Understanding Flatness in Generative Models: Its Role and Benefits

Taehwan Lee [1 *] **Kyeongkook Seo** [1 *] **Jaejun Yoo** [1 2] **Sung Whan Yoon** [1 2]

Abstract

Flat minima, known to enhance generalization and robustness in supervised learning, remain largely unexplored in generative models. In this work, we systematically investigate the role of loss surface flatness in generative models, both theoretically and empirically, with a focus on diffusion models. We establish a theoretical claim that flatter minima improve robustness against perturbations in target prior distributions, leading to benefits such as reduced exposure bias and improved resilience to model quantization, preserving generative performance even under strong quantization constraints. We further observe that Sharpness-Aware Minimization (SAM), which explicitly controls the degree of flatness, effectively enhances flatness in diffusion models even surpassing the indirectly promoting flatness methods—Input Perturbation (IP), ensembling-based approach like Stochastic Weight Averaging (SWA) and Exponential Moving Average (EMA)—are less effective. Through extensive experiments on CIFAR-10, LSUN Tower, and FFHQ, we demonstrate that flat minima in diffusion models indeed improve not only generative performance but robustness.

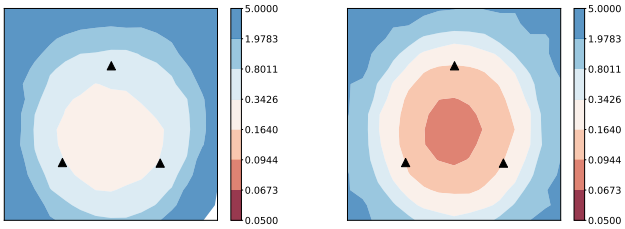
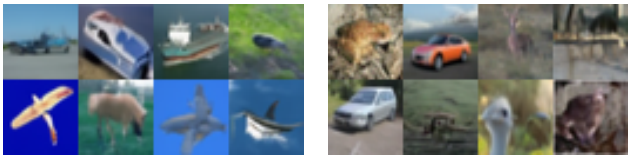


Figure 1. Loss surface of ADM (*left*) and ADM+SAM (*right*)

	FID ↓ (32 → 8)	$\|\epsilon_\theta\|^2$ gap ↓	LPF ↓
ADM	34.47 $\xrightarrow{+13.65}$ 48.02	+11.39	0.097
+SAM	**9.01** $\xrightarrow{-0.07}$ **8.94**	**+3.32**	**0.063**

Table 1. Comparison of the 1) Fréchet Inception Distance (FID) scores under 32- and 8-bit precision with (+/-) changes, 2) the gap of $\|\epsilon\|_2$ with training, 3) flatness metric called LPF for ADM and +SAM in the CIFAR-10 experiments. ↓: a lower value is preferred.



(a) ADM *(b)* ADM+SAM

Figure 2. Generated samples of 8-bit quantized models for CIFAR-10. Model parameters are directly converted from 32-bit to 8-bit after training, without further optimization or adaptation.

1. Introduction

What does it mean for a generative model to have a flat loss landscape? While flat minima have been extensively studied in supervised learning, where they are known to enhance generalization and robustness to distribution shifts, e.g., promoting stable label predictions under input shifts such as domain changes (Izmailov et al., 2018; Cha et al., 2021; Foret et al., 2021; Kwon et al., 2021; Bisla et al., 2022; Liu et al., 2022; Li et al., 2024b), their role in generative models remains largely unexplored.

At first glance, one might expect that flatness would be beneficial in generative models as well. However, generative models—particularly those based on diffusion processes—operate under fundamentally different principles: rather than mapping structured inputs to labels, they take noise as input and iteratively refine it into coherent outputs through a learned denoising process. This key difference motivates a critical question: *Do flat minima compel the generative model to produce similar contents regardless of input variations?* If so, such behavior would contradict the goal of generative diversity and potentially impair generalization. It is not evident whether flatness necessarily translates into reduced diversity in the generated samples. If not, it remains unclear how flatness affects generative modeling, and if it is

[1]Graduate School of Artificial Intelligence, Ulsan National Institute of Science and Technology [2]Department of Electrical Engineering, Ulsan National Institute of Science and Technology. Correspondence to: Jaejun Yoo <jaejun.yoo@unist.ac.kr>, Sung Whan Yoon <shyoon8@unist.ac.kr>.

