TRAINING FLOW-BASED GENERATIVE MODELS WITH ADAPTIVE CURRICULUM SAMPLER

Anonymous authors
Paper under double-blind review

ABSTRACT

The training of state-of-the-art flow-based generative models can be viewed as learning to denoise latents across a continuum of timesteps, or noise levels. Despite their empirical success, the role of timestep sampling in training and the relationship between flow-based generative models and conventional denoising methods remain insufficiently understood in existing literature. To answer this question, in this paper we present a principled view that interprets noise level as a proxy for sample difficulty in curriculum learning from a denoising perspective, thereby providing a natural basis for organizing samples with different noise levels into an easy-to-hard curricula. We show both theoretically and empirically that noise level provides a consistent measure for this curriculum-level difficulty, establishing it as a foundation for guiding the sampling distribution. This perspective further suggests that the timestep sampling should adapt to the model's evolving denoising ability during training. However, existing methods typically rely on entirely fixed or weakly adaptive strategies, limiting the generation performance. To address this issue, we further propose **CurFM**, a Curriculum-based adaptive sampler for flow-based generative models that progressively shifts the timestep distribution toward more challenging noise levels according to the model's current denoising capability, without relying on unstable loss signals. CurFM benefits in minimal modifications to existing flow-based training pipelines without changes during inference, while achieving faster convergence as well as competitive or superior performance under both v- and ϵ-*prediction* frameworks.

1 INTRODUCTION

Modern generative models (Rombach et al., 2022; Wan et al., 2025) can be unified under the continuous normalizing flow (CNF) framework (Chen et al., 2018), with common parameterizations such as *v-prediction* (Lipman et al., 2023) and ϵ-*prediction* (Ho et al., 2020). Despite their empirical success, it remains insufficiently understood that i) why they are more effective than classical denoising models and ii) what fundamentally distinguishes them from conventional denoising approaches (Zhang et al., 2017). Recent works (Gagneux et al., 2025; Li & He, 2025) revisit flow-based training from a denoising perspective and present that, after reparameterization (Equation 5), the primary difference between flow-based optimization and standard denoising objective lies in a timestep-dependent weighting that emphasizes low-noise samples in training. This naturally raises a question: *why such straightforward reweighting yields superior performance, and why lower-noise levels should be prioritized even though higher-noise samples appear more challenging and may seem to warrant greater attention?*

In this paper, we investigate this question via interpreting noise level as a principled proxy for sample difficulty in curriculum learning (Bengio et al., 2009; Wang et al., 2021). During training, the model encounters samples across a continuum of noise levels, with highly corrupted samples intuitively more difficult to denoise, corresponding to greater curriculum-level difficulty. From this perspective, timestep sampling determines how training emphasis is distributed across samples with different levels of denoising difficulty. Consequently, emphasizing lower-noise (easier) samples provides more stable optimization signals, allowing the model to progressively develop denoising ability and better handle challenging high-noise samples. We experimentally show that training dominated by high-noise (more difficult) samples leads to slower convergence and inferior performance, whereas prioritizing easier samples in early stages improves stability and facilitates overall learning.

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

Moreover, this perspective naturally suggests that the optimal timestep distribution should be aligned with the model's denoising ability, which evolves throughout training. Despite the recognized importance of timestep sampling (Esser et al., 2024), existing generative models (Xiao et al., 2025; Yang et al., 2025) typically employ fixed timestep distributions throughout training, such as uniform (Lipman et al., 2023) or logit-normal (Esser et al., 2024). While some works explore more flexible distributions (Hang et al., 2025; Sahoo et al., 2024), these approaches remain static and fail to adapt to the evolving capabilities of the model, leading to suboptimal optimization. Recent efforts on noise scheduling (Kim et al., 2025b;a) largely rely on regression-based losses, which are known to exhibit high variance and instability (Xu et al., 2023; Jeha et al., 2024). As illustrated in Figure 1, even after the model achieves converged performance (*i.e.*, epoch 600) in terms of Fréchet Inception Distance (FID), the training loss continues to decrease and fluctuate significantly, making it an unreliable signal for guiding timestep sampling or assessing convergence. Therefore, the key challenge is to design an adaptive sampling strategy that tracks the model's evolving denoising ability while remaining stable without relying on noisy loss signals.

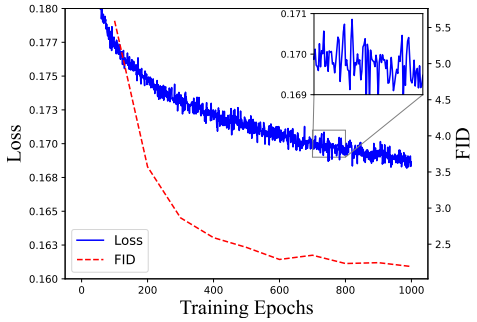


Figure 1: Comparison between loss and FID curves when training on CIFAR-10.

To tackle the challenge, we propose CurFM, a curriculum-based adaptive sampler for training flow-based generative models. The effectiveness of Curriculum Learning (Bengio et al., 2009) critically depends on reliable difficulty measures, which are particularly non-trivial to obtain in generative modeling due to highly stochastic training losses. Thus we present theoretical justification that noise level provides a principled measure of curriculum-level sample difficulty, thereby grounding our curriculum strategy. Building upon this, we introduce a scheduling mechanism that progressively exposes the model to more difficult noise levels based on its current denoising ability without relying on unstable loss signals, while mitigating catastrophic forgetting. Despite requiring i) merely minimal modifications to existing training pipelines and ii) no modification during inference, the proposed CurFM achieves faster convergence as well as comparable or superior performance to baseline samplers under both v- and ϵ-*prediction* frameworks. Our contributions are as follows:

- We present a principled view that interprets noise level as a proxy for sample difficulty in curriculum learning, providing insights into the effectiveness of flow-based generative training over conventional denoising methods and motivating adaptive timestep sampling.
- We theoretically justify noise level as a curriculum-level proxy for denoising difficulty and propose CurFM, a curriculum sampler that adapts to the model's evolving denoising ability.
- We empirically validate that the proposed CurFM achieves faster convergence as well as competitive or superior performance over baseline samplers across diverse settings.

2 Related Work

Flow-based Generative Models Modern generative models (Xiao et al., 2025; Blattmann et al., 2023b; Chen et al., 2024; Kong et al., 2024) are capable of synthesizing high-fidelity samples such as images. Early works (Rombach et al., 2022; Guo et al., 2023) adopt diffusion-based architectures (Ho et al., 2020; Song et al., 2020a), while more recent approaches (Esser et al., 2024; Wan et al., 2025) employ Flow Matching (Lipman et al., 2023; Liu et al., 2023). Nevertheless, both frameworks can be unified under CNF (Chen et al., 2018), where models learn flows that transform samples from a prior distribution into the target distribution using different prediction targets such as velocity or noise. Despite their strong empirical performance, the theoretical advantages of these frameworks over conventional denoising methods (Zhang et al., 2017) remain insufficiently studied.

Timestep Sampling in Flow-based Generative Models Several works (Gagneux et al., 2025; Li & He, 2025) reparameterize the flow objective to analyze its connection to conventional formulations, highlighting the importance of timestep sampling (Esser et al., 2024). However, existing flow-based generative models (Blattmann et al., 2023a; Labs et al., 2025) typically adopt fixed sampling distributions, such as uniform (Lipman et al., 2023) or logit-normal (Esser et al., 2024). More related studies explore alternative sampling strategies, including U-shaped (Lee et al., 2024) and Laplace

distribution (Hang et al., 2025). Other works investigate timestep-dependent weighting schemes, such as SNR-based weighting (Hang et al., 2023) and implicitly learned weighting conditioned on context (Sahoo et al., 2024). Nevertheless, these approaches remain static throughout training and fail to adapt to the model's evolving capability. Although recent works (Xu et al., 2026) explore dynamic noise-level scheduling, their effectiveness largely depends on regression-based losses, which can exhibit high variance and instability (Jeha et al., 2024). Consequently, existing methods remain suboptimal, relying predominantly on fixed or weakly adaptive timestep sampling and motivating a principled, more adaptive approach that accounts for the model's evolving denoising capability.

Curriculum Learning Curriculum Learning (CL) (Wang et al., 2021) is a training paradigm that follows an easy-to-hard progression, inspired by human learning processes. A key component of CL is the **difficulty measurer**, which estimates the difficulty of data, along with a **scheduler** that determines the ordering of these data during training. Representative approaches include Transfer Teacher (Hacohen & Weinshall, 2019), which employs a teacher model to assess sample difficulty, and Self-paced Learning (Kumar et al., 2010), which gradually expands the training set from easy to hard based on loss. CL has been widely applied across various domains, including computer vision (Gong et al., 2016; Lan et al., 2023), recommendation (Wang et al., 2022), and graph learning (Li et al., 2023). In the context of generative modeling, CL has also been explored to facilitate training, where curricula are typically constructed at the data or pixel level (Wu et al., 2025; Croitoru et al., 2025; Liang et al., 2025). In these approaches, samples are often ranked by prompt complexity or external reward models. However, such measures are largely heuristic and may not accurately reflect the intrinsic learning difficulty. This limitation is particularly pronounced in generative settings, where stochasticity makes reliable and well-aligned difficulty metrics difficult to define.

