55.7M parameters (U-Net-1 to U-Net-10). For each model, we compute $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ across varying distillation dataset sizes, following the distillation setting outlined in Section 3. In addition, we conduct experiments on the ImageNet dataset (Deng et al., 2009). Results for CIFAR-10 and ImageNet are reported in Figure 2 and Figure 3, respectively. Additional experimental details are provided in Section C.5. From these results, we observe the following:

MtoG transitions governed by the ratio $N/\sqrt{|\boldsymbol{\theta}|}$. As shown in Figure 2 (left), for a fixed model capacity $|\boldsymbol{\theta}|$, our metrics reveal a clear transition from memorization to generalization as the number of distillation samples N increases, consistent with prior experiments training diffusion models from scratch. (Zhang et al., 2024; Yoon et al., 2023). Moreover, in contrast to prior work that focuses solely on the effect of distillation sample size N, our results in Figure 2 (right) for CIFAR-10 and Figure 3 for ImageNet reveal a consistent quantitative scaling behavior when using our metric, governed by the ratio $N/\sqrt{|\boldsymbol{\theta}|}$ between data size and model capacity. Remarkably, both $\mathcal{E}_{\texttt{gen}}$ and $\mathcal{E}_{\texttt{mem}}$ metrics exhibit near-identical MtoG transition curves across models of varying sizes when plotted against this ratio. As such, analogous to the empirical scaling laws observed in large language models (Kaplan et al., 2020), this predictable trend provides practical guidance for the development of diffusion models, particularly when scaling up model size, data, or compute to achieve optimal performance gains.

4.2. Generalization across Learning Dynamics

Building on the findings in Section 4.1, we further examine the generalization behavior across different training regimes. Under the distillation setting for CIFAR-10 dataset in Section 3, we analyze the learning dynamics of a U-Net model with fixed model capacity (UNet-10 introduced in Section C.2) distilled with $N = 2^6$, 2^{12}, and 2^{16}, corresponding to the memorization, transition, and generalization regimes in Section 4.1, respectively. The model is distilled using stochastic gradient descent (SGD) for 500 epochs, during which we track $\mathcal{E}_{\texttt{mem}}$, $\mathcal{E}_{\texttt{gen}}$, $\ell_{\texttt{train}}$, and $\ell_{\texttt{test}}$ at each epoch. The results in Figure 4 reveal several notable generalization behaviors that align with phenomena previously observed in the training of overparameterized deep models (Zhang et al., 2017; Nakkiran et al., 2021):

Epochwise double descent of the generalization error in the generalization regime. In contrast, as shown in Figure 4 (right), distilling in the generalization regime ($N = 2^{16}$) reveals a clear instance of the *double descent* phenomenon (Nakkiran et al., 2021) in the generalization error. Specifically, the error initially decreases, then increases during intermediate distilling epochs, and finally decreases again as distilling approaches convergence. Notably, this non-monotonic behavior is not captured by the standard training and test losses $\ell_{\texttt{train}}$ and $\ell_{\texttt{test}}$, both of which decrease monotonically throughout distilling. This implies that extended distilling can improve generalization performance in the generalization regime.

Epochwise early learning behavior in memorization and transition regimes. As shown in Figure 4 (left & middle), in both the memorization ($N = 2^6$) and transition ($N = 2^{12}$) regimes, the generalization error initially decreases during distilling but reaches its minimum at an early epoch, after which it begins to increase again. This *early learning* (or early generalization) phenomenon becomes more salient as the distillation sample size increases from the memorization to the transition regime. As shown in the visualization at the bottom of Section 3, the model at Epoch 85 clearly exhibits generalization, whereas the model at Epoch 500 fails to generalize. This is also corroborated by the divergence of training loss $\ell_{\texttt{train}}$ and test loss $\ell_{\texttt{test}}$ at the top of the figure. It is worth mentioning that, although early learning behavior has been theoretically and visually demonstrated in previous works (Li et al., 2024d; 2023; Bonnaire et al., 2025), our work develops the first framework for precisely capturing this phenomenon.

4.3. Bias-variance Trade-off of the Generalization Error

In statistical learning theory, bias-variance trade-off is a classical yet fundamental concept in supervised learning which helps us understand and analyze the sources of prediction error in the model (Kohavi et al., 1996; Hastie et al., 2009; Yang et al., 2020; Belkin et al., 2019). Specifically, bias–variance decomposition expresses the expected generalization error as the sum of two components: (i) the *bias term*, which quantifies the discrepancy between the expected model prediction and the true function: high bias indicates systematic error or underfitting; and (ii) the *variance term*, which measures the prediction variability of the model across different training sets: high variance reflects sensitivity to data fluctuations or overfitting.

However, in unsupervised learning settings such as diffusion models, the notion of generalization error was not well-defined prior to our work, in contrast to the well-established definitions in supervised learning. As a result, bias–variance decomposition in this context remains largely unexplored. In this work, we address this gap through the generalization error measure $\mathcal{E}_{\texttt{gen}}$ (see Equation (8)), which admits a bias–variance decomposition analogous to that in the supervised setting, as we detail below.

Definition 3 (Bias-Variance Decomposition of $\mathcal{E}_{\texttt{gen}}$)**.** *Based on the same setup as Definition 2, we can decompose $\mathcal{E}_{\texttt{gen}}$ in Equation (8) as*

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{E}^2_{\texttt{gen}}\left(p_{\boldsymbol{\theta}(\mathcal{D})}\right)\right] = \mathcal{E}^2_{\texttt{bias}} + \mathcal{E}_{\texttt{var}} \tag{9}$$

where $p_{\boldsymbol{\theta}(\mathcal{D})}$ denotes the distribution induced by a diffusion model $\boldsymbol{\theta}(\mathcal{D})$ trained on a given training dataset $\mathcal{D}$ sampled from $p_{\texttt{data}}$. Specifically, the bias and variance terms are defined as:

$$\mathcal{E}_{\texttt{bias}} := \mathbb{E}_{\boldsymbol{x}_T}([\|\boldsymbol{\Psi}\circ\boldsymbol{\Phi}_{p_{\texttt{data}}}(\boldsymbol{x}_T) - \overline{\boldsymbol{\Psi}\circ\boldsymbol{\Phi}}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)\|_2^2])^{1/2},$$
$$\mathcal{E}_{\texttt{var}} := \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\boldsymbol{x}_T}[\|\boldsymbol{\Psi}\circ\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}(\mathcal{D})}}(\boldsymbol{x}_T) - \overline{\boldsymbol{\Psi}\circ\boldsymbol{\Phi}}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)\|_2^2],$$

with $\overline{\boldsymbol{\Psi}\circ\boldsymbol{\Phi}}_{p_{\boldsymbol{\theta}}}(\cdot) := \mathbb{E}_{\mathcal{D}}[\boldsymbol{\Psi}\circ\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}(\mathcal{D})}}(\cdot)]$.

Intuitively, our definitions of the bias term $\mathcal{E}_{\texttt{bias}}$ and the variance term $\mathcal{E}_{\texttt{var}}$ are both well-justified: (i) $\mathcal{E}_{\texttt{bias}}$ quantifies the systematic error between the learned distribution $p_{\boldsymbol{\theta}}$ and the ground-truth distribution $p_{\texttt{data}}$; and (ii) $\mathcal{E}_{\texttt{var}}$ captures the variability of model predictions across different training sets by measuring the distance between $p_{\boldsymbol{\theta}}$ and the mean $\overline{p_{\boldsymbol{\theta}}}$ which can be empirically estimated by averaging over multiple datasets $\mathcal{D}$ sampled from $p_{\texttt{data}}$. Experimental results, following the protocol in Section 3, are shown in Figure 5, with detailed settings in Section C.7.

In Figure 5 (a), when diffusion models are distilled in the generalization regime, the resulting generalization decomposition aligns with classical bias–variance theory from supervised learning: as model complexity increases, the bias term $\mathcal{E}_{\texttt{bias}}$ decreases while the variance term $\mathcal{E}_{\texttt{var}}$ increases, resulting in a U-shaped generalization error curve. Additionally, Figure 5 (b) further illustrates the effect of the distillation sample size N and number of parameters $|\boldsymbol{\theta}|$: increasing N reduces both $\mathcal{E}_{\texttt{bias}}$ and $\mathcal{E}_{\texttt{var}}$, thereby lowering the generalization error $\mathcal{E}_{\texttt{gen}}$, as expected; In contrast,

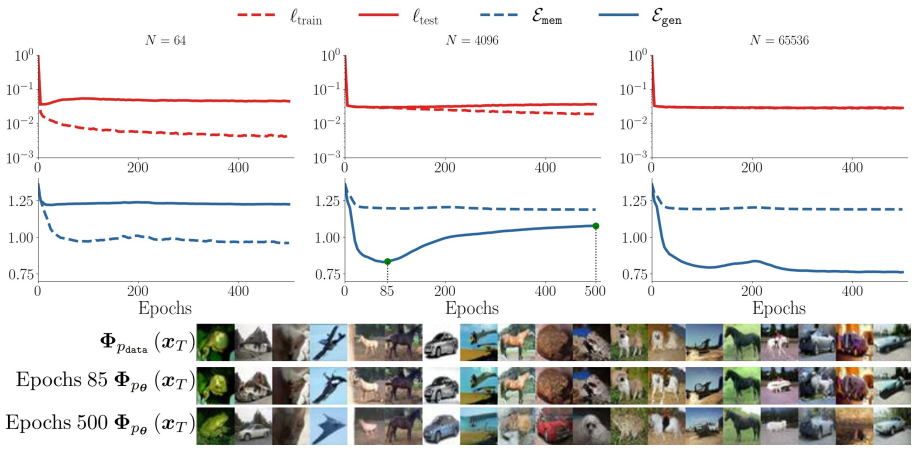


Figure 4. **Training dynamics of diffusion models in different regimes.** The top figure plots $\mathcal{E}_{\texttt{mem}}, \mathcal{E}_{\texttt{gen}}, \ell_{\texttt{train}}, \ell_{\texttt{test}}$ over training epochs for different different dataset sizes: $N = 2^6$ (left), 2^{12} (middle), 2^{16} (right). The bottom figure visualizes the generation when $N = 2^{12}$. The top row shows samples from the underlying distribution $\mathbf{\Phi}_{p_{\texttt{data}}}(\boldsymbol{x}_T)$, while the middle and bottom rows display outputs from the trained diffusion model $\mathbf{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)$ at epoch 85 and 500, respectively.

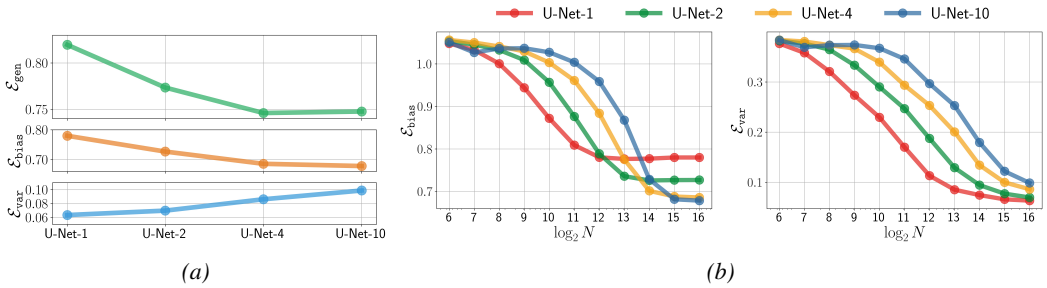


Figure 5. **Bias–Variance Trade-off.** (a) plots the generalization error $\mathcal{E}_{\texttt{gen}}$, bias $\mathcal{E}_{\texttt{bias}}$, and variance $\mathcal{E}_{\texttt{var}}$ across different network architectures with a fixed distillation sample size of $N = 2^{16}$. (b) shows $\mathcal{E}_{\texttt{bias}}$ and $\mathcal{E}_{\texttt{var}}$ as functions of the number of distillation samples N for various network architectures.

increasing $|\boldsymbol{\theta}|$ consistently increases $\mathcal{E}_{\texttt{var}}$, and its effect on $\mathcal{E}_{\texttt{bias}}$ depends on the size of N: it decreases $\mathcal{E}_{\texttt{bias}}$ when $N \geq 2^{15}$ but increases it when $N \leq 2^{11}$.

5. Discussion & Conclusion

While PFD is designed to evaluate generalization in distillation settings, we believe its impact extends beyond distillation and is highly relevant to diffusion model training:

Distillation is closely aligned with training from scratch. The teacher-student framework has been widely adopted to understand learning phenomena observed in real-world scenarios, both empirically (Betzalel et al., 2024; Lee et al., 2021; Saglietti et al., 2022) and theoretically (Advani et al., 2020; Goldt et al., 2019; d'Ascoli et al., 2020). In this work, the phenomena we observe, such as the MtoG transition (left column of Figure 2) and early learning dynamics (middle column of Figure 4), are consistent with observations reported in prior studies of models trained from scratch. This alignment suggests that distillation provides a tractable and controllable setting for understanding training in real-world.