Published as a paper at the 1st FoGen workshop, ICML 2026, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

indeed beneficial, how can we induce it?

As a preview of our analysis, Figure 1, 2 and Table 1 hint at the role and benefits of the generative models with flat minima. We find that SAM (Foret et al., 2021), an optimizer designed to explicitly promote flatter loss surfaces, significantly enhances flatness when applied to a baseline, i.e., Ablated Diffusion Model (ADM) (Dhariwal & Nichol, 2021), as shown in Figure 1 and Table 1 (measured using Low-Pass-Filter (LPF) flatness metric (Bisla et al., 2022)). Regarding the benefits of flatness, we observe improved FID, reduced exposure biases (as indicated by the $\|\epsilon\|^2$ gap (Ning et al., 2023b)), and lower degradation from model quantization (as reflected by the FID change from 32-bit to 8-bit precision and qualitative results in Figure 2).

Returning to the key question, we set out to investigate the implication of flat loss landscapes in generative models, particularly diffusion models (Ho et al., 2020). To make progress, we first seek to refine our core research question: *"Why might flatter minima be desirable in generative models, and what theoretical guarantees can we establish?"*

To address these questions, we conduct a theoretical and empirical analysis of flat minima in diffusion models. First, we establish a theoretical link between the model's parameter space and data probability density (**Theorem 3.3**), showing that flatter minima improve robustness to perturbations in the target prior distribution. As a consequence, we prove that the divergence between the true distribution and the learned distribution is upper-bounded, indicating that flat minima enable diffusion models to generalize not only to the training distribution but also to variations in perturbed data densities (**Theorem 3.8**). This insight suggests that flat diffusion models exhibit superior robustness when out-of-distribution is given, demonstrating their ability to cope with the error accumulation problem during the iterative sampling process, known as *exposure bias problem* (Ning et al., 2023c;b; Li et al., 2023a). Moreover, the flat loss landscape in the parameter space directly implies the robustness against model parameter changes, which indicates that the flatness offers the robustness of generative performance when the model has been quantized.

To complement our theoretical findings, we empirically examine the effects of explicit flatness regularization. Initially, we expect that applying SAM to diffusion models would yield improvements similar to those observed in discriminative settings. However, an intriguing discovery emerges: diffusion models are inherently much flatter than anticipated. Unlike discriminative models, where typical SAM trade-off values significantly impact on flatness, applying standard regularization strengths to diffusion models produced almost no discernible measurable change. Further analysis reveals that diffusion models naturally exhibit a flat loss landscape, likely due to the inherently diverse range of noise levels they are trained to denoise.

Building on this observation, we find that significantly stronger regularization is required in diffusion models for SAM to induce meaningful flatness, setting them apart from their discriminative counterparts.

To validate these findings, we conduct extensive experiments on CIFAR-10, LSUN Tower, and FFHQ. Our results show that flat minima in diffusion models consistently improve generative performance with enhanced generalization. This improved generalization offers several practical advantages. For instance, we observe that flat diffusion models better mitigate exposure bias, where errors in noise estimation accumulate over iterations, leading to more stable sampling dynamics. We also show that flat diffusion models result in a low quantization error where the model precision is reduced from 32-bit to 8-bit.

We summarize our contributions as follows. Our work provides the first systematic investigation into the role of flat minima in generative models, offering both theoretical insights and practical implications:

- We explore diffusion models through the lens of flat loss landscapes, revealing their inherent flatness.
- We establish a theoretical link between flat minima and generalization, proving that flatness improves robustness via an upper-bound on the divergence between the target and learned distributions.
- We empirically analyze IP (Ning et al., 2023c), SWA (Izmailov et al., 2018), and EMA, showing that while they improve FID, they do not significantly enhance flatness across our experiments. In contrast, SAM reshapes the loss surface and effectively promotes flatness in diffusion models.
- We demonstrate how these insights address practical challenges in generative modeling through extensive evaluations on CIFAR-10, LSUN Tower, and FFHQ, analyzing FID performance, exposure bias, model quantization error, and loss landscape properties.