3 Motivation

3.1 Preliminaries

Existing generative models can be unified under the CNF framework (Chen et al., 2018; Song et al., 2020b; Lipman et al., 2023; Ho et al., 2020). Specifically in training, given a data sample $x_1 \in \mathbb{R}^d$, a paired noise $\epsilon \sim \mathcal{N}(0, I)$, and a timestep $t \in (0, 1)$, an intermediate latent x_t is constructed as

$$x_t = \alpha(t)x_1 + \beta(t)\epsilon, \tag{1}$$

where $\alpha(t)$ and $\beta(t)$ are timestep-dependent coefficients satisfying $\alpha(1) = \beta(0) = 1$ and $\alpha(0) = \beta(1) = 0$. This formulation defines a probability path, named ***flow***, that transforms a simple prior distribution, *i.e.*, standard Gaussian, into the target data distribution. In this paper, we consider the straightforward and widely adopted linear path, i.e., $\alpha(t) = t$ and $\beta(t) = 1 - t$. Under this formulation, the training objective u_t can be broadly categorized into two prediction types:

1. *v-prediction* (Lipman et al., 2023): the model predicts the *velocity*, *i.e.*, $u_t \coloneqq x_1 - \epsilon$.
2. *ϵ-prediction* (Ho et al., 2020): the model predicts the *noise*, *i.e.*, $u_t \coloneqq \epsilon$.

The model is then optimized to matching the above flow formulation, termed as FM:

$$\mathcal{J}_{\text{FM-}v/\epsilon} = \mathbb{E}_{x_1, \epsilon \sim \mathcal{N}(0,1),\, t \in (0,1)} \left[\| u_\theta(x_t, t) - u_t \|_2^2 \right], \tag{2}$$

where $u_\theta(x_t, t)$ denotes the model prediction, whose form depends on the chosen training target, velocity v or noise ϵ, timestep t is sampled from $(0, 1)$, and $\theta \in \Theta$ represents the available hypotheses. Although the two parameterizations on training objectives differ mathematically, they are fundamentally equivalent in that both aim to model the same underlying probability path.

In standard practice, state-of-the-art flow-based generative models (Rombach et al., 2022; Yang et al., 2025; Wan et al., 2025; Kong et al., 2024) typically sample the timestep t from a fixed distribution. Common choices include the uniform distribution (Lipman et al., 2023) and the logit-normal distribution (Esser et al., 2024). However, the motivation and the implementation for adaptive distribution of timestep sampling remain largely underexplored.

3.2 Flow-based Generative Models as Denoisers

In this paper, we reinterpret existing generative models from a denoising perspective (Gagneux et al., 2025; Li & He, 2025). Specifically, let $D_\theta(x_t, t)$ denote a conventional denoiser that predicts the

clean image from a noisy latent. Such a model is typically trained to reconstruct clean samples via

$$\mathcal{J}^*_{\text{Denoise}} = \mathbb{E}_{x_1,\epsilon\sim\mathcal{N}(0,1),\, t\in(0,1)} \left[\|D_\theta(x_t,t) - x_1\|_2^2\right]. \tag{3}$$

Under this formulation, the predictions of flow-based generative models can be reparameterized as

$$u_{\theta\text{-}v}(x_t,t) = \frac{D_\theta(x_t,t) - x_t}{1-t}, \quad u_{\theta\text{-}\epsilon}(x_t,t) = \frac{x_t - tD_\theta(x_t,t)}{1-t}. \tag{4}$$

Substituting Equation 4 into Equation 2 yields

$$\mathcal{J}^*_{\text{FM-}v/\epsilon} = \mathbb{E}_{x_1,\epsilon\sim\mathcal{N}(0,1),\, t\in(0,1)} \left[\omega_{t\text{-}v/\epsilon} \cdot \|D_\theta(x_t,t) - x_1\|_2^2\right], \tag{5}$$

where $\omega_{t\text{-}v} = (1-t)^{-2}$ and $\omega_{t\text{-}\epsilon} = t^2(1-t)^{-2}$ are timestep-dependent weights induced by the reparameterization for *v-prediction* and *ϵ-prediction* frameworks, respectively.

Comparing $\mathcal{J}^*_{\text{Denoise}}$ and $\mathcal{J}^*_{\text{FM-}v/\epsilon}$, we find that generative model training is equivalent to optimizing a standard denoising objective with an additional timestep-dependent weighting. Notably, this weighting emphasizes larger timesteps, corresponding to lower noise, than conventional denoising formulations. As this additional weighting effectively reshapes the timestep distribution, it underscores the critical role of timestep sampling in training generative models. Although prior works (Esser et al., 2024; Li & He, 2025) have recognized its importance, they typically employ fixed sampling distributions, such as uniform or logit-normal, throughout training. As we demonstrate in the following two observations, such strategies may be suboptimal for optimizing flow-based generative models.

Observation 1: Noise level as measure of difficulty. A natural question arising from the reparameterization in Equation 5 is *why existing generative models achieve superior performance over conventional denoisers by merely introducing a timestep-dependent weighting*. To address this, we propose a principled hypothesis, supported by empirical validation and tailored theoretical analysis.

Intuitively, *denoising strength can serve as a proxy for difficulty*. For a typical denoiser, latents at smaller timesteps, or higher noise levels, are naturally harder to reconstruct, while those at larger timesteps are easier. Thus, training exposes the model to a spectrum of difficulty levels: samples are intrinsically similar but corrupted to varying degrees, inducing different denoising difficulties.

Meanwhile, all noise levels are required during inference, where the denoising trajectory starts from the initial noise and progressively reaches the clean data. This necessitates that the model acquire competence across the entire range of noise levels, from high-noise to low-noise regimes. The empirical success of FM objectives $\mathcal{J}^*_{\text{FM-}v/\epsilon}$ over conventional denoising objectives $\mathcal{J}^*_{\text{Denoise}}$, however, suggests that the timestep-dependent weighting plays an important role in shaping how the model learns across this range. In particular, the weighting effectively emphasizes lower-noise levels, raising a key question: *why does emphasizing easier samples improve performance, despite the fact that harder, high-noise samples might seem to require more attention?*

To investigate this issue, we conduct the experiment shown in Figure 2 to evaluate the denoising ability of models trained under different noise regimes. Specifically, we train five models $\mathcal{G}_{i\text{-}v/\epsilon}$, each restricted to a specific timestep interval $t \in [0.2i, , 0.2i + 0.2)$ for $0 \le i \le 4$, while keeping all other settings identical. We then evaluate each model at its respective training noise level, together with a normally trained model evaluated at all five noise levels. At each evaluation level, the model directly predicts the clean data from its one-step estimated velocity/noise, and performance is measured by FID.

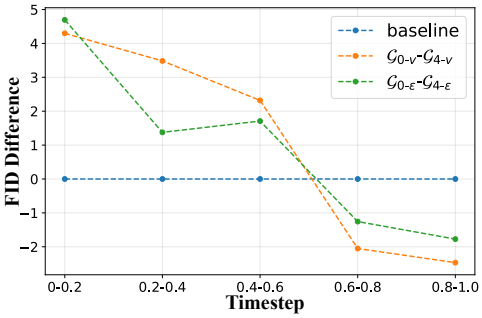


Figure 2: FID differences between the normally trained baseline and $\mathcal{G}_{0\text{-}v/\epsilon}$–$\mathcal{G}_{4\text{-}v/\epsilon}$ on CIFAR-10. Positive values indicate degraded performance, while negative values indicate improvement.

The results reveal a non-trivial pattern. Models trained on lower-noise regimes (*i.e.*, $\mathcal{G}_{3\text{-}v/\epsilon}$ and $\mathcal{G}_{4\text{-}v/\epsilon}$) exhibit stronger denoising performance than the normally trained model, whereas those trained on higher-noise regimes exclusively (*i.e.*, $\mathcal{G}_{0\text{-}v/\epsilon}$) perform significantly worse. At first glance, this appears counter-intuitive, as one might expect training solely on more challenging, high-noise

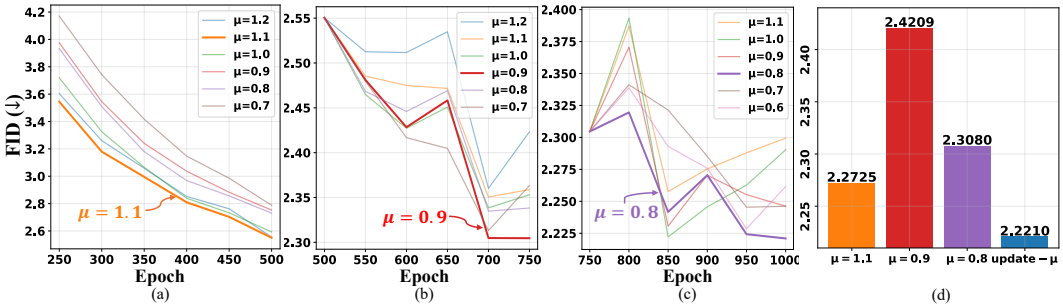


Figure 3: Stage-wise evolution of the model's denoising ability. (a)-(c) FID for models trained with different values of logit-normal mean μ across three training stages, *i.e.*, Epoch 0-500, 500-750 and 750-1000. The optimal μ progressively decreases ($1.1 \rightarrow 0.9 \rightarrow 0.8$), showing that the model's denoising ability evolves. (d) Comparison between static training at fixed means and the staged *update-*μ approach, demonstrating that gradually increasing noise levels can improve performance.

data to yield stronger denoising capability than the normally trained baseline. However, when interpreting noise level as a measure of sample difficulty, this phenomenon becomes more understandable: high-noise samples correspond to inherently difficult tasks, making direct optimization unstable and harder to converge, while the baseline, which is trained on a mixture of noise levels, can benefit from easier samples that stabilize learning and facilitate overall optimization. In contrast, lower-noise samples are easier to learn, allowing models to achieve stronger performance when trained in isolation. These findings suggest that directly focusing on the most difficult samples may not be optimal. Instead, they point to a more significant question: should the model be progressively exposed to harder noise levels during training? More fundamentally, *does the model's denoising ability evolve over the course of training?*

Observation 2: Denoising ability of the model evolves. In this section, we investigate further on whether the model's denoising ability improves throughout training. We conduct an experiment in which models are trained with *v-prediction* using different distributions of timesteps/noise levels. Concretely, we sample timesteps from a logit-normal distribution $\mathcal{L}\text{-}\mathcal{N}(\mu, \sigma^2)$, *i.e.*,

$$t = 1/\left[1 + \exp(-(\mu + \sigma \cdot \epsilon_0))\right], \quad \epsilon_0 \sim \mathcal{N}(0, 1), \tag{6}$$

and vary the mean μ to control the sampled noise level. A smaller μ shifts the distribution negatively, yielding smaller timesteps sampled on average, corresponding to higher noise levels.