Training from scratch is increasingly close to distillation. Modern diffusion models are often trained on large-scale datasets constructed from heterogeneous sources. Due to the scarcity of sufficiently high-quality real-world data, large-scale datasets such as LAION-5B (Schuhmann et al., 2022) inevitably contain a substantial fraction of synthetic data generated by foundation models. As a result, training from scratch increasingly resembles a teacher-student setting, where models implicitly learn from other pretrained models rather than purely from natural data.

Taken together, these facts motivate the use of PFD beyond explicit distillation settings, as a general tool for analyzing generalization in real-world large-scale diffusion models.

6. Impact Statement

This work advances the theoretical and empirical understanding of generalization in diffusion models by introducing Probability Flow Distance (PFD). Improved understand-

ing of generalization can help mitigate risks associated with memorization, including privacy leakage and copyright concerns. We do not anticipate direct negative societal impacts arising from this work; however, as with all advances in generative modeling, downstream applications should be developed and deployed responsibly.

References

Advani, M. S., Saxe, A. M., and Sompolinsky, H. High-dimensional dynamics of generalization error in neural networks. *Neural Networks*, 132:428–446, 2020.

Aithal, S. K., Maini, P., Lipton, Z., and Kolter, J. Z. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 134614–134644, 2024.

Alaa, A., Van Breugel, B., Saveliev, E. S., and Van Der Schaar, M. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*, pp. 290–306. PMLR, 2022.

Ambrogioni, L. In search of dispersed memories: Generative diffusion models are associative memory networks. *Entropy*, 26(5):381, 2024.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and equilibrium in generative adversarial nets (gans). In *International Conference on Machine Learning*, pp. 224–232. PMLR, 2017.

Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., and Zhu, J. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Belkin, M., Hsu, D., Ma, S., and Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Berthelot, D., Autef, A., Lin, J., Yap, D. A., Zhai, S., Hu, S., Zheng, D., Talbott, W., and Gu, E. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248*, 2023.

Betzalel, E., Penso, C., and Fetaya, E. Evaluation metrics for generative models: An empirical study. *Machine Learning and Knowledge Extraction*, 6(3):1531–1544, 2024.

Bhatia, R. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1lUOzWCW.

Block, A., Mroueh, Y., and Rakhlin, A. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Bonnaire, T., Urfin, R., Biroli, G., and Mezard, M. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=BSZqpqgqM0.

Borkar, J., Chadha, K., Mireshghallah, N., Zhang, Y., Veliche, I.-E., Mitra, A., Smith, D. A., Xu, Z., and Garcia-Olano, D. Memorization dynamics in knowledge distillation for language models. *arXiv preprint arXiv:2601.15394*, 2026.

Bortoli, V. D. Convergence of denoising diffusion models under the manifold hypothesis. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=MhK5aXo3gB. Expert Certification.

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramer, F., Balle, B., Ippolito, D., and Wallace, E. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Chen, H., Lee, H., and Lu, J. Improved analysis of score-based generative modeling: User-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pp. 4735–4763. PMLR, 2023a.

Chen, J., Hu, D., Huang, X., Coskun, H., Sahni, A., Gupta, A., Goyal, A., Lahiri, D., Singh, R., Idelbayev, Y., et al. Snapgen: Taming high-resolution text-to-image models for mobile devices with efficient architectures and training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 7997–8008, 2025.

Chen, M., Huang, K., Zhao, T., and Wang, M. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *Proceedings of the 40th International Conference on Machine Learning*, 2023b.

Chen, S., Chewi, S., Lee, H., Li, Y., Lu, J., and Salim, A. The probability flow ODE is provably fast. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023c. URL https://openreview.net/forum?id=KD6MFeWSAd.

Chen, S., Chewi, S., Li, J., Li, Y., Salim, A., and Zhang, A. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. In *The Eleventh International Conference on Learning Representations*, 2023d. URL https://openreview.net/forum?id=zyLVMgsZ0U_.

Chen, S., Daras, G., and Dimakis, A. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *International Conference on Machine Learning*, pp. 4462–4484. PMLR, 2023e.

Chen, Z. On the interpolation effect of score smoothing. *arXiv preprint arXiv:2502.19499*, 2025.

Chong, M. J. and Forsyth, D. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6070–6079, 2020.

Coddington, E. A. and Levinson, N. *Theory of ordinary differential equations*. McGraw-Hill New York, 1955.

d'Ascoli, S., Sagun, L., and Biroli, G. Triple descent and the two kinds of overfitting: Where & why do they appear? In *Advances in Neural Information Processing Systems*, volume 33, pp. 3058–3069, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Gao, R., Hoogeboom, E., Heek, J., Bortoli, V. D., Murphy, K. P., and Salimans, T. Diffusion models and gaussian flow matching: Two sides of the same coin. In *The Fourth Blogpost Track at ICLR 2025*, 2025a. URL https://openreview.net/forum?id=C8Yyg9wy0s.

Gao, X. and Zhu, L. Convergence analysis for general probability flow ODEs of diffusion models in wasserstein distances. In *The 28th International Conference on Artificial Intelligence and Statistics*, 2025. URL https://openreview.net/forum?id=EkO8rb3liX.

Gao, X., Nguyen, H. M., and Zhu, L. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of Machine Learning Research*, 26(43):1–54, 2025b. URL http://jmlr.org/papers/v26/24-0902.html.

Goldt, S., Advani, M., Saxe, A. M., Krzakala, F., and Zdeborová, L. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *arXiv preprint arXiv:2310.02664*, 2023.

Gu, X., Du, C., Pang, T., Li, C., Lin, M., and Wang, Y. On memorization in diffusion models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=D3DBqvSDbj.

Gulrajani, I., Raffel, C., and Metz, L. Towards GAN benchmarks which require generalization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HkxKH2AcFm.

Hastie, T., Tibshirani, R., Friedman, J., et al. The elements of statistical learning, 2009.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. CLIPScore: a reference-free evaluation metric for image captioning. In *Empirical Methods in Natural Language Processing*, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Hoeffding, W. Probability inequalities for sums of bounded random variables. *The Collected Works of Wassily Hoeffding*, pp. 409–426, 1994.

Hu, D., Gupta, A., Gabidolla, M., Sahni, A., Coskun, H., Li, Y., Idelbayev, Y., Mahmood, A., Lebedev, A., Lahiri, D., et al. Snapgen++: Unleashing diffusion transformers for efficient high-fidelity image generation on edge devices. *arXiv preprint arXiv:2601.08303*, 2026.

Hutchinson, M. F. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

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

Jolicoeur-Martineau, A., Piché-Taillefer, R., Mitliagkas, I., and des Combes, R. T. Adversarial score matching and improved sampling for image generation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=eLfqMl3z3lq.

Kadkhodaie, Z., Guth, F., Simoncelli, E. P., and Mallat, S. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ANvmVS2Yr0.

Kamb, M. and Ganguli, S. An analytic theory of creativity in convolutional diffusion models. *arXiv preprint arXiv:2412.20292*, 2024.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26565–26577, 2022a.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. <https://github.com/NVlabs/edm/tree/main>, 2022b.

Karras, T., Aittala, M., Lehtinen, J., Hellsten, J., Aila, T., and Laine, S. Analyzing and improving the training dynamics of diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24174–24184, 2024.

Kim, B.-K., Song, H.-K., Castells, T., and Choi, S. Bk-sdm: A lightweight, fast, and cheap version of stable diffusion. In *European Conference on Computer Vision*, pp. 381–399. Springer, 2024.

Kohavi, R., Wolpert, D. H., et al. Bias plus variance decomposition for zero-one loss functions. In *International Conference on Machine Learning*, volume 96, pp. 275–283. Citeseer, 1996.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Lee, H., Lu, J., and Tan, Y. Convergence for score-based generative modeling with polynomial complexity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 22870–22882, 2022.

Lee, S., Goldt, S., and Saxe, A. Continual learning in the teacher-student setup: Impact of task similarity. In *International Conference on Machine Learning*, pp. 6109–6119. PMLR, 2021.

Li, G., Huang, Y., Efimov, T., Wei, Y., Chi, Y., and Chen, Y. Accelerating convergence of score-based diffusion models, provably. In *Proceedings of the 41st International Conference on Machine Learning*, 2024a.

Li, G., Wei, Y., Chen, Y., and Chi, Y. Towards non-asymptotic convergence for diffusion-based generative models. In *The Twelfth International Conference on Learning Representations*, 2024b. URL https://openreview.net/forum?id=4VGEeER6W9.

Li, G., Wei, Y., Chi, Y., and Chen, Y. A sharp convergence theory for the probability flow odes of diffusion models. *arXiv preprint arXiv:2408.02320*, 2024c.

Li, P., Li, Z., Zhang, H., and Bian, J. On the generalization properties of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 2097–2127, 2023.

Li, X., Dai, Y., and Qu, Q. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024d. URL https://openreview.net/forum?id=Sk2duBGvrK.

Liang, J., Huang, Z., and Chen, Y. Low-dimensional adaptation of diffusion models: Convergence in total variation. *arXiv preprint arXiv:2501.12982*, 2025.

Liang, Y., Shi, Z., Song, Z., and Zhou, Y. Unraveling the smoothness properties of diffusion models: A gaussian mixture perspective. *arXiv preprint arXiv:2405.16418*, 2024.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Liu, X., Gong, C., and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=XVjTT1nw5z.

Loaiza-Ganem, G., Ross, B. L., Hosseinzadeh, R., Caterini, A. L., and Cresswell, J. C. Deep generative models through the lens of the manifold hypothesis: A survey and new connections. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a90WpmSi0I. Survey Certification, Expert Certification.

Lukoianov, A., Yuan, C., Solomon, J., and Sitzmann, V. Locality in image diffusion models emerges from data statistics. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=skunuOdavO.

Luo, C. Understanding diffusion models: A unified perspective. *arXiv preprint arXiv:2208.11970*, 2022.

Ma, X., Yu, R., Liu, S., Fang, G., and Wang, X. Diffusion model is effectively its own teacher. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 12901–12911, 2025.

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J., and Salimans, T. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14297–14306, 2023.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., and Sutskever, I. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Nichol, A. Q. and Dhariwal, P. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Nie, S., Guo, H. A., Lu, C., Zhou, Y., Zheng, C., and Li, C. The blessing of randomness: SDE beats ODE in general diffusion-based image editing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DesYwmUG00.

Niedoba, M., Zwartsenberg, B., Murphy, K., and Wood, F. Towards a mechanistic explanation of diffusion model generalization. *arXiv preprint arXiv:2411.19339*, 2024.

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. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt. Featured Certification.

Pham, B., Raya, G., Negri, M., Zaki, M. J., Ambrogioni, L., and Krotov, D. Memorization to generalization: Emergence of diffusion models from associative memory networks. In *New Frontiers in Associative Memories*, 2025. URL https://openreview.net/forum?id=IWZnhP3YgK.

Pizzi, E., Roy, S. D., Ravindra, S. N., Goyal, P., and Douze, M. A self-supervised descriptor for image copy detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14532–14542, 2022.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FjNys5c7VyY.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PmLR, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Rudin, W. Principles of mathematical analysis. 2021.

Saglietti, L., Mannelli, S., and Saxe, A. An analytical theory of curriculum learning in teacher-student networks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 21113–21127, 2022.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Salimans, T. and Ho, J. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294, 2022.

Skilling, J. The eigenvalues of mega-dimensional matrices. *Maximum Entropy and Bayesian Methods: Cambridge, England, 1988*, pp. 455–466, 1989.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6048–6058, 2023a.

Somepalli, G., Singla, V., Goldblum, M., Geiping, J., and Goldstein, T. Understanding and mitigating copying in diffusion models. *arXiv preprint arXiv:2305.20086*, 2023b.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=St1giarCHLP.

Song, Y., Durkan, C., Murray, I., and Ermon, S. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428, 2021b.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021c. URL https://openreview.net/forum?id=PxTIG12RRHS.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. In *International Conference on Machine Learning*, 2023.

Stein, G., Cresswell, J., Hosseinzadeh, R., Sui, Y., Ross, B., Villecroze, V., Liu, Z., Caterini, A. L., Taylor, E., and Loaiza-Ganem, G. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

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

Tinaz, B., Fabian, Z., and Soltanolkotabi, M. Emergence and evolution of interpretable concepts in diffusion models through the lens of sparse autoencoders. In *Second Workshop on Visual Concepts*, 2025. URL https://openreview.net/forum?id=plCImt7QaW.

Vastola, J. Generalization through variance: how noise shapes inductive biases in diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7lUdo8Vuqa.

Wang, B. An analytical theory of power law spectral bias in the learning dynamics of diffusion models. *arXiv preprint arXiv:2503.03206*, 2025.

Wang, B. and Vastola, J. J. The unreasonable effectiveness of gaussian score approximation for diffusion models and its applications. *arXiv preprint arXiv:2412.09726*, 2024.

Wang, F.-Y., Huang, Z., Bergman, A., Shen, D., Gao, P., Lingelbach, M., Sun, K., Bian, W., Song, G., Liu, Y., et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024a.

Wang, P., Zhang, H., Zhang, Z., Chen, S., Ma, Y., and Qu, Q. Diffusion models learn low-dimensional distributions via subspace clustering. *arXiv preprint arXiv:2409.02426*, 2024b.