2. Preliminaries and Backgrounds

2.1. Optimization for flatness

EMA. EMA is a weight-averaging method that blends a fraction of newly updated parameters with a fraction of previously accumulated parameters at each step. Formally, for trained model parameters θ, the EMA update is: $\theta_{\text{EMA}} \leftarrow (1-\lambda)\theta_{\text{EMA}} + \lambda\,\theta, \quad \lambda \in (0,1)$. Heuristically ensembling different models from multiple iterations during the course of training, reaches flat local minima and

Algorithm 1 Pseudo algorithm for SWA, EMA, SAM

Input: w_0: Initial weights, ρ: size of perturbation, c: Cycle length, λ: update momentum
Output: w, w_{SWA}, w_{EMA}

$w, w_{SWA}, w_{EMA} \leftarrow w_0$ {Initialize weights with w_0}
for $i = 0, ..., n-1$ **do**
  if SWA **then**
    $\alpha \leftarrow \alpha(i)$ {Calculate LR for the iteration}
  end if
  if SAM **then**
    $\hat{w} \leftarrow w + \rho \frac{\nabla \mathcal{L}_i(w)}{\|\nabla \mathcal{L}_i(w)\|}$ {Ascent step}
    $w \leftarrow w - \alpha \nabla \mathcal{L}_i(\hat{w})$ {Descent step}
  else
    $w \leftarrow w - \alpha \nabla \mathcal{L}_i(w)$
  end if
  if SWA & mod(i, c)=0 **then**
    $n_{models} \leftarrow i/c$ {Number of models}
    $w_{SWA} \leftarrow \frac{w_{SWA} \cdot n_{models} + w}{n_{models}+1}$ {Update average}
  else if EMA **then**
    $w_{EMA} \leftarrow (1-\lambda) \cdot w_{EMA} + \lambda \cdot w$ {Update average}
  end if
end for

improves generalization capacities (Granziol et al., 2020; Klinker, 2011; Li et al., 2024a).

SWA. SWA (Izmailov et al., 2018) stabilizes training by maintaining an ongoing average of model weights across training epochs. Rather than relying on the final parameters from a single run, SWA updates this average—typically during the latter part of training—yielding an overall smoothing effect on the loss landscape. Although SWA and EMA share the idea of parameter averaging, SWA is motivated by the search for flatter minima, which improves generalization.

SAM. SAM (Foret et al., 2021) is an optimization technique that promotes generalization by steering model parameters toward flatter regions of the loss landscape. It incorporates a sharpness term into the objective, defined as the maximal increase in loss within an ℓ_2-ball of radius ρ around the current parameters θ: $\max_{\|p\|_2 \leq \rho} \left[L(\theta + p) - L(\theta)\right]$. By minimizing this sharpness term alongside the original loss, SAM encourages solutions that are robust to small perturbations in the parameters, thus improving generalization. The pseudocode for all algorithms is provided in Algorithm 1.

IP. IP (Ning et al., 2023c) was introduced to address the exposure bias problem, which inherently arises in the autoregressive denoising process of diffusion models. They define exposure bias as a generalization failure to unseen inputs, caused by the accumulation of prediction errors over timesteps. To mitigate this issue, authors propose a smoothed diffusion model that satisfies a Lipschitz condition by training on randomly perturbed input data. This strategy aligns with the flat minima perspective, implicitly promoting generalization to nearby unseen inputs.

2.2. Score-based Generative Models

Score-based generative models (SGMs) (Song & Ermon, 2019; 2020; Song et al., 2020a;b) leverage the score function of a data distribution to iteratively generate samples by solving Stochastic Differential Equations (SDEs). Let $p_{\text{data}}(\mathbf{x})$ be the unknown data distribution. The *score function* is defined as the gradient of the log-density:

$$\nabla_{\mathbf{x}} \log p_{\text{data}}(\mathbf{x}). \tag{1}$$

Instead of modeling $p_{\text{data}}(\mathbf{x})$ explicitly, SGMs learn a neural network $s_\theta(\mathbf{x}, t)$ to approximate the score function under slight noise perturbations, t is timestep.

SGMs can be formulated as a stochastic process where noise is gradually added to the data in the forward direction and removed in the backward direction.