Specifically, in Figure 3(a), we first train the same base model on CIFAR-10 dataset for 500 epochs with $\mu \in [0.7, 1.2]$ and $\sigma = 1.2$. We observe that $\mu = 1.1$ achieves the best performance (lowest FID). Starting from the checkpoint corresponding to $\mu = 1.1$ at epoch 500, in Figure 3(b) we continue training the same model for another 250 epochs with $\mu \in [0.7, 1.2]$, selecting the optimal μ for the next training stage of the current best-performing model. We repeat this procedure at epoch 750 for another 250 epochs in Figure 3(c), with a total of 1000 epochs and the model eventually converging. Across these stages, the optimal value of μ progressively decreases ($1.1 \rightarrow 0.9 \rightarrow 0.8$), reflecting a gradual rise in the noise level handled by the model. In Figure 3(d), we further compare the final performance between variants trained with a static logit-normal distribution ($\mu \in \{1.1, 0.9, 0.8\}$) and the simple staged strategy described above (denoted as *update-*μ). Together, these results support the view that denoising ability evolves over time: mastering the current noise level better prepares the model for higher-noise inputs, while training at a static level may limit later performance. More results and experiments on CelebA-64 are provided in Appendix C.6.

4 METHOD

Based on the above observations, we conclude that the optimal training of flow-based models requires an adaptive, rather than fixed, timestep sampling distribution. Moreover, different timesteps correspond to different levels of sample difficulty from the view of denoising: intuitively, smaller timesteps introduce higher noise levels and are therefore more challenging for the model to handle.

These insights naturally motivate the design of an adaptive sampler for flow-based generative models, which can be effectively framed within the paradigm of CL (Bengio et al., 2009; Wang et al., 2021). In Section 4.1, we first provide a theoretical justification for interpreting noise level as a measure of sample difficulty within the curriculum learning framework. Building upon this, Section 4.2 introduces our proposed method, termed CurFM, which incorporates a curriculum-based adaptive sampler to improve the training of flow-based generative models, without relying on intrinsically stochastic signals such as regression-based training loss.

4.1 Theoretical Basis for Noise Level as a Curriculum Difficulty Proxy

In this section, we develop the theoretical foundation underlying the use of CL in our setting. While CL has demonstrated strong empirical effectiveness across a wide range of tasks (Wang et al., 2021; Graves et al., 2017; Zhou et al., 2024), its success critically depends on the proper implementation of the *difficulty measurer*. In our setting, we aim to formally justify the use of noise level as the proxy for sample difficulty in CL. Following prior CL work (Weinshall et al., 2018), we define the ideal difficulty for a general denoiser within the CL framework as follows.

Definition 1. *(Ideal Denoising Difficulty) The ideal denoising difficulty of a sample x_t is defined as its minimal loss with respect to the set of optimal hypotheses, i.e.,*

$$\{\mathcal{L}_{\bar{\theta}}(x_t, x_1)\} = \left\{ \|D_{\bar{\theta}}(x_t, t) - x_1\|_2^2 \right\}, \tag{7}$$

where $\bar{\theta}$ denotes the global optimum hypothesis that minimizes the denoising objective in Equation 3.

Moreover, it has been theoretically proved that, under linear hypotheses with regression loss, when samples are presented in an easy-to-hard order according to Definition 1, the expected convergence rate of gradient descent can be accelerated (Weinshall et al., 2018). Nevertheless, these works also point out that the ideal difficulty is rarely accessible in practice. This limitation is particularly pronounced in generative modeling, where the training loss, which is commonly used as a proxy for sample difficulty in prior curriculum learning works, often exhibits significant fluctuations in optimization (Xu et al., 2023; Jeha et al., 2024) (see Figure 1), and thus is not suitable to use as difficulty metric or indication of convergence.

In this work, we provide a principled view to address this challenge. Specifically, we leverage the inherent timestep sampling mechanism in existing flow-based models. Based on Observation 1, a latent x_t with a smaller timestep corresponds to a higher noise level and is therefore more difficult for the model to learn denoising. This observation naturally yields a simple yet practical proxy for sample difficulty, thereby establishing the foundation for incorporating CL into generative modeling.

However, a fundamental question remains: does this timestep-based difficulty induce the same ranking of samples as the *Ideal Denoising Difficulty*, in accordance with our intuition? Answering this question in full generality is nearly impossible, as it would require a precise understanding of the internal mechanisms of deep neural networks, which is an open problem that remains largely unresolved (Zhang et al., 2025; Chakraborty et al., 2017). Therefore, to obtain theoretical insights, we instead follow a simplified setting similar to Weinshall et al. (2018). Specifically, we consider a restricted hypothesis class Θ^* consisting of linear denoisers of the form:

$$D_{\theta \in \Theta^*}(x_t, t) = a^\top x_t + b = w^\top x_t. \tag{8}$$

In the last equality, we rewrite the affine transformation in augmented form, where $w = [a \quad b]^\top \in \mathbb{R}^{d+1}$, and x_t denotes $[x_t \quad 1]^\top$, with some abuse of notations. Then we establish the theorem:

Theorem 1. *Let the expectation of ideal denoising difficulty at timestep $t \in (0, 1)$ be defined as*

$$\mathcal{D}_{\bar{\theta}}(t) = \mathbb{E}_{x_1, \epsilon \sim \mathcal{N}(0,I)} \left[\mathcal{L}_{\bar{\theta}}(x_t, x_1) \right]. \tag{9}$$

Then, for any $0 < \alpha < \beta < 1$, it holds that

$$\mathcal{D}_{\bar{\theta}}(\alpha) > \mathcal{D}_{\bar{\theta}}(\beta). \tag{10}$$

The proof is provided in Appendix A. Theorem 1 formally establishes that samples with smaller timesteps (*i.e.*, higher noise levels) have greater *Ideal Denoising Difficulty*, indicating that they are more difficult to denoise under this criterion and thus correspond to higher difficulty levels within the curriculum learning framework, consistent with our intuition. This result provides theoretical justification for the curriculum strategy proposed in our method.

4.2 Flow-based Generative Models with Adaptive Curriculum Sampler

In this section, we describe the design of CurFM in detail. In particular, since the *Difficulty Measurer* has already been defined in Section 4.1, we focus on the other key component of Curriculum Learning: the *Scheduler*. A straightforward approach is to adopt a fixed, pre-defined schedule that gradually decreases the mean value of logit-normal distribution in Equation 6, thereby exposing the model to progressively more difficult denoising tasks, as demonstrated in the preliminary experiment in Figure 3. While this strategy introduces a dynamic and effective curriculum of noise levels, it suffers from a key limitation: the schedule must be manually tuned for each model and dataset. As a result, such fixed schedules are often impractical in real-world settings, as they require extensive hyper-parameter search and repeated retraining to identify suitable configurations, leading to substantial computational overhead and poor generalization across different scenarios.

At the core of this issue lies the question of when the model is ready to transition to more challenging noise levels. Building on Observation 1, we define the denoising ability at training stage k (measured at 20-epoch intervals) under the current sampling parameter μ_k as

$$\mathcal{A}(\mu_k) = -\mathcal{F}\big(\{D_{\theta\text{-}v/\epsilon}(x_t, t)\}_{t\sim\mathcal{L}\text{-}\mathcal{N}(\mu_k,\sigma^2)}, \{x_1\}\big), \tag{11}$$

where $\mathcal{F}$ denotes the metric function for evaluating denoised sample quality, instantiated in our case by the FID score. Intuitively, a higher value of $\mathcal{A}(\mu_k)$ indicates stronger denoising capability under the current noise regime. Here, $D_{\theta\text{-}v/\epsilon}(x_t, t)$ denotes the predicted clean image directly reconstructed from the model's prediction of velocity or noise, namely

$$D_{\theta\text{-}v}(x_t, t) = x_t + (1-t)u_{\theta\text{-}v}(x_t, t), \quad D_{\theta\text{-}\epsilon}(x_t, t) = (x_t - (1-t)u_{\theta\text{-}\epsilon}(x_t, t))/t. \tag{12}$$

We determine the transition point adaptively based on performance stagnation since the last update. Let k_0 denote the most recent step at which μ_{k_0} was updated. We trigger an update when

$$k = \min\big\{k' \mid \max_{0<j\le\bar{T}}\{\mathcal{A}(\mu_{k'-\bar{T}+j})\} - \mathcal{A}(\mu_{k'-\bar{T}}) < \delta, \quad k' - k_0 \ge \bar{T}\big\}, \tag{13}$$

where $\bar{T}$ is the minimum number of steps (patience) and $\delta = 1\text{e-}2$ is a small positive threshold. Once this condition is satisfied, the scheduler updates the sampling distribution with step $\bar{\mu}$ and resets:

$$\mu_{k+1} \leftarrow \mu_k - \bar{\mu}, \quad k_0 \leftarrow k+1, \tag{14}$$

thereby gradually shifting the distribution toward more difficult (higher-noise) regimes.