Xie, S., Xiao, Z., Kingma, D. P., Hou, T., Wu, Y. N., Murphy, K. P., Salimans, T., Poole, B., and Gao, R. EM distillation for one-step diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=rafVvthuxD.

Xu, Y., Zhao, Y., Xiao, Z., and Hou, T. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8196–8206, 2024.

Yang, R., Wang, Z., Jiang, B., and Li, S. Leveraging drift to improve sample complexity of variance exploding diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107662–107702, 2024.

Yang, R., Li, Y., Jiang, B., Chen, C., and Li, S. Multi-subspace multi-modal modeling for diffusion models: Estimation, convergence and mixture of experts. In *The Fourteenth International Conference on Learning*

Representations, 2026. URL https://openreview.net/forum?id=MPWIM6rxxU.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pp. 10767–10777. PMLR, 2020.

Ye, Z., Zhu, Q., Tao, M., and Chen, M. Provable separations between memorization and generalization in diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=42gfTZzyvV.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., and Freeman, B. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Yoon, T., Choi, J. Y., Kwon, S., and Ryu, E. K. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Zhang, H., Zhou, J., Lu, Y., Guo, M., Wang, P., Shen, L., and Qu, Q. The emergence of reproducibility and consistency in diffusion models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60558–60590. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhang24cn.html.

Zhang, Y., Chen, T., Wang, Z., Wu, Y. N., Zhou, M., and Leong, O. Restoration score distillation: From corrupted diffusion pretraining to one-step high-quality generation. *arXiv preprint arXiv:2505.13377*, 2025.

Zhang, Z., Li, X., Li, X., Shi, L., Wu, M., Tao, M., and Qu, Q. Generalization of diffusion models arises with a balanced representation space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=57THeGgNAN.

Zhu, Z., Locatello, F., and Cevher, V. Sample complexity bounds for score-matching: Causal discovery and generative modeling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 3325–3337, 2023.

The appendix is organized as follows. We first discuss related work in Section A. Next, we provide detailed proofs for Section 2 in Section B. Experimental settings and additional discussions for Section 3 and Section 4 are presented in Section C. We then offer further discussion related to $\mathcal{E}_{\texttt{mem}}$ in Section D. Finally, ablation studies for PFD are included in Section E.

A. Related Works

In this section, we briefly review related work on generalization metrics for diffusion models, discuss diffusion model generalizability, and cover the fundamentals of training diffusion models.

A.1. Generalization Metrics for Diffusion Models

Generalization metrics quantify the distance between the learned distribution and the underlying data distribution in diffusion models. To measure this distributional gap, theoretical works commonly employ metrics such as Kullback-Leibler (KL) divergence (Chen et al., 2023e; Nie et al., 2024; Li et al., 2023), total variation (TV) (Chen et al., 2023c; Li et al., 2024b;a; Yang et al., 2024; Li et al., 2024c; Liang et al., 2025), and Wasserstein distance (Gao & Zhu, 2025; Bortoli, 2022; Chen et al., 2023a; Gao et al., 2025b). However, these metrics are practically inefficient for diffusion models. Practical metrics focus on various perspective, including negative log-likelihood (NLL) (Song et al., 2021c), image generation quality: Fréchet inception distance (FID) (Heusel et al., 2017), inception score (IS) (Salimans et al., 2016), $\texttt{FD}_{\texttt{dinov2}}$ (Stein et al., 2024), maximum mean discrepancy (MMD) (Bińkowski et al., 2018), CLIP maximum mean discrepancy (CMMD) (Jayasumana et al., 2024); alignment: CLIPscore (Hessel et al., 2021), and precision, recall (Sajjadi et al., 2018; Kynkäänniemi et al., 2019). However, these practical metrics are not explicitly designed to evaluate the generalizability of diffusion models. Thus, there is a need for a generalization metric that are both theoretical grounded and practically efficient for diffusion models. To address this gap, we propose PFD, a novel generalization metric that is theoretically proven to be a valid distributional distance and can be efficiently approximated by its empirical version using a polynomial number of samples. In practice, PFD requires fewer samples for estimation and is the only existing metric that explicitly quantifies generalization in diffusion models.

A.2. Diffusion Model Generalizability

Recent works have shown that diffusion models transition from memorization to generalization as the number of training samples increases (Yoon et al., 2023; Zhang et al., 2024). With sufficient data, models trained with different architectures, loss functions, and even disjoint datasets can reproduce each other's outputs, indicating a strong convergence toward the underlying data distribution (Zhang et al., 2024; Kadkhodaie et al., 2024). To explain this strong generalization, (Kadkhodaie et al., 2024) attribute it to the emergence of a geometric-adaptive harmonic basis, while others argue that generalization arises from interpolation across the data manifold (Aithal et al., 2024; Chen, 2025). In contrast, (Vastola, 2025) explain generalization through inductive biases in the noise, whereas (Zhang et al., 2026) attribute it to the emergence of a balanced representation space. Theoretical insights by (Li et al., 2023) provide generalization bounds using KL-divergence under simplified models, (Ye et al., 2026) establish a dual theoretical separation explaining memorization and generalization. More recent efforts focus on characterizing the learned noise-to-image mapping for generalized diffusion models, either through Gaussian parameterizations (Li et al., 2024d; Wang & Vastola, 2024), mixture of low rank Gaussian parameterizations (Wang et al., 2024b), multi-subspace multi-modal modeling (Yang et al., 2026) or patch-wise optimal score functions (Niedoba et al., 2024; Kamb & Ganguli, 2024), and analyses based on statistical properties of image datasets (Lukoianov et al., 2025). In parallel, generalization has also been studied through the lens of associative memory (Pham et al., 2025; Ambrogioni, 2024) and sparse autoencoders (Tinaz et al., 2025). However, despite these theoretical analyses and qualitative insights, prior work lacks a quantitative framework for measuring generalizability. In this paper, we propose PFD, a metric that enables such quantitative evaluation. Using this measure, we uncover further insights into the generalization behavior of diffusion models, as discussed in Section 4.

A.3. Diffusion Model Distillation

Recent work on diffusion distillation has demonstrated substantial efficiency gains through both lightweight student architectures (Chen et al., 2025; Kim et al., 2024; Hu et al., 2026) and reduced sampling trajectories (Yin et al., 2024b;a; Salimans & Ho, 2022; Song et al., 2023; Meng et al., 2023; Sauer et al., 2024; Xie et al., 2024). Representative approaches include progressive distillation (Salimans & Ho, 2022; Berthelot et al., 2023), adversarial distillation (Jolicoeur-Martineau

et al., 2021; Xu et al., 2024; Sauer et al., 2024), and consistency-based distillation methods (Song et al., 2023; Wang et al., 2024a). Beyond architectural and sampling efficiency, Meng et al. (Meng et al., 2023) distill the classifier-free guidance score (Ho & Salimans, 2022) to further accelerate guidance at inference time. Moreover, diffusion distillation has been successfully applied beyond efficiency improvements, including text-to-3D generation (Poole et al., 2023), learning from corrupted data distributions (Zhang et al., 2025), and even achieving unexpected gains in generation quality (Ma et al., 2025). In this paper, we adopt the most straightforward distillation paradigm, where the student is trained on data generated by a teacher model. Nevertheless, PFD is orthogonal to the choice of distillation strategy and can be naturally extended to more advanced techniques, such as progressive, adversarial, or consistency-based distillation frameworks, offering a flexible foundation for future extensions.

A.4. Training Diffusion Models

To enable sampling via the PF-ODE (1), we train a neural network $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ to approximate the score function $\nabla \log p_t(\boldsymbol{x}_t)$ using denoising score matching loss (Song et al., 2021c):

$$\min_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \int_0^T \lambda_t \mathbb{E}_{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, T^2 \boldsymbol{I}_n)} \left[\left\| \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)} + t\boldsymbol{\epsilon}, t) + \boldsymbol{\epsilon}/t \right\|_2^2 \right] \mathrm{d}t, \tag{10}$$

λ_t denotes a scalar weight for the loss at t. Given the learned score function, the corresponding noise-to-image mapping is:

$$\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T) = \boldsymbol{x}_T - \int_T^0 t \boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) \mathrm{d}t. \tag{11}$$

Although alternative training objectives exist, such as predicting noise $\boldsymbol{x}_T$ (Ho et al., 2020), clean image $\boldsymbol{x}_0$ (Karras et al., 2022a), rectified flow $\boldsymbol{x}_T - \boldsymbol{x}_0$ (Liu et al., 2023) or other linear combinations of $\boldsymbol{x}_0$ and $\boldsymbol{x}_T$ (Salimans & Ho, 2022), prior works (Luo, 2022; Gao et al., 2025a) have shown that it is still possible to recover an approximate score function $\boldsymbol{s}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)$ from these methods.

B. Proof in Section 2

Proof of Theorem 1. It is trivial to show $\mathtt{PFD}(p, q) > 0$ for any $p \neq q$ and $\mathtt{PFD}(p, q) = \mathtt{PFD}(q, p)$, and thus we omit the proof.

- Proof of $p = q \Leftrightarrow \mathtt{PFD}(p, q) = 0$:
  - ($\Rightarrow$) If $p = q$, $\nabla \log p_t(\boldsymbol{x}_t) = \nabla \log q_t(\boldsymbol{x}_t)$, thus:

$$\mathrm{d}\boldsymbol{x}_t = -t \left(\nabla \log p_t(\boldsymbol{x}_t) - \nabla \log q_t(\boldsymbol{x}_t) \right) \mathrm{d}t = 0 \tag{12}$$

    Thus, $\boldsymbol{\Phi}_p(\boldsymbol{x}_T) - \boldsymbol{\Phi}_q(\boldsymbol{x}_T)$ is the solution of the ODE function Equation (12) with initial $\boldsymbol{x}_T = \mathbf{0}$. Thus $\boldsymbol{\Phi}_p(\boldsymbol{x}_T) - \boldsymbol{\Phi}_1(\boldsymbol{x}_T) = \mathbf{0}$ for all $\boldsymbol{x}_T$. Thus $\mathtt{PFD}(p, q) = 0$
  - ($\Leftarrow$) If $\mathtt{PFD}(p, q) = 0$ and $\boldsymbol{\Phi}_p, \boldsymbol{\Phi}_q$ are continuous function w.r.t $\boldsymbol{x}_T$, then we have $\boldsymbol{\Phi}_p(\boldsymbol{x}_T) = \boldsymbol{\Phi}_q(\boldsymbol{x}_T)$ for all $\boldsymbol{x}_T$. If $\boldsymbol{x}_0 = \boldsymbol{\Phi}(\boldsymbol{x}_T)$, from the transformation of probability identities, we have:

$$p(\boldsymbol{x}_0) = \frac{\partial}{\partial [\boldsymbol{x}_0]_1} \cdots \frac{\partial}{\partial [\boldsymbol{x}_0]_n} \int_{\{\boldsymbol{\epsilon} | \boldsymbol{\Phi}(\boldsymbol{\epsilon}) \leq \boldsymbol{x}_0\}} p_{\mathcal{N}}(\boldsymbol{\epsilon}) \mathrm{d}^n \boldsymbol{\epsilon}, \tag{13}$$

    where $[\boldsymbol{x}_0]_i$ denotes the i-th element of $\boldsymbol{x}_0$, $\boldsymbol{f}(\boldsymbol{\epsilon}) \leq \boldsymbol{x}_0$ denotes the element wise less or equal. $p_{\mathcal{N}}(\cdot)$ is the probability

density function (PDF) of Gaussian distribution $\mathcal{N}\left(\mathbf{0}, T^2\boldsymbol{I}_n\right)$. Thus, for all $\boldsymbol{x}_0$ we have:

$$\begin{aligned} p(\boldsymbol{x}_0) - q(\boldsymbol{x}_0) &= \frac{\partial}{\partial[\boldsymbol{x}_0]_1} \cdots \frac{\partial}{\partial[\boldsymbol{x}_0]_n} \int_{\{\boldsymbol{\epsilon}|\boldsymbol{\Phi}_p(\boldsymbol{\epsilon})\leq \boldsymbol{x}_0\}} p_{\mathcal{N}}(\boldsymbol{\epsilon})\mathrm{d}^n\boldsymbol{\epsilon} \\ &\quad - \frac{\partial}{\partial[\boldsymbol{x}_0]_1} \cdots \frac{\partial}{\partial[\boldsymbol{x}_0]_n} \int_{\{\boldsymbol{\epsilon}|\boldsymbol{\Phi}_q(\boldsymbol{\epsilon})\leq \boldsymbol{x}_0\}} p_{\mathcal{N}}(\boldsymbol{\epsilon})\mathrm{d}^n\boldsymbol{\epsilon}, \\ &= \frac{\partial}{\partial[\boldsymbol{x}_0]_1} \cdots \frac{\partial}{\partial[\boldsymbol{x}_0]_n} \int_{\{\boldsymbol{\epsilon}|\boldsymbol{\Phi}_p(\boldsymbol{\epsilon})\leq \boldsymbol{x}_0\}} p_{\mathcal{N}}(\boldsymbol{\epsilon})\mathrm{d}^n\boldsymbol{\epsilon} \\ &\quad - \frac{\partial}{\partial[\boldsymbol{x}_0]_1} \cdots \frac{\partial}{\partial[\boldsymbol{x}_0]_n} \int_{\{\boldsymbol{\epsilon}|\boldsymbol{\Phi}_p(\boldsymbol{\epsilon})\leq \boldsymbol{x}_0\}} p_{\mathcal{N}}(\boldsymbol{\epsilon})\mathrm{d}^n\boldsymbol{\epsilon}, \\ &= 0, \end{aligned} \tag{14}$$

so $p = q$.