Forward SDE. The forward diffusion process is defined by the following SDE:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \tag{2}$$

where $f(\mathbf{x}, t)$ is the drift coefficient that controls deterministic changes, $g(t)$ is the diffusion coefficient that controls noise intensity, and $d\mathbf{w}$ denotes a standard Wiener process.

Reverse SDE. Using the learned score function, we can reverse the diffusion process to generate new samples. The backward SDE is given by:

$$d\mathbf{x} = \left[f(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})\right] dt + g(t)d\bar{\mathbf{w}}, \tag{3}$$

where $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$ is the time-dependent score function, approximated by $s_\theta(\mathbf{x}, t)$, and $d\bar{\mathbf{w}}$ is a standard Wiener process running in reverse time. To successfully estimate the score function, the score matching objective is defined as:

$$\mathcal{L}_{\text{SGM}} = \mathbb{E}_t \left[\lambda(t) \cdot \mathbb{E}_{p_t(\mathbf{x})} \left[\|s_\theta(\mathbf{x}, t) - \nabla_{\mathbf{x}_t} \log p_t(\mathbf{x})\|_2^2\right]\right]. \tag{4}$$

Diffusion models. Diffusion models (DPMs) (Sohl-Dickstein et al., 2015; Ho et al., 2020) can be viewed as a discrete-time approximation of SGMs, with the following forward and backward diffusion process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) := \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{5}$$

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}), \tag{6}$$

$$\tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t, \tag{7}$$

$$\tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t \tag{8}$$

where β_t is a predefined parameter by variance scheduling, $\alpha_t := 1 - \beta_t$, and $\bar{\alpha}_t := \Pi_{s=1}^{t} \alpha_s$.

Exposure bias. Previous works (Ning et al., 2023c;b; Li et al., 2023a) have raised an input discrepancy problem between the training and sampling phases, known as *exposure bias problem*. Specifically, during training, the DPM is provided with ground truth noisy images, whereas during sampling, it receives a denoised image that came from the previous timestep. The presence of errors in these inputs, amplified through the iterative denoising process, makes it challenging for the diffusion model to accurately predict noise. As these errors accumulate over iterative timesteps, they cause the generated samples to significantly deviate from the desired trajectory.

3. Theoretical Analysis on Flatness

We here theoretically analyze the effect of flat minima in SGMs, with an emphasis on diffusion models.

3.1. Notations

Our analysis is built upon the settings of SGM, which is a fundamental form of diffusion models. Beyond the brief preliminaries in Section 2.2, we introduce additional notations and formulations to be used in our analysis.

Following prior theoretical analysis (Li et al., 2023b), we formulate the simplified score model $s_\theta(\cdot, \cdot)$ as a random feature model consisting of an encoder and a decoder.

$$s_{\boldsymbol{\theta}}(\mathbf{x}, t) := \frac{1}{m} \boldsymbol{\theta} \sigma\big(\mathbf{W}^\top \mathbf{x} + \mathbf{U}^\top \mathbf{e}_t\big), \tag{9}$$

where $\mathbf{x} \in \mathbb{R}^{d \times 1}$ is the given input, $\mathbf{e}_t$ is the timestep embedding for t, a group of parameters includes $\boldsymbol{\theta} \in \mathbb{R}^{d \times m}$, $\mathbf{W} \in \mathbb{R}^{d \times m}$ and $\mathbf{U} \in \mathbb{R}^{d_e \times m}$, and a few positive integers include d, m, and d_e. By following the training procedures of diffusion models, $\boldsymbol{\theta}$ is the learnable parameter, while others, i.e., $\mathbf{W}$ and $\mathbf{U}$, which respectively embed $\mathbf{x}$ and $\mathbf{e}_t$, are set to be frozen. For a given timestep t, the score matching loss objective defined in Equation (4)

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, t, p_t) := \| s_{\boldsymbol{\theta}}(\mathbf{x}, t) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x}) \|_2^2, \tag{10}$$

where $p_t(\mathbf{x})$ is the probability density function for the target distribution at time t.

3.2. Mathematical Claims

For the main claims, we follow the formulation of Equation (10), while omitting timestep t without loss of generality. Our mathematical claims are valid for all timesteps. We start from defining flat minima in SGMs (**Definition 3.1**), and the robustness against the gap of ground truth prior $p(\mathbf{x})$ and the estimated prior $\hat{p}(\mathbf{x})$ (**Definition 3.2**). For the flat minima definition, we rehearse the definition of flat minima for the supervised learning case (SHI et al., 2021) and extend it to the Δ-flat minima of SGMs.