Importantly, evaluating denoising ability requires only a single-step prediction, making the proposed mechanism computationally efficient and introducing negligible overhead to training. We formalize the overall procedure in Algorithm 2, where $\bar{T}$ and $\bar{\mu}$ control the patience and the step size of the distribution mean update, respectively. To mitigate potential forgetting, we retain the logit-normal distribution (Equation 6) as the base distribution and adaptively adjust its mean throughout training.

Compared to standard FM, CurFM requires only minimal changes to the training pipeline, introducing little additional complexity while leaving inference unchanged. Despite its simplicity, this adaptive curriculum mechanism effectively aligns the sampling strategy with the model's evolving denoising ability, leading to improved training efficiency and performance for flow-based models.

5 Experiment

In this section, we first describe the implementation and evaluation settings of CurFM. We then present quantitative comparisons with baseline samplers across multiple datasets in Section 5.2, along with extensive ablations, time-complexity and sentivity analysis to assess the effectiveness and practicality of CurFM in Section 5.3. Additional experiments, including discussions on curriculum schedule, base distribution and visualization of the induced curricula, are provided in Appendix C.

5.1 Experimental Setup

We evaluate CurFM on CIFAR-10 (Krizhevsky et al., 2009), CelebA 64×64 (Liu et al., 2015), and ImageNet 128×128 (Deng et al., 2009), comparing with common timestep samplers including logit-normal (Esser et al., 2024), uniform (Lipman et al., 2023), as well as recent methods Laplace (Hang et al., 2025), ANT (Kim et al., 2025b), and DTDC (Kim et al., 2025a). We report FID-50K (Heusel et al., 2017) as evaluation metric. For CurFM, we set $\bar{T} = 2$ and $\bar{\mu} = 0.05$. All methods use identical model architectures. Other details for implementation and baselines are deferred to Appendix B.

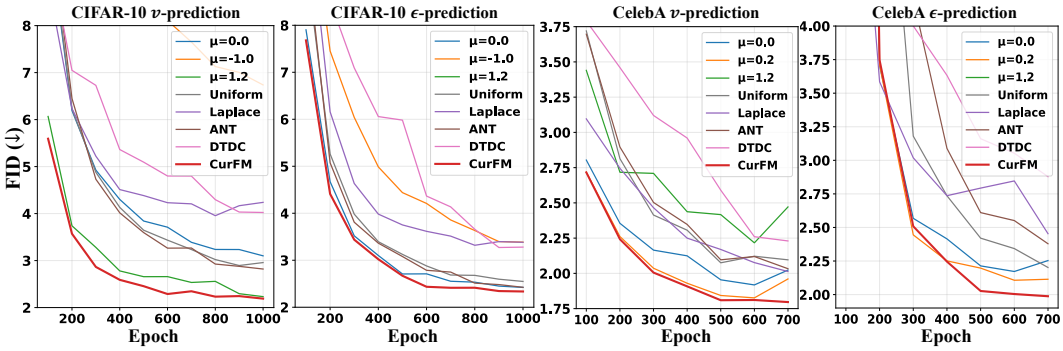


Figure 4: Comparison between CurFM and baselines under both v- and ϵ-prediction frameworks.

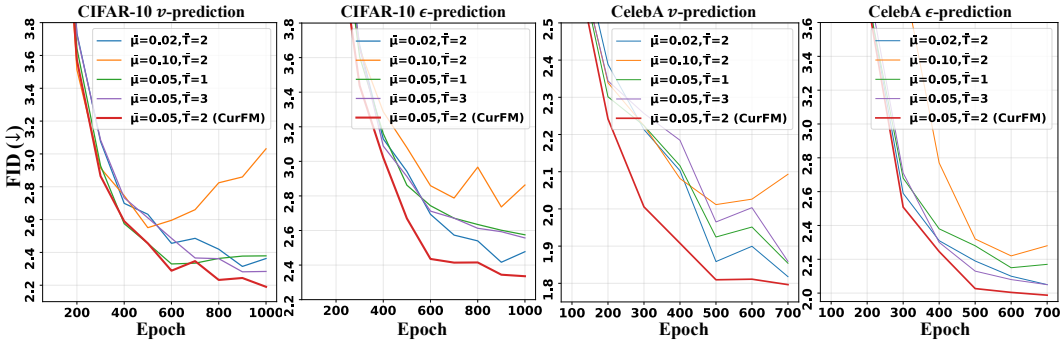


Figure 5: Hyper-parameter sentivity analysis on $\bar{T}$ and $\bar{\mu}$ of CurFM.

5.2 Main Results

In this section, we quantitatively compare CurFM with state-of-the-art baseline samplers on CIFAR-10 and CelebA under both v- and ϵ-prediction frameworks (Figure 4). Results on ImageNet are deferred to Appendix C.1. We denote baseline samplers corresponding to logit-normal timestep distributions with different means by $\mu = 0.0$, $\mu = -1.0$, etc., which reflect varying emphasis on noise levels. For CIFAR-10, following standard practice, all models are trained for 1000 epochs, while for CelebA, training runs for 700 epochs, and by which point most methods have converged.

We first examine convergence behavior. CurFM consistently converges faster: on CIFAR-10, it reaches convergence in approximately 600 epochs, whereas baseline methods typically require 800 epochs or more. A similar pattern holds on CelebA, where CurFM converges around 500 epochs compared to 600+ epochs for the baselines. This advantage is clearly reflected in Figure 4, where CurFM maintains lower FID values throughout most of the training, indicating improved efficiency.

Beyond convergence speed, CurFM also achieves competitive or superior final performance. For instance, under the v-prediction setting, it attains an FID of 2.19 on CIFAR-10 and 1.79 on CelebA, outperforming all baseline samplers. This suggests that CurFM more effectively schedules training across noise levels with the adaptive timestep sampler, aligning the learning process with the model's capability at different stages and thereby improving overall denoising performance. We further observe that the advantages on convergence speed and final performance hold across different training objectives, *i.e.*, both *velocity* and *noise* prediction. This consistency highlights the generality of the proposed curriculum-based adaptive sampling strategy beyond a specific dataset or formulation.

5.3 Analysis

Hyper-parameter Sensitivity In this section, we evaluate the sensitivity of CurFM to the hyper-parameters $\bar{T}$ and $\bar{\mu}$. Specifically, we vary $\bar{T} \in \{1, 2, 3\}$ with $\bar{\mu} = 0.05$, and $\bar{\mu} \in \{0.02, 0.05, 0.1\}$ with $\bar{T} = 2$, on CIFAR-10 and CelebA under both v- and ϵ-prediction frameworks. From the results in Figure 5, we observe that a large step size ($\bar{\mu} = 0.1$) degrades both convergence speed and

Table 1: Comparison of computational time (in seconds) between CurFM and the base sampler.

	CIFAR-10	CelebA	ImageNet
Baseline	1.96K	4.18K	25.18K
CurFM	2.08K	4.29K	25.92K
Increase.	6.2%	2.6%	2.9%

Table 2: Quantitative comparison on the performance of super-resolution on CelebA-HQ.

Sampler	FID-10K↓	PSNR↑	SSIM ↑
Uniform	8.3	18.2	0.57
Logit-Normal	7.6	20.1	0.51
CurFM	**5.7**	**22.6**	**0.61**

final performance, likely due to overly abrupt transitions in noise levels that destabilize training. In contrast, a small step size ($\bar{\mu} = 0.02$) leads to slower convergence, as the curriculum progresses too conservatively. A similar trend is observed with large $\bar{T}$, where increased patience delays adaptation and slows training. Comparing $\bar{T} = 1$ and $\bar{T} = 2$, the latter generally achieves more stable and faster convergence with better final performance. Based on these results, we adopt $\bar{T} = 2$ and $\bar{\mu} = 0.05$ as the setting for CurFM in Section 5.2, which provide a good balance between stable transition and efficient convergence. The choice of evaluation interval is further discussed in Appendix C.3.

Ablation Study We conduct an ablation study on variants of the proposed CL strategy. Specifically, we consider two variants: (1) *loss-based*, which uses the training loss instead of FID as the readiness metric to guide the curriculum; and (2) *hard-to-easy*, which reverses the scheduling direction toward lower-noise regimes. The results on CIFAR-10 are shown in Figure 6. We observe that *loss-based* consistently underperforms CurFM, likely due to the high instability of the regression-based loss, while *hard-to-easy* achieves inferior final performance, likely because its distribution eventually concentrates on low-noise regimes, where further improvement becomes limited. These results support the effectiveness of both the proposed readiness metric and scheduling direction.

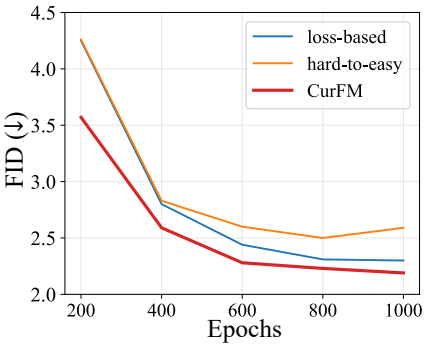


Figure 6: Ablation study on the two variants of CurFM.

Time Complexity Analysis We compare the computational cost of CurFM with the base sampler (uniform distribution) in Table 1. For a single training stage, CurFM introduces only a marginal overhead, increasing training time by 6.2% on CIFAR-10, 2.6% on CelebA, and 2.9% on ImageNet. This additional overhead is expected, as CurFM requires estimating the model's evolving denoising ability during training. However, as this estimation involves only a single forward pass, the computational costs are negligible. Overall, CurFM remains efficient while delivering consistent performance gains, and notably, introduces no additional cost during inference.