- Proof of $\texttt{PFD}(p, q) \leq \texttt{PFD}(p, p') + \texttt{PFD}(p', q)$:

$$\begin{aligned} &\texttt{PFD}(p, q) \\ =& \left(\mathbb{E}_{\boldsymbol{x}_T\sim\mathcal{N}(0,T^2\boldsymbol{I})}\left[\left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T\right)\right\|_2^2\right]\right)^{1/2} \\ \leq& \left(\mathbb{E}_{\boldsymbol{x}_T\sim\mathcal{N}(0,T^2\boldsymbol{I})}\left[\left(\left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_{p'}\left(\boldsymbol{x}_T\right)\right\|_2 + \left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_{p'}\left(\boldsymbol{x}_T\right)\right\|_2\right)^2\right]\right)^{1/2} \\ \leq& \left(\mathbb{E}_{\boldsymbol{x}_T\sim\mathcal{N}(0,T^2\boldsymbol{I})}\left[\left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T\right)\right\|_2^2\right]\right)^{1/2} \\ &+ \left(\mathbb{E}_{\boldsymbol{x}_T\sim\mathcal{N}(0,T^2\boldsymbol{I})}\left[\left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T\right)\right\|_2^2\right]\right)^{1/2} \\ =& \texttt{PFD}(p, p') + \texttt{PFD}(p', q) \end{aligned} \tag{15}$$

□

Lemma 1. *Under Assumption 1, for all $\boldsymbol{x}_T \in \mathcal{N}(\mathbf{0}, T^2\boldsymbol{I}_n)$, as $T \to \infty$, we have:*

$$\left\|\boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T\right)\right\|_2 \leq \exp\left(\frac{LT_\xi^2}{2}\right)\xi + \frac{\epsilon}{L}\left(\exp\left(\frac{LT_\xi^2}{2}\right) - 1\right), \tag{16}$$

where ξ is a numerical constant and a finite timestep T_ξ depending only on ξ.

Proof of Lemma 1. Let $\boldsymbol{\phi}_t, t \in [0, T]$ denotes the ODE trajectory:

$$\begin{aligned} \boldsymbol{\phi}_t &= \boldsymbol{x}_t^p - \boldsymbol{x}_t^q, \\ \boldsymbol{x}_t^p &= \boldsymbol{x}_T - \int_T^t \tau\nabla_{\boldsymbol{x}} \log p_\tau(\boldsymbol{x}_\tau^p)\mathrm{d}\tau, \\ \boldsymbol{x}_t^q &= \boldsymbol{x}_T - \int_T^t \tau\nabla_{\boldsymbol{x}} \log q_\tau(\boldsymbol{x}_\tau^q)\mathrm{d}\tau, \end{aligned} \tag{17}$$

From the definition, $\boldsymbol{\phi}_0 = \boldsymbol{\Phi}_p\left(\boldsymbol{x}_T\right) - \boldsymbol{\Phi}_q\left(\boldsymbol{x}_T\right)$. Because $\lim_{T\to\infty} \boldsymbol{\phi}_t = \boldsymbol{x}_T - \boldsymbol{x}_T = \mathbf{0}$, from the $\epsilon - \delta$ definition of the limit, given $\boldsymbol{x}_T$, and a constant ξ , there exists a finite T_ξ related to ξ such that:

$$\|\boldsymbol{\phi}_t\|_2 \leq \xi \quad \text{for all } t \geq T_\xi. \tag{18}$$

As $t \leq T_\xi$, we have:

$$\begin{aligned} &\frac{\mathrm{d}\boldsymbol{\phi}_t}{\mathrm{d}t} = -t\left(\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t^p) - \nabla_{\boldsymbol{x}} \log q_t(\boldsymbol{x}_t^q)\right), \\ &\left\|\boldsymbol{\phi}_{T_\xi}\right\|_2 \leq \xi. \end{aligned} \tag{19}$$

Apply Assumption 1 to Equation (19), we could obtain the following integral inequality w.r.t $\|\boldsymbol{\phi}_t\|_2$:

$$\begin{aligned}\frac{\mathrm{d}\,\|\boldsymbol{\phi}_t\|_2}{\mathrm{d}t} &\leq \left\|\frac{\mathrm{d}\boldsymbol{\phi}_t}{\mathrm{d}t}\right\|_2 \\ &\leq t\,\|\nabla_{\boldsymbol{x}}\log p_t(\boldsymbol{x}_t^p) - \nabla_{\boldsymbol{x}}\log q_t(\boldsymbol{x}_t^q)\|_2 \\ &\leq t\left(\epsilon + L\,\|\boldsymbol{\phi}_t\|_2\right), \\ \left\|\boldsymbol{\phi}_{T_\xi}\right\|_2 &\leq \xi,\ \ 0 \leq t \leq T_\xi,\end{aligned} \tag{20}$$

where the first inequality comes from the fact that $\frac{\mathrm{d}\,\|\boldsymbol{\phi}_t\|_2}{\mathrm{d}t} \leq \left\|\frac{\mathrm{d}\boldsymbol{\phi}_t}{\mathrm{d}t}\right\|_2$. From Grönwall's inequality (Coddington & Levinson, 1955), we could solve $\|\boldsymbol{\Phi}_p(\boldsymbol{x}_T) - \boldsymbol{\Phi}_q(\boldsymbol{x}_T)\|_2 = \|\boldsymbol{\phi}_0\|_2 \leq \exp(\frac{LT_\xi^2}{2})\xi + \frac{\epsilon}{L}\left(\exp(\frac{LT_\xi^2}{2}) - 1\right)$. □

Proof of Theorem 2. Let $\boldsymbol{X} := \|\boldsymbol{\Phi}_p(\boldsymbol{x}_T) - \boldsymbol{\Phi}_q(\boldsymbol{x}_T)\|_2^2$. From Lemma 1,

$$0 \leq \boldsymbol{X} \leq \kappa^2(L,\epsilon),$$

with $\kappa(L,\epsilon) := \exp\left(\frac{LT_\xi^2}{2}\right)\xi + \frac{\epsilon}{L}\left(\exp\left(\frac{LT_\xi^2}{2}\right) - 1\right)$. From Hoeffding's inequality (Hoeffding, 1994), we have:

$$\mathbb{P}\left(\left|\mathbb{E}[\boldsymbol{X}] - \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{X}_i\right| \geq \gamma\right) \leq 2\exp\left(-\frac{2M\gamma^2}{\kappa^4(L,\epsilon)}\right), \tag{21}$$

with M samples to achieve γ accuracy. Thus, we could guarantee $\mathbb{P}\left(\left|\mathbb{E}[\boldsymbol{X}] - \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{X}_i\right| \leq \gamma\right)$ with probability η, when:

$$M \geq \frac{\kappa^4(L,\epsilon)}{2\gamma^2}\log\frac{2}{\eta}. \tag{22}$$

Because

$$\left|\mathtt{PFD}(p,q) - \hat{\mathtt{PFD}}(p,q)\right| = \left|\sqrt{\mathbb{E}[\boldsymbol{X}]} - \sqrt{\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{X}_i}\right| \tag{23}$$

$$\leq \sqrt{\left|\mathbb{E}[\boldsymbol{X}] - \frac{1}{M}\sum_{i=1}^{M}\boldsymbol{X}_i\right|}. \tag{24}$$

We could guarantee that $\mathbb{P}\left(\left|\mathtt{PFD}(p,q) - \hat{\mathtt{PFD}}(p,q)\right| \leq \gamma\right)$ with probability η, when:

$$M \geq \frac{\kappa^4(L,\epsilon)}{2\gamma^4}\log\frac{2}{\eta}. \tag{25}$$

□

Example 1. *The Wasserstein-2 distance* $\mathrm{W}_2(\cdot,\cdot)$ *is the lower bound of the probability flow distance, i.e.,*

$$\mathrm{W}_2(p,q) \leq \mathtt{PFD}(p,q), \tag{26}$$

Specifically, let p *and* q *be multivariate Gaussian distribution* $\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, $\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$, *respectively, where* $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2 \in \mathbb{R}^n$ *and* $\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2 \in \mathbb{R}^{n\times n}$. *The* PFD *is given by*

$$\mathtt{PFD}(p,q) = \left(\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2 + \left\|\boldsymbol{\Sigma}_1^{1/2} - \boldsymbol{\Sigma}_2^{1/2}\right\|_F\right)^{1/2}, \tag{27}$$

under this case, the equality $\mathtt{PFD}(p,q) = W_2(p,q)$ *holds when* $\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1$.

Proof of Example 1. Proof of $W_2(p,q) \leq \texttt{PFD}(p,q)$. From the definition of Wasserstein-2 distance:

$$W_2(p,q) = \inf_{\gamma \in \Gamma(p,q)} \left(\mathbb{E}_{(\boldsymbol{x}_p, \boldsymbol{x}_q) \sim \gamma} \left\| \boldsymbol{x}_p - \boldsymbol{x}_q \right\|_2^2 \right)^{1/2}, \tag{28}$$

where $\Gamma(p,q)$ is the set of all couplings of p and q. As proofed by (Song et al., 2021c), the noise-to-image mapping $\boldsymbol{\Phi}_p$ and $\boldsymbol{\Phi}_q$ pushes the Gaussian distribution $\mathcal{N}\left(\boldsymbol{0}, T^2 \boldsymbol{I}_n\right)$ to the p and q distribution respectively. Thus we could find the coupling $\gamma_{\texttt{PFD}} \coloneqq (\boldsymbol{\Phi}_p, \boldsymbol{\Phi}_q)_{\#} \mathcal{N}\left(\boldsymbol{0}, T^2 \boldsymbol{I}_n\right)$, i.e., the pushforward of $\mathcal{N}\left(\boldsymbol{0}, T^2 \boldsymbol{I}_n\right)$ by $(\boldsymbol{\Phi}_p, \boldsymbol{\Phi}_q)$, such that

$$\texttt{PFD}(p,q) = \left(\mathbb{E}_{(\boldsymbol{x}_p, \boldsymbol{x}_q) \sim \gamma_{\texttt{PFD}}} \left\| \boldsymbol{x}_p - \boldsymbol{x}_q \right\|_2^2 \right)^{1/2} \geq W_2(p,q) \tag{29}$$

When distribution $p(\boldsymbol{x})$ is Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$. Thus $p_t(\boldsymbol{x})$ is $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} + \sigma_t^2 \boldsymbol{I}_n\right)$, thus the score function could be caluclated as,

$$\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}) = \left(\boldsymbol{\Sigma} + t^2 \boldsymbol{I}_n\right)^{-1} (\boldsymbol{\mu} - \boldsymbol{x}). \tag{30}$$

By plugging in Equation (30) to Equation (2), we could obtain the ODE equation w.r.t $\boldsymbol{x}$:

$$\mathrm{d}\boldsymbol{x} = -t \left(\boldsymbol{\Sigma} + t^2 \boldsymbol{I}_n\right)^{-1} (\boldsymbol{\mu} - \boldsymbol{x}) \, \mathrm{d}t, . \tag{31}$$

The above ODE equation has a close form solution:

$$\boldsymbol{x}_t = \boldsymbol{\mu} + \boldsymbol{U} \operatorname{diag}\left(\left[\sqrt{\frac{\lambda_1 + t^2}{\lambda_1 + T^2}}, \ldots, \sqrt{\frac{\lambda_n + t^2}{\lambda_n + T^2}} \right] \right) \boldsymbol{U}^\top (\boldsymbol{x}_T - \boldsymbol{\mu}) \tag{32}$$

where $\boldsymbol{U}, \lambda_k, k \in [n]$ are singlar value decomposition of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \boldsymbol{U} \operatorname{diag}([\lambda_1, \ldots, \lambda_n]) \boldsymbol{U}^\top$. $\operatorname{diag}(\cdot)$ convert a vector in $\mathbb{R}^n$ into diagonal matrix $\mathbb{R}^{n \times n}$, and $\boldsymbol{x}_T \sim \mathcal{N}\left(\boldsymbol{0}, T^2 \boldsymbol{I}_n\right)$. Let $\boldsymbol{x}_T = T\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. As $t = 0$ and $T \to \infty$, we have:

$$\boldsymbol{x}_t = \left(\boldsymbol{I}_n - \boldsymbol{U} \operatorname{diag}\left(\left[\sqrt{\frac{\lambda_1 + t^2}{\lambda_1 + T^2}}, \ldots, \sqrt{\frac{\lambda_n + t^2}{\lambda_n + T^2}} \right] \right) \boldsymbol{U}^\top \right) \boldsymbol{\mu}, \tag{33}$$

$$+ \boldsymbol{U} \operatorname{diag}\left(\left[T\sqrt{\frac{\lambda_1 + t^2}{\lambda_1 + T^2}}, \ldots, T\sqrt{\frac{\lambda_n + t^2}{\lambda_n + T^2}} \right] \right) \boldsymbol{U}^\top \boldsymbol{x}_T, \tag{34}$$

$$= \boldsymbol{\mu} + \boldsymbol{U} \operatorname{diag}\left(\left[\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n} \right] \right) \boldsymbol{U}^\top \boldsymbol{x}_T, \tag{35}$$

$$= \boldsymbol{\mu} + \boldsymbol{\Sigma}^{1/2} \boldsymbol{x}_T = \boldsymbol{\Phi}(\boldsymbol{x}_T). \tag{36}$$