Definition 3.1. (Δ-*flat minima*) Let us consider a SGM with loss function $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, p)$. A minimum $\boldsymbol{\theta}^*$ is Δ-flat minima when the following constraints are hold:

$$\forall\, \delta \in \mathbb{R}^{d \times m} \text{ s.t. } \|\delta\|_2 \leq \Delta,\ \ \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^* + \delta, p) = l^*$$
$$\exists\, \delta \in \mathbb{R}^{d \times m} \text{ s.t. } \|\delta\|_2 > \Delta,\ \ \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^* + \delta, p) > l^*,$$

where $l^* := \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, p)$ and $\Delta \in \mathbb{R}^+$.[1]

To extend to robustness against the erroneous estimation of prior probability, we further define the $\mathcal{E}$-distribution gap robustness of SGMs, where $\mathcal{E}$ indicates the divergence between the ground truth p and the estimated $\hat{p}$:

Definition 3.2. ($\mathcal{E}$-*distribution gap robustness*) A minimum $\boldsymbol{\theta}^*$ is $\mathcal{E}$-distribution gap robust when the following constraints are hold:

$$\forall\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p \| \hat{p}) \leq \mathcal{E},\ \ \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, \hat{p}) = l^*$$
$$\exists\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p \| \hat{p}) > \mathcal{E},\ \ \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, \hat{p}) > l^*,$$

where $D(\cdot \| \cdot)$ is the divergence between two probability density functions, $\hat{p}$ is the perturbed prior distribution of $\mathbf{x}$, and $\mathcal{E}$ is a positive real number.

Let us then bridge the flatness and the robustness. First, based on the definitions of flat minima (**Definition 3.1**) and distributional gap robustness (**Definition 3.2**), we try to formulate how the perturbation of the parameter, i.e., $\boldsymbol{\theta} + \delta$, links to the perturbed prior distribution $\hat{p}$, with the given equality as follows:

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\theta} + \delta, p) = \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \hat{p}). \tag{11}$$

This means how the parameter perturbation is translated into the shift of the prior, while keeping the loss unchanged. Let us provide a mathematical claim for it:

Theorem 3.3. *(A perturbed distribution) For a given prior distribution of $p(\mathbf{x})$ and the δ-perturbed minimum, i.e., $\boldsymbol{\theta}+\delta$, the following $\hat{p}(\mathbf{x})$ satisfies the equality (11):*

$$\hat{p}(\mathbf{x}) = e^{-I(\mathbf{x}, \delta)} p(\mathbf{x}), \tag{12}$$

where $I(\mathbf{x}, \delta) := \frac{1}{2} \mathbf{x}^\top (\delta \mathbf{W}^\top) \mathbf{x} + \mathbf{x}^\top \delta (\mathbf{U}^\top \mathbf{e}) + C$ with $\delta \mathbf{W}^\top$ being symmetric, and $C \in \mathbb{R}$ is set to satisfy $\int_{\mathbb{R}^d} e^{-I(\mathbf{x}, \delta)} p(\mathbf{x}) d\mathbf{x} = 1$.

For analytical convenience, **Theorem 3.3** derives a perturbed density under symmetric $\delta \mathbf{W}^\top$, showing that parameter perturbation translates into a "scaling" of probability

[1] $\forall$ means "for all," $\exists$ means "there exists," and $\mathbb{R}^+$ indicates the set of positive real numbers.

density. Note that C absorbs the normalizing constant. However, the scaling depends on various parameters. Thus, it is non-trivial to imagine the behavior of $\hat{p}$.