Conditional Generation We further evaluate CurFM on a conditional generation task, namely image super-resolution from 16×16 to 128×128. Following the protocol established in Saharia et al. (2022b), we train the model on FFHQ (Karras et al., 2019) and evaluate on CelebA-HQ (Karras et al., 2018). All models use the same architecture and are trained for 100K iterations. As shown in Table 2, CurFM consistently outperforms both baseline samplers across all metrics. These results demonstrate that the proposed curriculum-based adaptive sampler extends effectively to conditional generation, improving performance by better aligning training with the model's denoising capability.

6 Conclusion

In this paper, we examine flow-based generative models through the lens of curriculum learning for denoising. We show that, under a suitable reparameterization, their training objective differs from conventional denoising objectives primarily through timestep-dependent weighting. We further establish that noise level provides a principled proxy for curriculum-level sample difficulty, with higher noise corresponding to greater difficulty. Motivated by these findings, we propose CurFM, a curriculum-based sampler that adaptively adjusts the timestep distribution to the model's evolving denoising capability. Experiments show that CurFM accelerates convergence while achieving comparable or improved performance across datasets under both v- and ϵ-*prediction* frameworks.

AI USE STATEMENT

In this work, we used generative AI tools to *[edit the research paper to improve readability]*. We did not use generative AI tools for any other tasks requiring disclosure. We reviewed all AI-assisted content and manually verified all text refined or edited by these tools. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed pseudocode, a comparison between CurFM and standard FM in Figure 7, and comprehensive hyper-parameter settings in Table 3, enabling reproduction of our results. All datasets used in this work, namely, CIFAR-10, CelebA 64×64, and ImageNet 128×128, are publicly available. For the theoretical results, we provide the proof of Theorem 1 in Appendix A.

REFERENCES

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22563–22575, 2023b.

Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: A survey of results. In *2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI)*, pp. 1–6. IEEE, 2017.

Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. *Advances in neural information processing systems*, 31, 2018.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, Nicu Sebe, and Mubarak Shah. Curriculum direct preference optimization for diffusion and consistency models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2824–2834, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Anne Gagneux, Ségolène Martin, Rémi Gribonval, and Mathurin Massias. The generation phases of flow matching: a denoising perspective. *arXiv preprint arXiv:2510.24830*, 2025.

Chen Gong, Dacheng Tao, Stephen J Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. Pmlr, 2017.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Guy Hacohen and Daphna Weinshall. On the power of curriculum learning in training deep networks. In *International conference on machine learning*, pp. 2535–2544. PMLR, 2019.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

Tiankai Hang, Shuyang Gu, Jianmin Bao, Fangyun Wei, Dong Chen, Xin Geng, and Baining Guo. Improved noise schedule for diffusion training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4796–4806, 2025.

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

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

Paul Jeha, Will Sussman Grathwohl, Michael Riis Andersen, Carl Henrik Ek, and Jes Frellsen. Variance reduction of diffusion model's gradients with taylor approximation-based control variate. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. URL https://openreview.net/forum?id=YqFIzHAfbk.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hk99zCeAb.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Jin-Young Kim, Hyojun Go, Soonwoo Kwon, and Hyun-Gyoon Kim. Denoising task difficulty-based curriculum for training diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=96GMFXsbJE.

Myunsoo Kim, Donghyeon Ki, Seong-Woong Shim, and Byung-Jun Lee. Adaptive non-uniform timestep sampling for accelerating diffusion model training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2513–2522, 2025b.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742.

Xiaohan Lan, Yitian Yuan, Hong Chen, Xin Wang, Zequn Jie, Lin Ma, Zhi Wang, and Wenwu Zhu. Curriculum multi-negative augmentation for debiased video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 1213–1221, 2023.

Sangyun Lee, Zinan Lin, and Giulia Fanti. Improving the training of rectified flows. *Advances in neural information processing systems*, 37:63082–63109, 2024.

Haoyang Li, Xin Wang, and Wenwu Zhu. Curriculum graph machine learning: A survey. *arXiv preprint arXiv:2302.02926*, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Yijun Liang, Shweta Bhardwaj, and Tianyi Zhou. Diffusion curriculum: Synthetic-to-real data curriculum via image-guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1697–1707, October 2025.

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

Xingchao Liu, Chengyue Gong, 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.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Remi Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022a. URL https://openreview.net/forum?id=08Yk-n5l2Al.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022b.

Subham S Sahoo, Aaron Gokaslan, Chris De, and Volodymyr Kuleshov. Diffusion models with learned adaptive noise. *Advances in Neural Information Processing Systems*, 37:105730–105779, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Valentin I Spitkovsky, Hiyan Alshawi, and Dan Jurafsky. From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 751–759, 2010.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Xin Wang, Yudong Chen, and Wenwu Zhu. A survey on curriculum learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):4555–4576, 2021.

Xin Wang, Hong Chen, Yuwei Zhou, Jianxin Ma, and Wenwu Zhu. Disentangled representation learning for recommendation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):408–424, 2022.

Daphna Weinshall, Gad Cohen, and Dan Amir. Curriculum learning by transfer learning: Theory and experiments with deep networks. In *International conference on machine learning*, pp. 5238–5246. PMLR, 2018.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Yilun Xu, Shangyuan Tong, and Tommi S. Jaakkola. Stable target field for reduced variance score estimation in diffusion models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WmIwYTd0YTF.

Yixian Xu, Shengjie Luo, Liwei Wang, Di He, and Chang Liu. Diagnosing and improving diffusion models by estimating the optimal loss value. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=X7JfjLKKLQ.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LQzN6TRFg9.

Kai Zhang, Wangmeng Zuo, Shuhang Gu, and Lei Zhang. Learning deep cnn denoiser prior for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3929–3938, 2017.

Xinjie Zhang, Tenggan Zhang, Lei Sun, Jinming Zhao, and Qin Jin. Exploring interpretability in deep learning for affective computing: a comprehensive review. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(7):1–28, 2025.

Yuwei Zhou, Zirui Pan, Xin Wang, Hong Chen, Haoyang Li, Yanwen Huang, Zhixiao Xiong, Fangzhou Xiong, Peiyang Xu, Wenwu Zhu, et al. Curbench: curriculum learning benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

A THEOREM PROOFS

Theorem 1. *Let the expectation of ideal denoising difficulty at timestep $t \in (0,1)$ be defined as*

$$\mathcal{D}_{\bar{\theta}}(t) = \mathbb{E}_{x_1,\epsilon\sim\mathcal{N}(0,I)}\left[\mathcal{L}_{\bar{\theta}}(x_t, x_1)\right]. \tag{15}$$

Then, for any $0 < \alpha < \beta < 1$, it holds that

$$\mathcal{D}_{\bar{\theta}}(\alpha) > \mathcal{D}_{\bar{\theta}}(\beta). \tag{16}$$

Proof. Without loss of generality, suppose the clean image $x_1 \in \mathbb{R}^d$ has zero mean and covariance matrix $\Sigma = \mathbb{E}[x_1 x_1^\top]$ which is positive definite. Recall that $\epsilon \sim \mathcal{N}(0, I)$ is an independent Gaussian noise, and for a fixed timestep $\alpha \in (0,1)$, the corresponding noisy latent is computed as

$$x_\alpha = \alpha x_1 + (1-\alpha)\epsilon. \tag{17}$$

Then its *Ideal Denoising Difficulty* is calculated as

$$\mathcal{L}_{\bar{\theta}}(x_\alpha, x_1) = ||D_{\bar{\theta}}(x_\alpha, \alpha) - x_1||_2^2, \tag{18}$$

where $\bar{\theta}$ denotes the global optimal hypothesis that minimizes Equation 3. Importantly, the denoiser takes as input not only the noisy latent x_α but also the corresponding noise level (timestep) α. Consequently, the global optimum remains optimal for any specific $\alpha \in (0,1)$; *i.e.*, the parameter $\bar{\theta}$ simultaneously minimizes the conditional objective at each fixed timestep α. Then we have

$$\begin{aligned}
\mathcal{D}_{\bar{\theta}}(\alpha) = \mathcal{J}^*_{\text{Denoise}}(\bar{\theta}, \alpha) &= \min_{\theta\in\Theta^*} \mathcal{J}^*_{\text{Denoise}}(\theta, \alpha) && (19)\\
&= \min_{\theta\in\Theta^*} \mathbb{E}_{x_1,\epsilon\sim\mathcal{N}(0,I)}\left[||D_\theta(x_\alpha,\alpha) - x_1||_2^2\right] && (20)\\
&= \min_{w} \mathbb{E}_{x_1,\epsilon\sim\mathcal{N}(0,I)}\left[||w^\top x_\alpha - x_1||_2^2\right]. && (21)
\end{aligned}$$

The optimal solution to Equation 21 can be derived through normal equations:

$$\begin{aligned}
\bar{w} &= \arg\min_{w} \mathbb{E}_{x_1,\epsilon\sim\mathcal{N}(0,I)}\left[||w^\top x_\alpha - x_1||_2^2\right] && (22)\\
&= \mathbb{E}[x_\alpha x_\alpha^\top]^{-1}\mathbb{E}[x_\alpha x_1^\top] && (23)\\
&= \left(\alpha^2\Sigma + (1-\alpha)^2 I\right)^{-1}\alpha\Sigma. && (24)
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathcal{D}_{\bar{\theta}}(\alpha) &= \operatorname{tr}\left(\mathbb{E}\left[(\bar{w}^\top x_\alpha - x_1)(\bar{w}^\top x_\alpha - x_1)^\top\right]\right) && (25)\\
&= \operatorname{tr}\left(\bar{w}^\top \mathbb{E}\left[x_\alpha x_\alpha^\top\right]\bar{w} - \bar{w}^\top\mathbb{E}[x_\alpha x_1^\top] - \mathbb{E}[x_1 x_\alpha^\top]\bar{w} + \mathbb{E}[x_1 x_1^\top]\right) && (26)\\
&= \operatorname{tr}\left(\Sigma - \alpha^2\Sigma\left(\alpha^2\Sigma + (1-\alpha)^2 I\right)^{-1}\Sigma\right). && (27)
\end{aligned}$$