Thus, plugging in Definition 1, we have:

$$\texttt{PFD}(p,q) = \left(\mathbb{E}_{\boldsymbol{x}_T \sim \mathcal{N}(0, T^2 \boldsymbol{I})} \left[\left\| \boldsymbol{\Phi}_1(\boldsymbol{x}_T) - \boldsymbol{\Phi}_2(\boldsymbol{x}_T) \right\|_2^2 \right] \right)^{1/2} \tag{37}$$

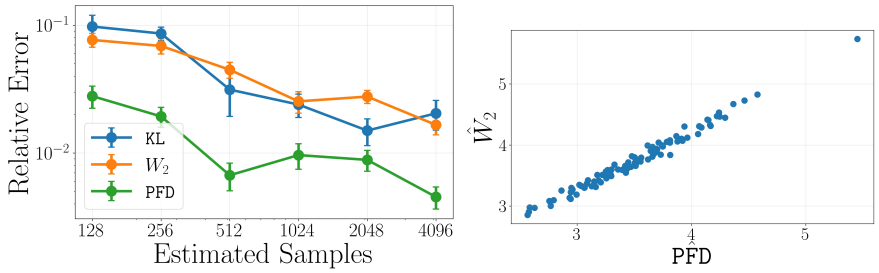
$$= \left(\mathbb{E}_{\boldsymbol{x}_T \sim \mathcal{N}(0, T^2 \boldsymbol{I})} \left[\left\| \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_1^{1/2} \boldsymbol{x}_T - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_2^{1/2} \boldsymbol{x}_T \right\|_2^2 \right] \right)^{1/2} \tag{38}$$

$$= \left(\left\| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right\|_2^2 + \left\| \boldsymbol{\Sigma}_1^{1/2} - \boldsymbol{\Sigma}_2^{1/2} \right\|_F^2 \right)^{1/2} \tag{39}$$

$$= \left(\left\| \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 \right\|_2^2 + \operatorname{Tr}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_2^{1/2} \right) \right)^{1/2} \tag{40}$$

From Wasserstein-2 distance for Gaussian distribution p, q has close form solution:

$$W_2(p,q) = \left(||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||_2^2 + \operatorname{Tr}\left(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2 - 2\left(\boldsymbol{\Sigma}_1^{1/2} \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1^{1/2} \right)^{1/2} \right) \right)^{1/2}. \tag{41}$$



(a) *(b)*

Figure 6. **Comparison of different metrics on synthetic datasets.** The figure illustrates (a) the sample efficiency of KL, W_2, and PFD under a Gaussian distribution, and (b) the correlation between W_2 and PFD under a mixture of Gaussians.

From Lemma 2, we have $W_2(p,q) \leq \mathtt{PFD}(p,q)$. And specifically, $W_2(p,q) = \mathtt{PFD}(p,q)$ when $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{\Sigma}_2\mathbf{\Sigma}_1$. □

Lemma 2. *Given two positive semi-definite matrix* $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2 \in \mathbb{R}^{n\times n}$,

$$0 \leq \mathrm{Tr}\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right) \leq \mathrm{Tr}\left(\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2}\right)^{1/2}\right). \tag{42}$$

Proof of Lemma 2. Because $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ are positive semi-definite matrix, $\mathrm{Tr}\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right) \geq 0$ and

$$\mathrm{Tr}\left(\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2\mathbf{\Sigma}_1^{1/2}\right)^{1/2}\right) = \mathrm{Tr}\left(\sqrt{\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right)\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right)^{\top}}\right) = \left\|\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right\|_*, \tag{43}$$

where $||\cdot||_*$ is the nuclear norm (trace norm). From trace norm inequality ((Bhatia, 2013) Chapter IV, Section 2), for a random matrix $\boldsymbol{M}$, $\mathrm{Tr}(\boldsymbol{M}) \leq \|\boldsymbol{M}\|_*$. Thus, we have:

$$\mathrm{Tr}\left(\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right) \leq \left\|\mathbf{\Sigma}_1^{1/2}\mathbf{\Sigma}_2^{1/2}\right\|_*. \tag{44}$$

□

C. Experiments

In this section, we provide experimental details and additional discussion of the main results.

C.1. Comparison of different metrics on synthetic datasets.

To support the claim that PFD is both sample-efficient and a meaningful distributional metric, we conduct numerical experiments on synthetic datasets, comparing PFD with the KL divergence (KL) and the 2-Wasserstein distance (W_2). Specifically, in Figure 6a, we evaluate the sample efficiency of these methods under multivariate Gaussian distributions. We set the dimension of the Gaussian distributions to 5, randomly generate their means and variances, and repeat the experiment 10 times. By varying the number of samples used for estimation from 128 to 4096, denoted by M, we report the relative error $|\mathcal{E} - \hat{\mathcal{E}}|/|\mathcal{E}|$, where $\mathcal{E}$ denotes the ground-truth value and $\hat{\mathcal{E}}$ its empirical estimate. In Figure 6b, we examine the correlation between estimates of W_2 and PFD on mixtures of Gaussian distributions. We consider mixtures of 5 Gaussian components, each with 5 dimensions. Both metrics are estimated using $M = 4096$ samples, and the experiment is repeated 100 times to generate the plotted results.

As shown in Figure 6a, PFD exhibits significantly better sampling efficiency than both KL and W_2. With $M = 4096$ samples, PFD achieves a relative error of approximately 4×10^{-2}, whereas KL and W_2 incur errors on the order of 2×10^{-1}. Moreover, the computational complexity of PFD is $O(M)$, compared to $O(M^2)$ for W_2. Consequently, under the same sampling budget, PFD is substantially more efficient in both estimation accuracy and computational cost.

As illustrated in Figure 6b, when measuring the distance between two mixture-of-Gaussian distributions, the estimated $\hat{\mathtt{PFD}}$ exhibits a strong linear correlation with $\hat{W}_2$ (with correlation coefficient 0.992). This result indicates that PFD

Table 2. **U-Net architectures details.**

Name	Dimensions for encoder blocks	Number of residual blocks	Number of parameters $\lvert\boldsymbol{\theta}\rvert$
U-Net-1	[32, 32, 32]	4	0.9M
U-Net-2	[64, 64, 64]	4	3.5M
U-Net-3	[96, 96, 96]	4	7.9M
U-Net-4	[128, 128, 128]	4	14.0M
U-Net-5	[80, 160, 160]	4	17.1M
U-Net-6	[160, 160, 160]	3	17.8M
U-Net-7	[160, 160, 160]	4	21.8M
U-Net-8	[192, 192, 192]	4	31.3M
U-Net-9	[224, 224, 224]	4	42.7M
U-Net-10	[256, 256, 256]	4	55.7M

captures meaningful distributional distance, which align with 2-Wasserstein distance, even for more complex, multimodal distributions.

C.2. Network Architecture Details

In this subsection, we provide details of the U-Net architectures, as summarized in Table 2. The U-Net follows an encoder-decoder design, where the encoder comprises multiple encoder blocks. The column "**Dimensions for encoder blocks**" indicates the feature dimensions of each encoder block, while "**Number of residual blocks**" specifies how many residual blocks are used within each encoder block. The decoder is symmetric to the encoder. For further architectural details, please refer to (Karras et al., 2022b). By varying the encoder block dimensions and the number of residual blocks, we scale the U-Net model from 0.9M to 55.7M parameters.

C.3. Evaluation Protocol

In this subsection, we provide details of the evaluation protocol introduced in Section 3, as well as the comparison between the synthetic dataset from the teacher model and the real dataset.

Experiment settings for evaluation protocol. The teacher model $\boldsymbol{\theta}_t$ and the student model $\boldsymbol{\theta}$ share a similar U-Net architecture (Ronneberger et al., 2015) with different numbers of parameters, as introduced in Section C.2. The teacher model, with UNet-10 architecture, is trained on the CIFAR-10 dataset (Krizhevsky et al., 2009) using the EDM noise scheduler (Karras et al., 2022a), with a batch size of 128 for 1,000 epochs. The student model [1] is trained using the variance-preserving (VP) noise scheduler (Ho et al., 2020), under the same training hyperparameters. We use one A40 GPU with 48 GB video random access memory (VRAM) for all experiments. We generated three subsets of initial noise $\{\boldsymbol{x}^{(i)}_{\texttt{train},T}\}_{i=1}^N, \{\boldsymbol{x}^{(i)}_{\texttt{gen},T}\}_{i=1}^M, \{\boldsymbol{x}^{(i)}_{\texttt{test},T}\}_{i=1}^M \overset{\text{iid}}{\sim} \mathcal{N}(0, T^2\boldsymbol{I}_n)$. The training and test datasets are produced using the teacher model:

$$\mathcal{D} \coloneqq \{\boldsymbol{x}^{(i)}_{\texttt{train}}\}_{i=1}^N = \{\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}_t}}(\boldsymbol{x}^{(i)}_{\texttt{train},T})\}_{i=1}^N, \;\; \mathcal{D}_{\texttt{test}} \coloneqq \{\boldsymbol{x}^{(i)}_{\texttt{test}}\}_{i=1}^M = \{\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}_t}}(\boldsymbol{x}^{(i)}_{\texttt{test},T})\}_{i=1}^M.$$

To evaluate the student model, we generate an evaluation dataset from itself:

$$\mathcal{D}_{\texttt{gen}} \coloneqq \{\boldsymbol{x}^{(i)}_{\texttt{gen}}\}_{i=1}^M = \{\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}^{(i)}_{\texttt{gen},T})\}_{i=1}^M.$$

All samples are generated using the second-order Heun solver (Karras et al., 2022a) with 18 sampling steps. We vary the number of training samples N from 2^6 to 2^{16} in powers of two. M is set to 50,000 for the experiments in Section C.4, and 10,000 for the rest.

Experiment settings for validating the synthetic dataset with real real-world dataset. We evaluate `FID` and $\mathcal{E}_{\texttt{mem}}$ for diffusion models with UNet-4 architecture, trained separately on the synthetic dataset $\mathcal{D}$ and CIFAR-10 training dataset. We keep the number of training dataset N the same for these two settings, ranging from 2^6 to 2^{15}, with a power of 2. Then we

[1] The architecture of the student model varies across experiments and will be described in detail for each specific case.

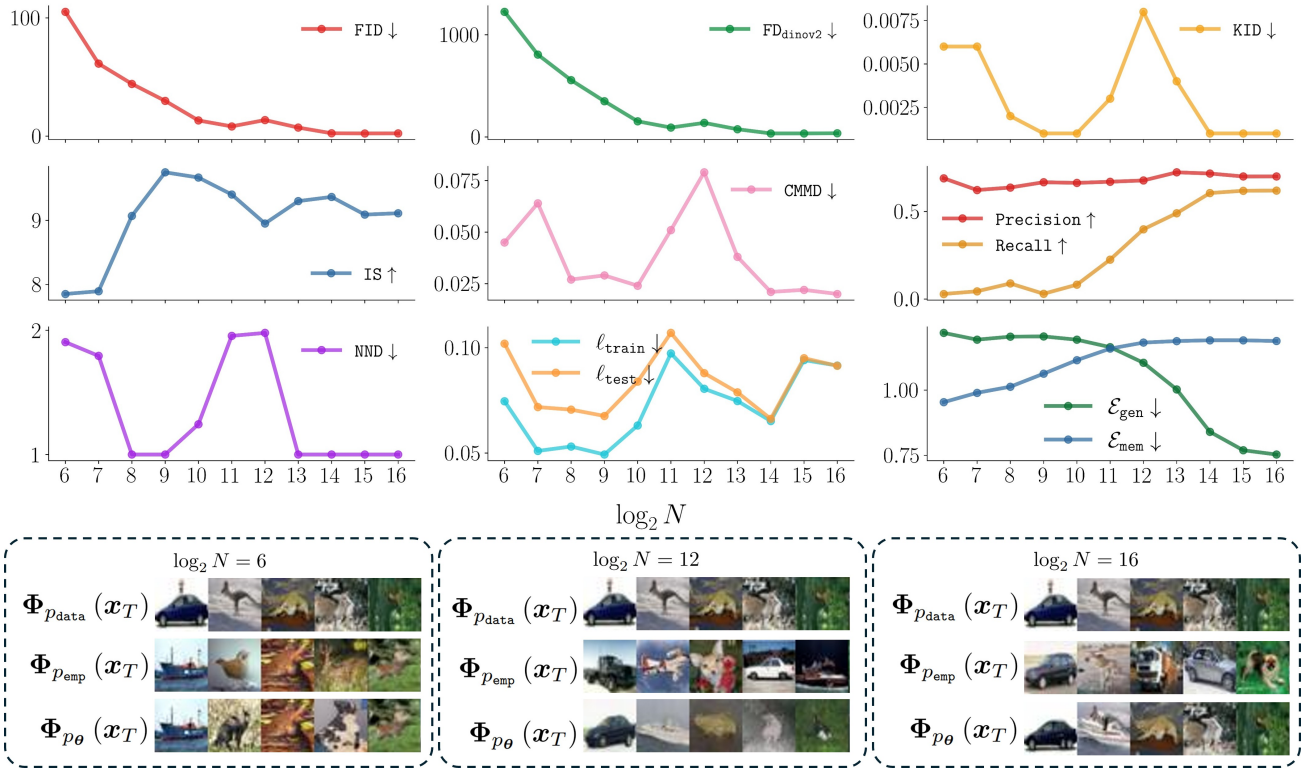


Figure 7. **Comparison of practical metrics on the MtoG transition.** The top figure plots multiple evaluation metrics as functions of $\log_2 N$. The bottom figure visualizes the generation under three numbers of training samples ($2^6, 2^{12}, 2^{16}$). For each setting, the figure shows generations from the underlying distribution (top row), empirical data distribution (middle row), and the learned distribution from the diffusion model (bottom row). Each column corresponds to the same initial noise.

evaluate the FID between $\mathcal{D}_{\texttt{gen}}$ and $\mathcal{D}_{\texttt{test}}$ (CIFAR-10 test dataset) for the synthetic (real-world) setting, with $M = 10000$. To evaluate $\mathcal{E}_{\texttt{mem}}$, we use the initial noise $\{\boldsymbol{x}_{\texttt{gen}}^{(i)}\}_{i=1}^{M}$.