Let us narrow our view to the diffusion model, where it predicts a noise signal sampled from a Gaussian distribution, i.e., $\mathcal{N}(0, \mathbf{I})$. It changes our focus on handling the noise distribution, i.e., ϵ, rather than the data distribution, i.e., p. The following Cor. describes how the perturbed $\hat{\epsilon}$ looks like:

Corollary 3.4. *(Diffusion version of **Theorem 3.3**) For a given prior Gaussian distribution of noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and the δ-perturbed minimum, i.e., $\boldsymbol{\theta}+\delta$, the following $\hat{\epsilon}$ satisfies the equality* (11)*:*

$$\hat{\epsilon} = e^{-I(\mathbf{x},\delta)}\epsilon = \mathcal{N}(\boldsymbol{\mu}_\delta, \Sigma_\delta), \tag{13}$$

where $\Sigma_\delta := \left(\mathbf{I} + \frac{\delta \mathbf{W}^\top}{m}\right)^{-1}$, $\boldsymbol{\mu}_\delta := \frac{1}{m}\Sigma_\delta \delta \mathbf{U}^\top \mathbf{e}$.

Remark 3.5. **($\hat{\epsilon}$ becomes perturbed Gaussian)** For the diffusion models, we emphasize that the perturbation in the parameter space, δ, leads to the perturbation of distribution, which follows the Gaussian distribution.

Next, let us consider all possible perturbations within the ball of norm, i.e., $\|\delta\|_2 \leq \Delta$ and $\delta \mathbf{W}^\top = \mathbf{W}\delta^\top$. The resulting set of perturbed distributions induced by the parameter perturbation is defined as follows:

Definition 3.6. (A set of perturbed distributions) For a given distribution of $p(\mathbf{x})$, a set of distributions $\hat{\mathcal{P}}(\mathbf{x}; p, \Delta)$ is defined as the set of perturbed distributions $\hat{p}(\mathbf{x})$:

$$\hat{\mathcal{P}}(\mathbf{x}; p, \Delta) := \left\{ e^{-I(\mathbf{x},\delta)} p(\mathbf{x}) \middle| \|\delta\|_2 \leq \Delta, \delta \mathbf{W}^\top = \mathbf{W}\delta^\top \right\}. \tag{14}$$

From the definition, Δ-flat minimum flattens all loss values of the perturbed parameters. Consequently, the flat minimum achieves flat loss values for all possible distributions within $\hat{\mathcal{P}}$, as formally shown in the following proposition:

Proposition 3.7. *(A link from Δ-flatness to $\hat{\mathcal{P}}$) A Δ-flat minimum $\boldsymbol{\theta}^*$ achieves the flat loss values for all distributions sampled from the set of perturbed distributions:*

$$\forall\, p \sim \hat{\mathcal{P}}(\mathbf{x}; p, \Delta), \;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, p) = l^* \tag{15}$$

$$\exists\, p \nsim \hat{\mathcal{P}}(\mathbf{x}; p, \Delta), \;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, p) > l^*. \tag{16}$$

We finally link flatness to the distributional gap, i.e., connecting **Definition 3.1** and **3.2**. Let us anticipate the distribution in $\hat{\mathcal{P}}$ with the maximal divergence from p. The distribution gap $\mathcal{E}$ is then upper bounded by the divergence of the most outreached distribution:

Theorem 3.8. *(A link from Δ-flatness to $\mathcal{E}$-gap robustness) A Δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:*

$$\mathcal{E} \leq \max_{\hat{p} \sim \hat{\mathcal{P}}(\mathbf{x}; p, \Delta)} D(p \| \hat{p}). \tag{17}$$

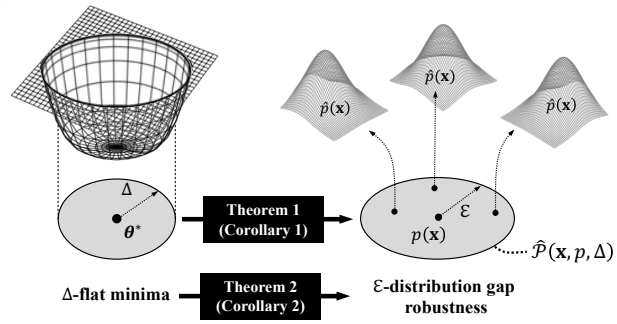


Figure 3. A conceptual illustration of theoretical analysis. **Theorem 3.3** (**Corollary 3.4** for diffusion model) translates the perturbation in the parameter space into the set of perturbed distributions. **Theorem 3.8** (**Corollary 3.9** for diffusion model) shows that flat minima lead to robustness against the distribution gap.