Since Σ is positive definite, we can diagonalize it: $\Sigma = U\Lambda U^\top$ with U orthogonal and $\Lambda = \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_d)$, $\lambda_i > 0$. Then

$$\begin{aligned}
\Sigma\left(\alpha^2\Sigma + (1-\alpha)^2 I\right)^{-1}\Sigma &= \Sigma U\left(\alpha^2\Lambda + (1-\alpha)^2 I\right)^{-1} U^\top\Sigma && (28)\\
&= U\Lambda\left(\alpha^2\Lambda + (1-\alpha)^2 I\right)^{-1}\Lambda U^\top. && (29)
\end{aligned}$$

Substituting, we have

$$\mathcal{D}_{\bar{\theta}}(\alpha) = \operatorname{tr}\left(\Lambda - \alpha^2 U\Lambda\left(\alpha^2\Lambda + (1-\alpha)^2 I\right)^{-1}\Lambda U^\top\right). \tag{30}$$

The trace can be computed as the sum over diagonal entries, *i.e.*,

$$\begin{aligned}
\mathcal{D}_{\bar{\theta}}(\alpha) &= \sum_{i=1}^{d}\left(\lambda_i - \frac{\alpha^2\lambda_i^2}{(1-\alpha)^2 + \alpha^2\lambda_i}\right) && (31)\\
&= \sum_{i=1}^{d}\frac{(1-\alpha)^2\lambda_i}{(1-\alpha)^2 + \alpha^2\lambda_i}. && (32)
\end{aligned}$$

Denote $f_i(\alpha) = \frac{(1-\alpha)^2\lambda_i}{(1-\alpha)^2+\alpha^2\lambda_i}$, we have

$$\frac{d}{d\alpha}f_i(\alpha) = \frac{-2(1-\alpha)\lambda_i[(1-\alpha)^2+\alpha^2\lambda_i] - (1-\alpha)^2\lambda_i[-2(1-\alpha)+2\alpha\lambda_i]}{[(1-\alpha)^2+\alpha^2\lambda_i]^2} \tag{33}$$

$$= -\frac{2\alpha(1-\alpha)\lambda_i^2}{[(1-\alpha)^2+\alpha^2\lambda_i]^2} < 0, \quad \forall\alpha \in (0,1). \tag{34}$$

Therefore, each $f_i(\alpha)$ is strictly decreasing on $(0,1)$, $1 \le i \le d$.

Finally, we show that for any $0 < \alpha < \beta < 1$,

$$\mathcal{D}_{\bar{\theta}}(\alpha) = \sum_{i=1}^{d} f_i(\alpha) > \sum_{i=1}^{d} f_i(\beta) = \mathcal{D}_{\bar{\theta}}(\beta). \tag{35}$$

□

B Implementation Details

Algorithm 1 Training of standard **FM**.

Input: Clean image data $\{x_1\}$.
Parameter: θ: parameters for the base generative model.
BEGIN MAIN FUNCTION:

for k in $\{1, 2, \cdots, T\}$ **do**
  Forward data $\{x_1\}$.
  Sample $\epsilon \sim \mathcal{N}(0,1)$.
  Sample t from (0,1).
  $x_t \leftarrow$ GETLATENT(ϵ, x_1, t).
  loss$\leftarrow$ GETLOSS(ϵ, x_1, x_t, t).
  $\theta \leftarrow \arg\min_\theta$(loss).
end for

Algorithm 2 Training of **CurFM**.

Input: Clean image data $\{x_1\}$.
Parameter: θ: parameters for the base generative model.
BEGIN MAIN FUNCTION:
Initialize hyper-parameter $\bar{T}$ and $\bar{\mu}$.
for k in $\{1, 2, \cdots, T\}$ **do**
  Forward data $\{x_1\}$.
  Sample $\epsilon \sim \mathcal{N}(0,1)$.
  if $\max_{0<j\le\bar{T}}\{\mathcal{A}(\mu_{k-\bar{T}+j})\} - \mathcal{A}(\mu_{k-\bar{T}}) < \delta$ **and** $k - k_0 \ge \bar{T}$ **then**
    $\mu_{k+1} \leftarrow \mu_k - \bar{\mu}, \quad k_0 \leftarrow k+1.$
  else
    $\mu_{k+1} \leftarrow \mu_k$.
  end if
  Sample $t \sim \mathcal{L}$-$\mathcal{N}(\mu_k, \sigma^2)$.
  $x_t \leftarrow$ GETLATENT(ϵ, x_1, t).
  loss$\leftarrow$ GETLOSS(ϵ, x_1, x_t, t).
  $\theta \leftarrow \arg\min_\theta$(loss).
end for

Figure 7: Comparison between the training procedures of FM and CurFM. CurFM introduces a simple yet fundamental modification to FM, requiring only a minimal change in the implementation.

B.1 Training and Evaluation

Following Lipman et al. (2023), all models, including baselines and CurFM, are trained using the same UNet (Dhariwal & Nichol, 2021) architecture specified in Table 3. For CurFM, we set the hyper-parameters $\bar{T} = 2$, $\bar{\mu} = 0.05$, and use an evaluation interval of 20 epochs, based on the analysis in Section 5.3 and Section C.3. To mitigate potential forgetting, we adopt a Logit-Normal distribution $\mathcal{L}$-$\mathcal{N}(\mu, \sigma^2)$ as the base sampling distribution and apply our curriculum strategy by adjusting its mean, thereby controlling the exposure to noise levels of varying difficulty throughout training. We fix $\sigma = 1.2$, and set the initial value of $\mu_0 = 1.2, 0.2$ and 1.2 for CIFAR-10, CelebA, and ImageNet, respectively, which are widely-used settings in practice. The overall training procedures for CurFM and standard flow-based generative models are illustrated in Figure 7. As

Table 3: Hyper-parameters used for training base models on the three datasets, CIFAR-10, CelebA and ImageNet.

	CIFAR-10	**CelebA**	**ImageNet**
Channels	256	256	256
Depth	2	3	3
Channels Multiple	1,2,2,2	1,2,2,2	1,1,2,3,4
Heads	4	4	4
Heads Channels	64	64	64
Attention Resolution	16	16,8	32,16,8
Dropout	0.0	0.0	0.0
Effective Batch Size	512	256	32
GPUs	1	4	4
Epochs	1000	700	500
Learning Rate	5e-4	5e-4	1e-4
Learning Rate Scheduler	Polynominal Decay	Polynominal Decay	Polynominal Decay
ODE Method	heun2	heun2	dopri5
ODE Option	$\mathrm{nfe} = 50$	$\mathrm{nfe} = 50$	$\mathrm{atol} = \mathrm{rtol} = 1\mathrm{e}^{-5}$

shown, CurFM introduces only minimal changes to the training pipeline: it adaptively modifies the timestep sampling distribution according to the model's evolving denoising capability, while leaving the architecture and other hyper-parameters unchanged. Importantly, CurFM incurs no additional overhead at inference time, as the modification is applied exclusively during training. All the models are trained using Nvidia A100 GPUs with 40GB memory.

We evaluate CurFM on CIFAR-10 (Krizhevsky et al., 2009), CelebA 64×64 (Liu et al., 2015), and ImageNet 128×128 (Deng et al., 2009), covering benchmarks of varying scales. We compare our method against a range of baseline samplers, detailed in Section B.2. Performance is measured using FID-50K (Heusel et al., 2017), with evaluations conducted every 100 epochs to produce the performance curves shown in Figure 4. The hyper-parameter sensitivity analysis in Figure 5 and the ablation study in Figure 6 follow the same experimental settings as the main results. We do not employ classifier-free guidance during inference. For sampling, we use the `heun2` ODE solver with 50 steps on CIFAR-10 and CelebA, and `dopri5` with absolute and relative tolerances of $1\mathrm{e}{-5}$ on ImageNet.