C.4. Comparison with Practical Metrics for Generalization Evaluation

In this subsection, we expand upon the experiment presented in Section 3, which compares our proposed metric with practical metrics for evaluating generalization. We compare $\mathcal{E}_{\texttt{gen}}$ and $\mathcal{E}_{\texttt{mem}}$ with well-used generative model metrics, inluding FID, $\texttt{FD}_{\texttt{DINOv2}}$, KID, CMMD, Precision, Recall, NND, IS. We also including the training and testing loss $\ell_{\texttt{train}}, \ell_{\texttt{test}}$ (Equation (10)) as comparison. We evaluating their ability in capturing the MtoG transition, under the evaluation protocol proposed in Section 3.

Metric	**Dataset(s)**
FID, $\texttt{FD}_{\texttt{DINOv2}}$, KID, CMMD, Precision, Recall, NND	$\mathcal{D}_{\texttt{gen}}$ vs. $\mathcal{D}_{\texttt{test}}$
$\texttt{FID}_{\texttt{train}}$, $\texttt{FD}_{\texttt{DINOv2,train}}$	$\mathcal{D}$ vs. $\mathcal{D}_{\texttt{test}}$
IS	$\mathcal{D}_{\texttt{gen}}$
$\ell_{\texttt{train}}$	$\mathcal{D}$
$\ell_{\texttt{test}}$	$\mathcal{D}_{\texttt{test}}$
$\mathcal{E}_{\texttt{mem}}, \mathcal{E}_{\texttt{gen}}$	$\{\boldsymbol{x}_{\texttt{gen},T}^{(i)}\}_{i=1}^{M}$

Table 3. Datasets used to evaluate each metric.

We use UNet-10 for the student model in this experiment. We summarized datasets used by these metrics in Table 3. Results are shown in Figure 7, summarized into one sentence, only $\mathcal{E}_{\texttt{gen}}$ and $\mathcal{E}_{\texttt{mem}}$ could quantitatively capture this transition. We

include detailed discussions below:

Results discussions. Figure 7 (bottom) is consistent with prior empirical observations (Zhang et al., 2024; Yoon et al., 2023): In the memorization regimes ($N = 2^6$), $p_{\boldsymbol{\theta}}$ tends to memorize the empirical distribution $p_{\texttt{emp}}$, resulting in similar generation between $\boldsymbol{\Phi}_{p_{\texttt{emp}}}(\boldsymbol{x}_T)$ and $\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)$; in the transition regime ($N = 2^{12}$), the model lacks sufficient capacity to memorize and the sample complexity is inadequate for generalization, leading to poor-quality generations $\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)$; in the generalization regimes ($N = 2^{16}$), $p_{\boldsymbol{\theta}}$ captures the underlying distribution $p_{\texttt{data}}$, and the generations $\boldsymbol{\Phi}_{p_{\texttt{data}}}(\boldsymbol{x}_T)$ and $\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)$ are closely aligned.

As shown in Figure 7 (top), when N increases, $\mathcal{E}_{\texttt{mem}}$ consistently increases and $\mathcal{E}_{\texttt{gen}}$ consistently decreases. This aligns with our intuition: as sample complexity grows, models tend to generalize and memorize less. In contrast, all other metrics fail to capture this transition effectively. The reasons can be summarized as follows:

- **FID, $\text{FD}_{\texttt{DINOv2}}$, KID, IS, and CMMD are sensitive to generation quality.** Image quality metrics, including FID, $\text{FD}_{\texttt{DINOv2}}$, KID, IS, and CMMD, show degradation in performance at $N = 2^{12}$. This drop is primarily due to degraded visual quality in the generated samples, as visualize in Figure 7 (bottom-middle). However, at this sample complexity, the generated data still captures low-level features such as colors and structures from the underlying distribution. This is evident from the visual similarity between $\boldsymbol{\Phi}_{p_{\texttt{data}}}(\boldsymbol{x}_T)$ and $\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}(\boldsymbol{x}_T)$, suggesting the model have some generalizability. In comparison, only $\mathcal{E}_{\texttt{gen}}$ decreases consistently around $N = 2^{12}$, indicating it captures generalizability better than others despite visual degradation.

- **FID, $\text{FD}_{\texttt{DINOv2}}$ and Recall are sensitive to diversity.** The monotonic trends for FID, $\text{FD}_{\texttt{DINOv2}}$ and Recall are due to their sensitivity to the diversity of $\mathcal{D}_{\texttt{gen}}$, rather than their ability to measure generalizability. At small N, the model memorizes the training samples, resulting in $\mathcal{D}_{\texttt{gen}}$ closely resembling $\mathcal{D}$ and exhibiting significantly lower diversity than $\mathcal{D}_{\texttt{test}}$, since $N \ll M$. Under these conditions, FID, $\text{FD}_{\texttt{DINOv2}}$ are large because they are biased towards the diversity of the evaluation samples (as proved in (Chong & Forsyth, 2020)). Meanwhile, Recall is low because the the support of $\mathcal{D}_{\texttt{test}}$ is limited, reducing the probability that samples drawn from $\mathcal{D}_{\texttt{gen}}$ lie within the support of $\mathcal{D}_{\texttt{test}}$. In contrast, $\mathcal{E}_{\texttt{gen}}$ measures generalizability by directly quantifying the distance between the generation from the learned distribution and the underlying distribution and is less affected by the diversity of the generated samples.

- **NND and ℓ fail to capture the generalizability.** The NND, originally designed for assessing the generalization of GANs, is sensitive to image quality and increases during the transition regime. Additionally, it produces identical values across a wide range of sample sizes (e.g., $N = 2^8, 2^9, 2^{13}, 2^{14}, 2^{15}, 2^{16}$), making it unreliable for evaluating generalization in diffusion models. Similarly, neither the training loss $\ell_{\texttt{train}}$ nor the test loss $\ell_{\texttt{test}}$ exhibits a consistent decreasing trend as N increases, indicating that these losses do not directly reflect either memorization or generalization. While the loss gap $\ell_{\texttt{test}} - \ell_{\texttt{train}}$ does tend to decrease with larger N, it cannot serve as a robust generalization metric either. This is because even a randomly initialized model $\boldsymbol{\theta}$ can exhibit a small loss gap.

In conclusion, $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ are the only metrics that could capture the MtoG transition for diffusion models. They evaluate the generalization (memorization) by directly measuring the distance between the learned distribution by the diffusion model and the underlying (empirical) distribution. Unlike other metrics, they are less affected by the quality or diversity of the evaluating samples.

C.5. Scaling Behaviors of the MtoG Transition

In this subsection, we provide detailed experimental settings for Section 4.1, along with additional experiments to further investigate the MtoG transition across more architectures (e.g., Transformer-based models (Bao et al., 2023)). We also investigate the scaling behavior of the MtoG transition under the DINOv2 descriptor.

Experiment settings. The detailed architectures of the student models, from U-Net-1 to U-Net-10, are provided in Section C.2, with model sizes ranging from 0.9M to 55.7M parameters. We scale up the architectures by increasing the dimensionality of the encoder blocks and the number of residual blocks. For the ImageNet experiments, we adopt the U-Net architectures proposed in (Karras et al., 2024), referred to as U-Net-11 and U-Net-12. These models contain 124.2M and 295.9M parameters, respectively.

MtoG transition between U-Net and transformer architecture. To further investigate the impact of network architecture, we compare the U-Net architecture with the transformer-based UViT (Bao et al., 2023). Specifically, we use the U-Net-9

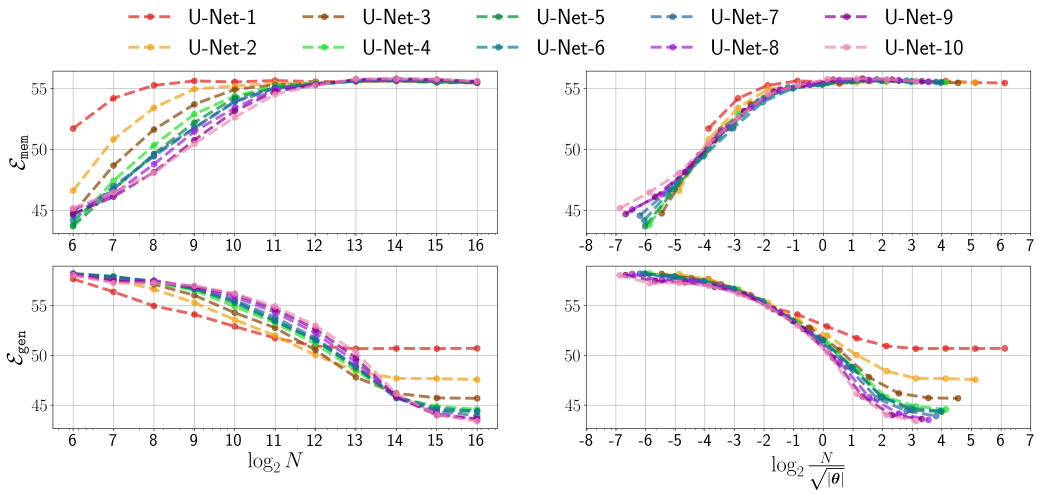


Figure 8. **Scaling behavior in the MtoG transition under DINOv2 descriptor.** $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ plotted against $\log_2(N)$ for a range of U-Net architectures (U-Net-1 to U-Net-10). Right: the same metrics plotted against $\log_2(N/\sqrt{|\boldsymbol{\theta}|})$, where $|\boldsymbol{\theta}|$ is the number of model parameters.

from Table 2, containing 42.7M parameters, and design the UViT model with comparable parameters of 44.2M. Both models are trained for 1000 epochs. Using the same experimental setup described in Section 4.1, we plot the MtoG transition curves for both U-Net and UViT, as shown in Figure 9.

As illustrated in Figure 9, with a similar number of parameters and the same training data sizes, UViT exhibits a higher $\mathcal{E}_{\texttt{mem}}$ in the memorization regime ($2^6 \leq N \leq 2^{10}$) and a higher $\mathcal{E}_{\texttt{gen}}$ in the generalization regime ($2^{11} \leq N \leq 2^{15}$), suggesting a lower model capacity compared to U-Net under these conditions. However, when provided with sufficient training data ($N = 2^{16}$), UViT achieves a lower $\mathcal{E}_{\texttt{gen}}$, demonstrating better generalization performance. This observation is consistent with prior findings on transformer architectures in classification tasks: transformer-based models, lacking the inductive biases inherent to CNNs, tend to generalize poorly when trained on limited data (Dosovitskiy et al., 2021).

Scaling behavior of the MtoG transition under the DINOv2 descriptor. The scaling behavior under the DINOv2 descriptor is shown in Figure 8. Both $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ exhibit trends consistent with those observed under the SSCD descriptor (see Figure 2). The only difference is that, under the DINOv2 descriptor, models with varying parameter sizes show greater differentiation in the generalization regime compared to those under the SSCD descriptor. Further discussion on this can be found in the ablation study on image descriptors in Section E.2.