B.2 Baselines

We compare our method with multiple baselines, with different sampling distributions for timesteps during training, specifically,

- Uniform (Lipman et al., 2023): Timesteps are sampled from a uniform distribution over $(0, 1)$, *i.e.*, $t \sim \mathcal{U}(0, 1)$.
- Logit-Normal (Esser et al., 2024): Timesteps are sampled from a Logit-Normal distribution, *i.e.*, $t \sim \mathcal{L}\text{-}\mathcal{N}(\mu, \sigma^2)$.
- Laplace (Hang et al., 2025): Hang et al. (Hang et al., 2025) investigate several timestep sampling strategies parameterized in the log signal-to-noise ratio space, where $\lambda = \log \mathrm{SNR}(t)$, and report that a Laplace distribution yields the strongest empirical performance. Specifically, λ is sampled from a Laplace distribution

$$p(\lambda) = \frac{1}{2b} \exp\left(-\frac{|\lambda - \mu|}{b}\right), \tag{36}$$

  with $\mu = 0$ and $b = 1$. In the main results, we adopt this sampling strategy as one of the baseline methods.
- ANT (Kim et al., 2025b): Timesteps are sampled from a Beta distribution $\mathcal{B}(a, b)$, where the parameters a and b are predicted by a neural network. The network is optimized to maximize the reduction of the training objective during optimization. However, due to the

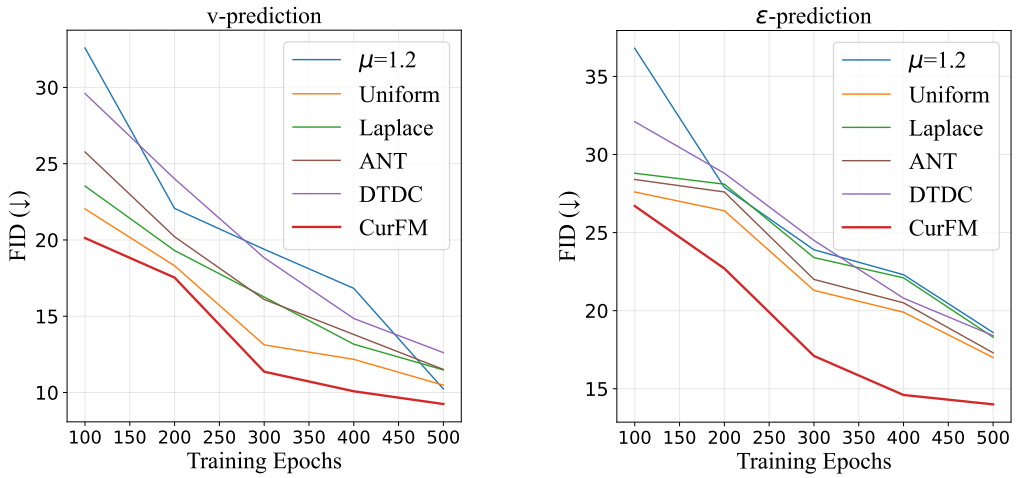


Figure 8: Quantitative comparison between CurFM and baselines on ImageNet 128×128 under both v- and ϵ-prediction frameworks.

highly stochastic nature of the regression-based objectives used in generative modeling, the resulting training signals are often noisy and unstable, which may limit their effectiveness for guiding timestep scheduling.

- DTDC (Kim et al., 2025a): Kim et al. (Kim et al., 2025a) apply the BabyStep curriculum strategy (Spitkovsky et al., 2010) to diffusion model training. Specifically, the timestep interval $(0, 1)$ is partitioned into N groups, where earlier groups correspond to lower timesteps (*i.e.*, higher noise levels) and later groups correspond to higher timesteps (*i.e.*, lower noise levels). Training then progressively transitions from earlier groups to later ones. Their method differs from ours in two important aspects. First, their transition schedule is also determined using noisy regression-based training signals. Second, they heuristically regard small timesteps as easier samples and therefore train on them first, whereas our theoretical analysis shows that small timesteps correspond to higher denoising difficulty, making them inherently more challenging for the model to learn.

C Extended Results

In this section, we provide additional results and analyses of CurFM. Specifically,

- In Section C.1, we provide additional quantitative comparisons between CurFM and the baselines on ImageNet 128×128 under both v- and ϵ-prediction frameworks.
- In Section C.2, we investigate different curriculum schedules, including *Geometric*, *Root*, and *Linear*, as well as alternative base distributions beyond the logit-normal distribution.
- In Section C.3, we analyze the choice of evaluation interval for the CurFM scheduler.
- In Section C.4, we visualize the actual curricula induced by CurFM on three datasets.
- In Section C.5, we investigate the generality of CurFM to text-to-image generation, beyond the unconditional settings in the main results and the conditional super-resolution setting in Table 2.
- In Section C.6, we provide additional results on CelebA for Observation 2, complementing the experiment in Figure 3 from the main paper.

C.1 Quantitative Comparison on ImageNet Dataset

In this section, we provide quantitative comparison between CurFM and baseline samplers on the high-resolution ImageNet 128×128 dataset, complementing the CIFAR-10 and CelebA results presented in Figure 4 of the main paper. We adopt both the *v-prediction* and ϵ-prediction frameworks,

Table 4: Comparison of FID metric on CIFAR-10 between different curriculum schedules.

Epoch	200	400	600	800	1000
Geometric	**3.55**	2.66	2.31	2.26	2.24
Root	3.63	2.65	2.36	2.27	2.32
Linear (CurFM)	3.57	**2.59**	**2.28**	**2.23**	**2.19**

Table 5: Comparison of FID metric on CIFAR-10 using Laplace as base distribution.

Epoch	200	400	600	800	1000
Laplace	6.23	4.51	4.23	3.95	4.23
+Geometric	5.24	**4.02**	3.66	3.65	3.60
+Root	**5.18**	4.13	3.73	3.62	3.61
+Linear	5.21	4.07	**3.61**	**3.60**	**3.58**

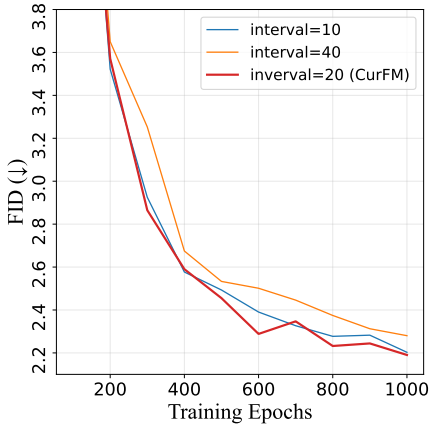


Figure 9: Performance analysis of CurFM with different evaluation intervals $\in \{10, 20, 40\}$.

and all models are trained for 500 epochs. From the results in Figure 8, we observe that CurFM achieves the fastest convergence speed among all methods, consistently attaining the lowest FID values throughout training. This trend is consistent with the observations in the main paper and further demonstrates the effectiveness of the proposed curriculum-based adaptive sampling strategy on larger-scale datasets.

C.2 Extended Analysis on Curriculum Schedules and Base Distribution

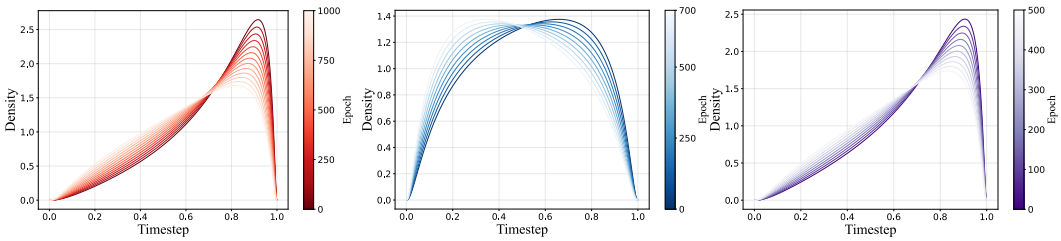
We further investigate the influence of the curriculum schedule and base timestep distribution. First, we evaluate geometric, root-2, and linear schedules (the latter used in CurFM) under the logit-normal (L-N) distribution and v-prediction on CIFAR-10 using FID-50K. The results are reported in Table 4. These schedules achieve similar performance, suggesting that although they induce different timestep updates, the resulting scheduler behavior is primarily determined by the model's current denoising ability, ultimately leading to similar adaptive curricula.

Second, to examine the generality of CurFM beyond the logit-normal distribution, we adopt a Laplace-based timestep distribution and perform curriculum scheduling by adaptively adjusting its mean, analogous to the L-N case. From the results in Table 5, we observe that CurFM consistently improves over the vanilla Laplace baseline, while different curriculum schedules exhibit similar trends to those observed under the logit-normal distribution.

Finally, the goal of our method is not to identify an *optimal* timestep distribution, but to adapt existing sampling strategies according to the model's evolving denoising ability, using noise level as a principled proxy for curriculum-level difficulty. Thus, we use the widely adopted L-N distribution as a natural baseline. Nevertheless, these experiments demonstrate that our adaptive mean adjustment readily generalizes to other base distributions and consistently improves training performance.

C.3 Selection of Evaluation Interval

In this section, we discuss the choice of evaluation interval, *i.e.*, number of epochs for each training stage, in CurFM. A smaller evaluation interval requires more frequent estimation of the model's actual denoising capability, resulting in higher computational overhead, whereas an excessively large



(a) CIFAR-10 (b) CelebA (c) ImageNet

Figure 10: Visualization of the derived curricula of timestep sampling distribution on CIFAR-10, CelebA and ImageNet, respectively.

Metric	**CLIP-Text**			**CLIP-Image**		
Training Steps	step=3000	step=6000	step=9000	step=3000	step=6000	step=9000
Baseline (FM)	0.195	0.213	0.231	0.487	0.503	0.508
CurFM	**0.207**	**0.228**	**0.259**	**0.512**	**0.529**	**0.538**

Table 6: Comparison of CLIP-Text and CLIP-Image scores for standard FM training and CurFM at different training steps.

interval reduces the responsiveness of the curriculum schedule and may lead to suboptimal adaptation during training.

To study this trade-off, we compare CurFM with evaluation intervals selected from $\{10, 20, 40\}$ on CIFAR-10, as shown in Figure 9. The results indicate that a large interval (*i.e.*, 40) slows convergence and leads to worse final performance. In contrast, intervals of 10 and 20 exhibit similar behavior throughout training, while the latter requires fewer evaluations thus less computational overhead. Considering both performance and efficiency, we adopt an evaluation interval of 20 in our setting.