C.6. Early Learning and Double Descent in Learning Dynamics

In this subsection, we build on the discussion from Section 4.2. In Figure 4, we evaluate $\ell_{\texttt{train}}$ and $\ell_{\texttt{test}}$ across the three training regimes. Notably, the gap $\ell_{\texttt{test}} - \ell_{\texttt{train}}$ emerges as a practical heuristic for identifying the training regime: In the memorization regime, the gap increases steadily with training; In the transition regime, the gap remains near zero during early training (when generalization improves) and increases for further training (when generalization degrades); in the generalization regime, the gap remains close to zero throughout training. While $\ell_{\texttt{test}} - \ell_{\texttt{train}}$ is not a strict measure of generalization, it proves to be a useful empirical indicator of training regimes for diffusion models. Practically, by setting aside a test dataset to estimate this gap, we can more effectively identify the training regime for diffusion models.

Training dynamics of diffusion models under the DINOv2 descriptor. The training dynamics under the DINOv2 descriptor are shown in Figure 10. Both $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ exhibit trends consistent with those observed under the SSCD descriptor for $N = 64$ and $N = 4096$ (see Figure 4). For $N = 65536$, $\mathcal{E}_{\texttt{gen}}$ still displays a double descent pattern under the DINOv2 descriptor; however, instead of a rise between the two drops, the curve remains relatively flat.

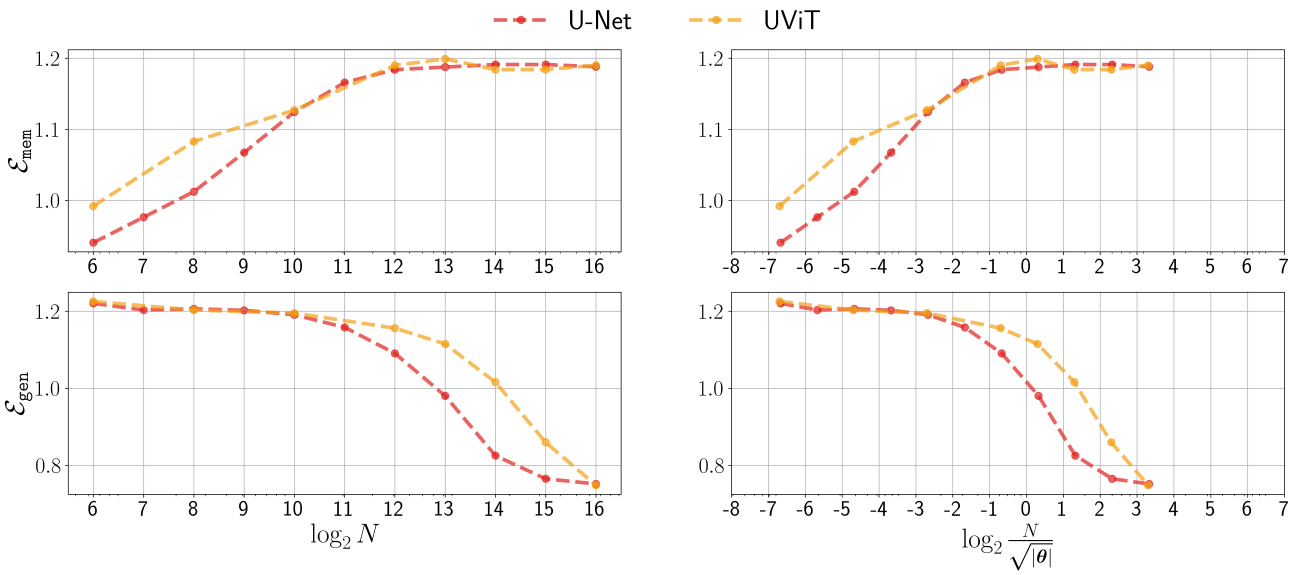


Figure 9. **Comparison of scaling behavior between UNet and Transformer architectures in the MtoG transition.** $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ plotted against $\log_2(N)$ for U-Net architecture (U-Net-9) and UViT architecture. Right: the same metrics plotted against $\log_2(N/\sqrt{|\boldsymbol{\theta}|})$, where $|\boldsymbol{\theta}|$ is the number of model parameters.

C.7. Bias-Variance Decomposition of Generalization Error

To approximate $\overline{\boldsymbol{\Psi} \circ \boldsymbol{\Phi}}_{p_{\boldsymbol{\theta}}}(\cdot)$, we independently sample two training datasets, $\mathcal{D}_1$ and $\mathcal{D}_2$, for each specified number of training samples N. We then train two student models, $\boldsymbol{\theta}(\mathcal{D}_1)$ and $\boldsymbol{\theta}(\mathcal{D}_2)$, using these datasets. The quantity $\overline{\boldsymbol{\Psi} \circ \boldsymbol{\Phi}}_{p_{\boldsymbol{\theta}}}(\cdot)$ is approximated as follows:

$$\overline{\boldsymbol{\Psi} \circ \boldsymbol{\Phi}}_{p_{\boldsymbol{\theta}}}(\cdot) \approx \frac{1}{2}(\boldsymbol{\Psi} \circ \boldsymbol{\Phi}_{p_{\boldsymbol{\theta}(\mathcal{D}_1)}}(\cdot) + \boldsymbol{\Psi} \circ \boldsymbol{\Phi}_{p_{\boldsymbol{\theta}(\mathcal{D}_2)}}(\cdot)). \tag{45}$$

Bias-Variance Decomposition of Generalization Error under the DINOv2 Descriptor. The bias-variance decomposition under the DINOv2 descriptor is shown in Figure 11. Overall, the results are consistent with those observed under the SSCD descriptor, with two differences: (1) for $N = 65536$, $\mathcal{E}_{\texttt{gen}}$ does not exhibit a U-shaped curve under the DINOv2 descriptor; and (2) $\mathcal{E}_{\texttt{bias}}$ for U-Net-1 and U-Net-2 does not decrease monotonically, instead, it first decreases and then increases.

D. Further Discussions of $\mathcal{E}_{\texttt{mem}}$

In this section, we present the mathematical formulation for estimating $\mathcal{E}_{\texttt{mem}}$ and compare it with the existing memorization metric.

Empirically estimate $\mathcal{E}_{\texttt{mem}}$. As described in Definition 1 and Definition 2, estimating $\mathcal{E}_{\texttt{mem}}$ requires access to the mapping $\boldsymbol{\Phi}_{p_{\texttt{emp}}}(\cdot)$. According to Equation (2), this mapping is determined by the score function of the empirical distribution, denoted as $\nabla \log \hat{p}_t(\boldsymbol{x}_t)$. Based on prior works (Karras et al., 2022a; Zhang et al., 2024; Gu et al., 2023), the score function of the empirical distribution has a closed-form expression:

$$\nabla \log \hat{p}_t(\boldsymbol{x}_t) = \frac{1}{T^2}\left(\frac{\mathbb{E}_{\boldsymbol{x}\sim p_{\texttt{emp}}}[\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}, T^2\boldsymbol{I}_n)\cdot \boldsymbol{x}]}{\mathbb{E}_{\boldsymbol{x}\sim p_{\texttt{emp}}}[\mathcal{N}(\boldsymbol{x}_t; \boldsymbol{x}, T^2\boldsymbol{I}_n)]} - \boldsymbol{x}_t\right), \tag{46}$$

where $p_{\texttt{emp}}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\boldsymbol{x} - \boldsymbol{y}^{(i)})$ corresponds to the empirical distribution over the training dataset ${\boldsymbol{y}^{(i)}}_{i=1}^{N}$. This formulation allows us to numerically compute $\nabla \log \hat{p}_t(\boldsymbol{x}_t)$ for any given t. Subsequently, we can use a numerical solver to estimate the integral in Equation (2), thereby enabling the estimation of $\mathcal{E}_{\texttt{mem}}$.

Comparison between existing memorization metric and $\mathcal{E}_{\texttt{mem}}$. Previous works (Yoon et al., 2023; Zhang et al., 2024)

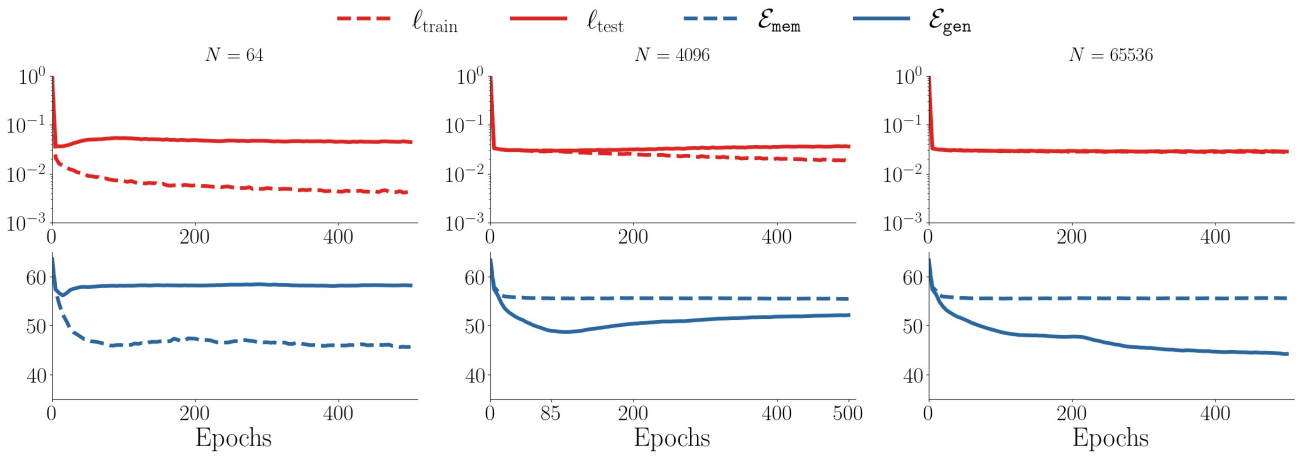


Figure 10. **Training dynamics of diffusion models under DINOv2 descriptor in different regimes.** The figure plots $\mathcal{E}_{\texttt{mem}}, \mathcal{E}_{\texttt{gen}}, \ell_{\texttt{train}}, \ell_{\texttt{test}}$ over training epochs for different different dataset sizes: $N = 2^6$ (left), 2^{12} (middle), 2^{16} (right).

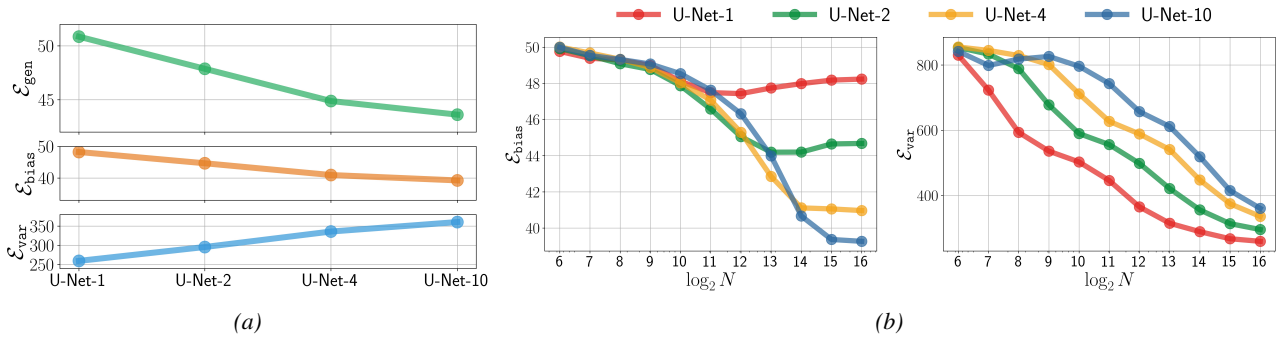


Figure 11. **Bias–Variance Trade-off under DINOv2 descriptor.** (a) plots the generalization error $\mathcal{E}_{\texttt{gen}}$, bias $\mathcal{E}_{\texttt{bias}}$, and variance $\mathcal{E}_{\texttt{var}}$ across different network architectures with a fixed training sample size of $N = 2^{16}$. (b) shows $\mathcal{E}_{\texttt{bias}}$ and $\mathcal{E}_{\texttt{var}}$ as functions of the number of training samples N for various network architectures.

define memorization metirc as:

$$\texttt{M Distance}\left(p_{\boldsymbol{\theta}}\right) \coloneqq \mathbb{E}_{\boldsymbol{x}_T}\left[\min_{\boldsymbol{x}\sim p_{\texttt{emp}}} \left\|\boldsymbol{\Psi}\left(\boldsymbol{x}\right) - \boldsymbol{\Psi} \circ \boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}\left(\boldsymbol{x}_T\right)\right\|_2\right], \tag{47}$$

A generated sample $\boldsymbol{\Phi}_{p_{\boldsymbol{\theta}}}\left(\boldsymbol{x}_T\right)$ is a memorized sample if it is close enough to one of the sample $\boldsymbol{x}$ from $p_{\texttt{emp}}$. It is easy to show that $\mathcal{E}_{\texttt{mem}}$ is a more strict metric than `M Distance`, i.e. "$\mathcal{E}_{\texttt{mem}}\left(p^{\boldsymbol{\theta}}\right) = 0$" is a sufficient but not necessary condition for "`M Distance` $\left(p^{\boldsymbol{\theta}}\right) = 0$". We propose $\mathcal{E}_{\texttt{mem}}$ in order to unify the definition of memorization and generalization.