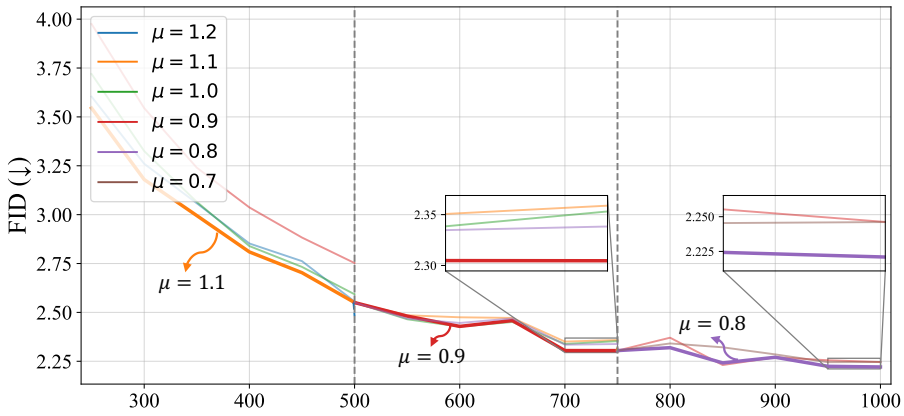
C.4 Visualization Results of Induced Curriula

In this section, we visualize the curricula induced by our proposed CurFM on CIFAR-10 (Figure 10(a)), CelebA (Figure 10(b)), and ImageNet (Figure 10(c)). Note that we use the same hyperparameter setting as in Section 5.2. We adopt the v-prediction framework and plot the density of the timestep sampling distribution across different training epochs, where darker colors correspond to earlier epochs and lighter colors correspond to later ones. Across all datasets, we observe a clear progression throughout training: the sampling distribution gradually shifts toward more challenging regions associated with higher noise levels, reflecting the model's progressively improving denoising capability.

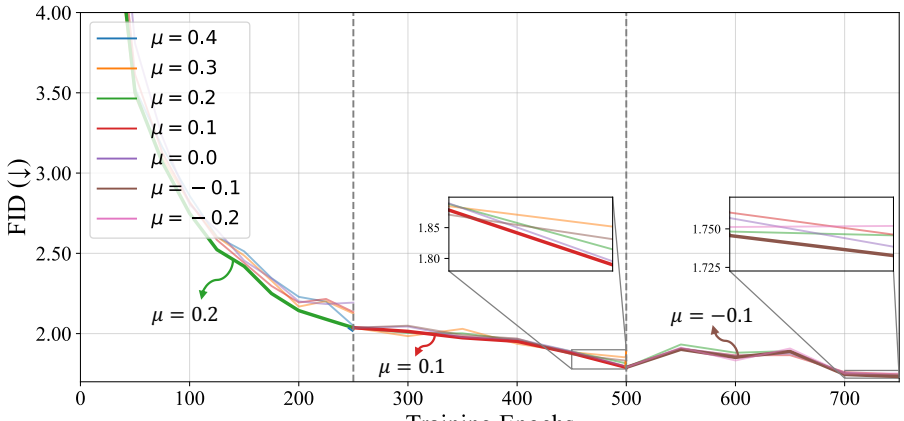
C.5 Generality of CurFM to Text-to-Image Generation

We further investigate the generality of CurFM for text-to-image (T2I) generation. In our framework, the capability measure can be instantiated according to the specific generation task. For experiments in the main results, we use one-step reconstruction FID, which directly evaluates the model's current denoising ability and provides a suitable signal for curriculum progression. We have validated this choice across both unconditional generation (CIFAR-10, CelebA, and ImageNet) and conditional generation (FFHQ super-resolution). We then examine whether FID remains effective for T2I generation.

Specifically, we train MinImagen, a public and simplified implementation of Imagen (Saharia et al., 2022a), from scratch on the publicly available relaion-pop dataset containing 580K text-image pairs. We randomly select 500 text-image pairs as the evaluation set and compare standard logit-normal timestep sampling with CurFM, using one-step reconstruction FID-1K as the curriculum signal. We set the curriculum interval to 500 steps and keep all other settings consistent with the main



(a) Results on CIFAR-10.



(b) Results on CelebA.

Figure 11: Stage-wise evolution of the model's denoising ability throughout training.

experiments. The model is trained with a batch size of 16, and we evaluate generated samples using CLIP-Text and CLIP-Image scores throughout training.

As shown in Table 6, CurFM consistently outperforms the baseline under both CLIP-Text and CLIP-Image metrics throughout training, despite neither metric being used to guide the curriculum. This suggests that one-step reconstruction FID can serve as an effective proxy for the evolving denoising capability in text-conditioned image generation, extending beyond the unconditional settings considered in the main experiments.

These results demonstrate the potential applicability of CurFM to broader generative settings. For text-to-video generation, metrics such as Fréchet Video Distance (FVD) (Unterthiner et al., 2018) could serve as candidate capability measures. More generally, CurFM provides an adaptive scheduling framework in which the specific capability measure can be selected according to the generation task.

C.6 Extended Results on Observation 2

In Observation 2, we argue that the model's denoising capability evolves throughout training. In this section, we provide additional experimental results corresponding to Figure 3 in the main paper to further support this observation. Specifically, we extend the experiments to the CelebA dataset while keeping all other experimental settings unchanged.

We divide training into three stages: 0–500, 500–750, and 750–1000 epochs for CIFAR-10, and 0–250, 250–500, and 500–750 epochs for CelebA. Within each stage, we train the same base model

using Logit-Normal timestep sampling distribution with different values of μ. Models in each stage are initialized from the best-performing checkpoint of the previous stage. We evaluate the models every 100 epochs on CIFAR-10, and every 50 epochs on CelebA. The results are presented in Figure 11.

From the results, we observe that the optimal value of μ gradually decreases over the course of training ($1.1 \rightarrow 0.9 \rightarrow 0.8$ for CIFAR-10 and $0.2 \rightarrow 0.1 \rightarrow -0.1$ for CelebA), indicating that the model progressively shifts toward handling higher noise levels more effectively. These findings further support Observation 2 and demonstrate its consistency across different datasets.

D VISUALIZATIONS

In this section, we present qualitative samples generated by CurFM. Specifically, Figure 12 and Figure 14 show samples generated under the v-prediction framework on CIFAR-10 and CelebA, respectively. Similarly, Figure 13 and Figure 15 present samples generated using the ϵ-prediction framework. Finally, Figure 16 illustrates super-resolution results on the CelebA-HQ test set.

E LIMITATIONS AND FUTURE WORK

While CurFM provides a principled perspective for understanding the effectiveness of flow-based generative models relative to conventional denoising approaches through the lens of denoising difficulty, several limitations remain.

First, due to computational constraints, we are unable to evaluate CurFM on even higher-resolution settings such as ImageNet 256×256. Nevertheless, we note that we have validated the effectiveness of the proposed method across datasets of varying scales, including CIFAR-10, CelebA 64×64, and ImageNet 128×128, under both v- and ϵ-prediction frameworks.

Second, in our theoretical analysis, we follow Weinshall et al. (2018) and adopt a simplified linear hypothesis. We note that theoretically justifying curriculum difficulty in the general case remains an open problem in prior CL literature and may require a deeper understanding of the internal mechanisms of deep neural networks and the underlying learning paradigm. Nevertheless, our analysis provides sufficient insight to justify the use of noise level as a sample difficulty measure within the curriculum learning framework. We further conduct extensive experiments across multiple datasets and diverse settings to empirically validate the effectiveness and practicality of our method.

More broadly, the primary goal of this work is to understand flow-based generative models through the lens of denoising and to explain their effectiveness relative to conventional denoising methods. Building on this understanding, we propose a curriculum-based adaptive sampler that schedules training according to the model's evolving denoising capability. Exploring more sophisticated curriculum strategies therefore remains a promising direction for future work.



Figure 12: Non-cruated samples generated by CurFM trained on CIFAR-10 using *v-prediction*.



Figure 13: Non-cruated samples generated by CurFM trained on CIFAR-10 using ϵ-*prediction*.



Figure 14: Non-cruated samples generated by CurFM trained on CelebA 64×64 using *v-prediction*.



Figure 15: Non-cruated samples generated by CurFM trained on CelebA 64×64 using ϵ-*prediction*.

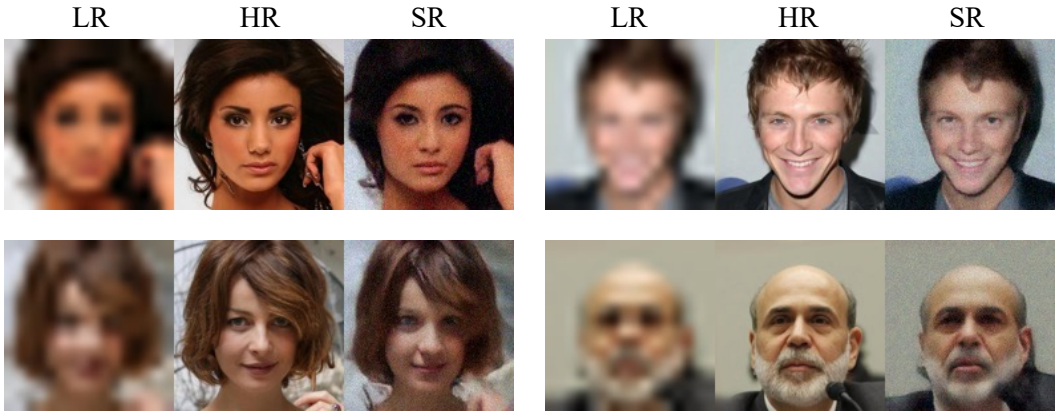


Figure 16: Super-resolution results ($16 \times 16 \rightarrow 128 \times 128$) generated by CurFM on the CelebA-HQ test set. LR, HR, and SR denote the resized low-resolution input image, the ground-truth high-resolution image, and the super-resolved image generated by CurFM from the LR input, respectively.