E. Ablation Study

In this section, we present ablation studies on the evaluation protocol, examining the effects of different noise schedulers and sampling methods (Section E.1), image descriptors (Section E.2), sample sizes for evaluation (Section E.3), and teacher models (Section E.4).

E.1. Sampling Methods

In this subsection, we present ablation studies on various noise schedulers and sampling strategies. Specifically, we evaluate the performance of the following methods: Variance Preserving (VP) (Song et al., 2021c), Variance Exploding (VE) (Song et al., 2021c), iDDPM (Nichol & Dhariwal, 2021) + DDIM (Song et al., 2021a), and EDM (Karras et al., 2022a). The specific form of $f(t), g(t)$ used in each approach are detailed in Table 1 of (Karras et al., 2022a). Additionally, each method

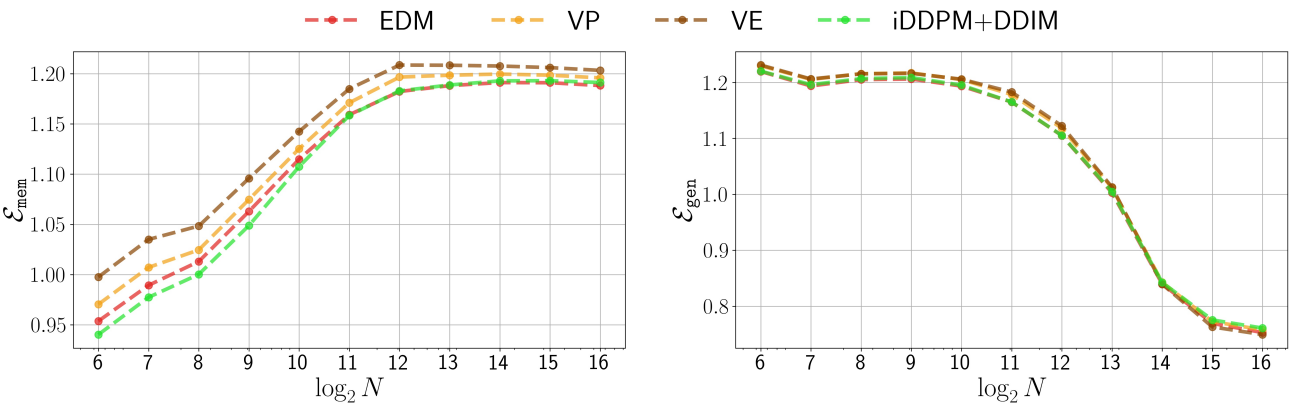


Figure 12. **Comparison of different sampling methods.** $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ plotted against $\log_2(N)$ for different sampling methods, including: EDM, VP, VE, iDDPM+DDIM.

also differs in its choice of ODE solver and timestep discretization strategy. For sampling, we use 256 steps for VP, 1000 for VE, 100 for iDDPM + DDIM, and 18 for EDM. All experiments are conducted under the evaluation protocol described in Section 3, where we estimate the $\mathcal{E}_{\texttt{gen}}$ under different training samples N. The student models use the UNet-10 architecture. During the ablation study, both the teacher and student models use the same sampling method[2] as specified above.

As shown in Figure 12, different samplers yield highly consistent results, demonstrating that PFD can be extended to various noise schedules, i.e., different choices of $f(t)$ and $g(t)$.

E.2. Image Descriptors

In this subsection, we present ablation studies on the image descriptor $\mathbf{\Psi}$ used in Equation (3). The descriptors evaluated include DINOv2 (Oquab et al., 2024), InceptionV3 (Szegedy et al., 2016), CLIP (Radford et al., 2021), SSCD (Pizzi et al., 2022), and the identity function. All experiments follow the evaluation protocol described in Section 3, where we estimate both $\mathcal{E}_{\texttt{mem}}$ and $\mathcal{E}_{\texttt{gen}}$ across varying training sample sizes N and different student model architectures: U-Net-1, U-Net-2, U-Net-4, and U-Net-10.

As shown in Figure 13, different feature embeddings reveal a consistent trend in the memorization-to-generalization (MtoG) transition across various U-Net architectures. With limited training samples, smaller models exhibit lower generalization scores. Conversely, with sufficient training data, larger models tend to have lower generalization scores. When comparing with $\mathcal{E}_{\texttt{gen}}$ measured in pixel space (i.e., using the identity function as the descriptor), we observe that $\mathcal{E}_{\texttt{gen}}$ values are nearly identical across diffusion architectures when $N \geq 2^{15}$. In this regime, all models have learned low-level image features such as color and structure; however, only the larger models capture high-level perceptual details. Because pixel-space measurements fail to reflect these high-level features, they yield similar $\mathcal{E}_{\texttt{gen}}$ values regardless of model size. Therefore, it is better to evaluate $\mathcal{E}_{\texttt{gen}}$ in a feature space, which better captures perceptual differences between models.

Different feature descriptors mainly differ in the generalization regime. Specifically, $\mathcal{E}_{\texttt{gen}}$ varies the most across architectures when using the DINOv2 descriptor, and the least when using the SSCD descriptor. This is because each descriptor capture different aspects of the image. SSCD focuses on detecting duplicate content and is more sensitive to low-frequency features, while DINOv2 emphasizes perceptual quality and captures high-frequency features. Diffusion models with limited capacity tend to learn low-frequency information first, as it it more easier to learn (Wang, 2025). As a result, under the SSCD descriptor, different architectures show more similar $\mathcal{E}_{\texttt{gen}}$ values, since they are all primarily capturing the same low-frequency information in the early training stages.

E.3. Evaluation Sample Number

In this subsection, we present ablation studies on the number of samples M used by $\hat{\texttt{PFD}}$ to approximate PFD, as defined in Equation (4). All experiments follow the evaluation protocol described in Section 3, where we estimate $\mathcal{E}_{\texttt{gen}}$ across varying

[2]Note that the noise scheduler used for sampling could differ from that used during training.

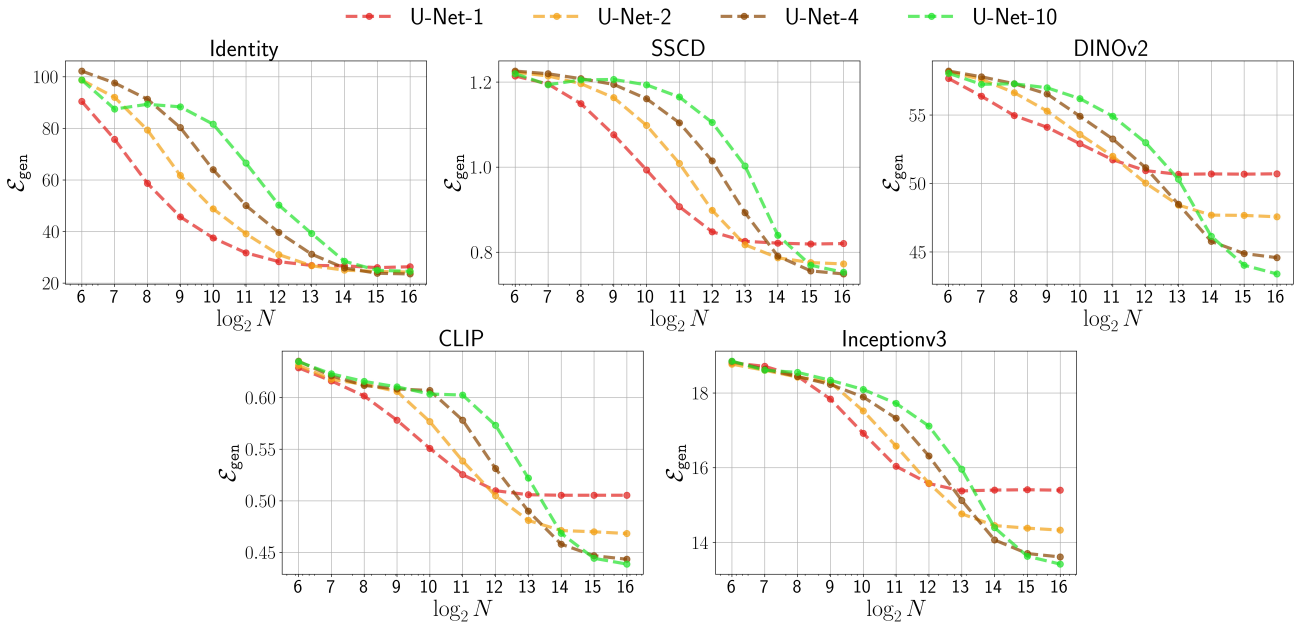


Figure 13. **Comparison between different image descriptors.** $\mathcal{E}_{\text{gen}}$ plotted against $\log_2(N)$ for a range of U-Net architectures (U-Net-1, U-Net-2, U-Net-4, U-Net-10) using different image descriptors, including identity function, SSCD, DINOv2, CLIP, Inceptionv3.

training sample sizes N and different student model architectures: U-Net-1, U-Net-2, U-Net-4, and U-Net-10. We vary $M \in \{10, 32, 100, 316, 1000, 3163, 10000\}$, and for each setting, generate 5 independent sets of $\{\boldsymbol{x}^{(i)}_{\text{gen},T}\}_{i=1}^{M}$ initial noise estimate $\mathcal{E}_{\text{gen}}$, computing both the mean and variance.

As shown in Figure 14, the variance of $\mathcal{E}_{\text{gen}}$ approaches zero as M increases to 10,000, indicating that when $M \geq 10000$, the empirical estimate of $\mathcal{E}_{\text{gen}}$ converges to its value over the underlying distribution. This result holds consistently across different model architectures.

E.4. Teacher Model Architecture

We end this section by examining how different teacher models affect the evaluation protocol. Specifically, we consider three types of diffusion models: EDM, Rectified Flow (Rect) (Liu et al., 2023), and UViT. Using the CIFAR-10 dataset, we train three teacher models, one for each of these diffusion types. For each teacher model, we then evaluate all three diffusion models as student models. We report their corresponding $\mathcal{E}_{\text{gen}}$ values. Both teacher and student models use the same sampling method, the second-order Heun solver with 18 steps.

As shown in Figure 15, the $\mathcal{E}_{\text{gen}}$ is approximately 0.7 when both the student and teacher models are selected from EDM or UViT. However, $\mathcal{E}_{\text{gen}}$ increases to around 0.8 when either the student or teacher model is Rect. According to its original paper, Rect has the poorest generation quality among the three, as measured by `FID`. This suggests that the teacher model should possess strong generative performance to serve as an underlying distribution that is close to the real-world data distribution. Therefore, in this paper, we adopt EDM as the teacher model, as it achieves the lowest `FID` among the three models.

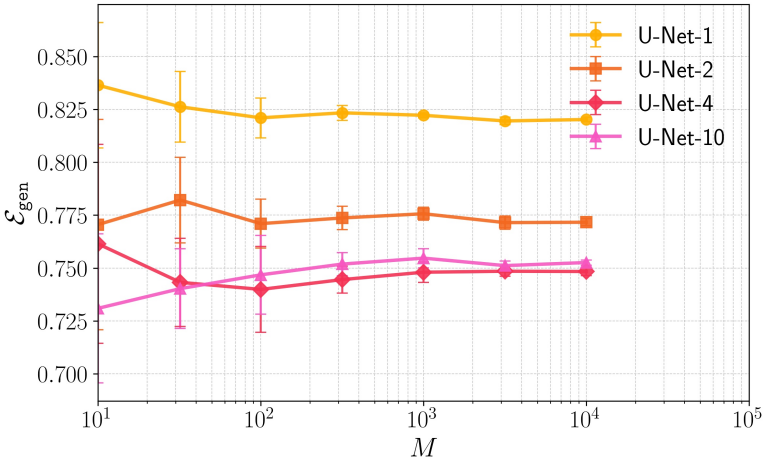


Figure 14. **Comparison across evaluation sample sizes.** The mean and variance of $\mathcal{E}_{\text{gen}}$ are plotted against the number of evaluation samples M for various U-Net architectures (U-Net-1, U-Net-2, U-Net-4, U-Net-10), with a fixed number of training samples $N = 2^{16}$.

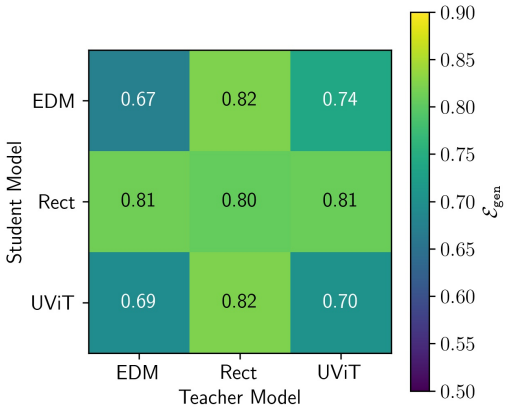


Figure 15. **Comparison of different teacher models.**The figure shows the $\mathcal{E}_{\text{gen}}$ values for various student models (EDM, Rect, UViT) trained using different teacher models (EDM, Rect, UViT), with a fixed training data size of $N = 2^{16}$.