Let us further manipulate the upper bound for the diffusion model case. From **Corollary 3.4**, $\hat{p}$ is a form of Gaussian distribution; thus, it is possible to achieve the closed form of the upper bound as follows:

Corollary 3.9. *(Diffusion version of **Theorem 3.8**) For a diffusion model, a Δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:*

$$\begin{aligned} \mathcal{E} &\leq \max_{\|\delta\|_2 \leq \Delta} \frac{1}{2}\left[\log|\Sigma_\delta| - d + tr(\Sigma_\delta^{-1}) + \boldsymbol{\mu}_\delta{}^\top \Sigma_\delta{}^{-1} \boldsymbol{\mu}_\delta \right] \\ &\leq \frac{1}{2}\left[\sum_{i=1}^{d} (\sigma_i - \log \sigma_i) - d + \frac{\sigma_d}{m^2} \|\mathbf{U}^\top \mathbf{e}\|_2^2 \Delta^2 \right], \end{aligned} \tag{18}$$

where σ_i is an eigenvalue of Σ_δ^{-1} with the increasing order of $\sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_d$.

Remark 3.10. **(Flatter minima enhance distribution gap robustness)** It is noteworthy that flatter minima with a large Δ flatten the loss values for the far-pushed away perturbed distribution $\hat{p}$ with large divergence, leading to the larger bound of $\mathcal{E}$. It indicates that flattening the loss surface makes the generative model zero-force the loss values of diversified prior distribution, thus enhancing the robustness against the prior distribution gap.

Flatness reduces exposure bias (a gap of $\hat{\epsilon}$ and ϵ): A clear advantage of flatness is robustness to exposure bias, where the errors in noise estimation accumulate over iterations, thus severely deteriorating the generative performance. In formula, $\hat{\epsilon}$ deviates from ϵ. We argue that a generative model with a flat minimum suppresses the loss values of perturbed estimation, so that the error accumulation is sufficiently relieved. To be shown in experiments, we empirically confirm that flat minima tend to show a smaller exposure bias, leading to robust generative performance.

Flatness becomes robust to model quantization (a compression from $\boldsymbol{\theta}$ to $\hat{\boldsymbol{\theta}}$): For another benefit, flat minima

FID Score	Dataset	CIFAR-10 (32x32)		LSUN Tower (64x64)		FFHQ (64x64)	
	T'	20 steps	100 steps	20 steps	100 steps	20 steps	100 steps
Algorithms	ADM	34.47	8.80	36.65	8.57	30.81	7.53
	+EMA	10.63	4.06	7.87	2.49	19.03	6.19
	+SWA	11.00	3.78	8.72	2.31	17.93	5.49
	+IP	20.11	7.23	25.77	7.00	15.03	13.55
	+IP+EMA	9.10	3.46	7.66	2.43	11.72	4.00
	+IP+SWA	9.04	3.07	8.55	2.34	12.99	**3.54**
	+SAM	9.01	3.83	16.02	4.79	11.59	5.29
	+SAM+EMA	**7.00**	3.18	6.66	2.30	**11.41**	5.04
	+SAM+SWA	7.27	**2.96**	**6.50**	**2.27**	12.15	4.17

Table 2. FID Scores for ADM baselines with different algorithms on CIRAR-10, LSUN Tower, and FFHQ datasets. We use DDPM sampling with shorter respacing timesteps, $T' = 20, 100$.

make the model robust to the performance degradation caused by the model quantization.

Recently, quantization of generative models has become crucial, when enabling real-time generative applications (Wu et al., 2025; Li et al., 2023c; 2021). We claim that flat minima inherently bolster the robustness against the degradation due to the quantization, because the quantized model, $\hat{\boldsymbol{\theta}}$, can be viewed as a perturbed version of the full-precision parameter, i.e., $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} + \Delta$; thus the loss remains flat for the perturbation (Li et al., 2021). In experiments, we clearly demonstrate that the diffusion model with flatness is remarkably resilient to degradation from quantization compared to other methods.

A conceptual overview of our theoretical analysis is illustrated in Figure 3. The proof of **Theorem 3.3**, and **Corollary 3.4**, **3.9** are given in Appendix 1 in Supplementary.

4. Experiments

4.1. Experiments Settings

Datasets. Following (Ning et al., 2023c), we use CIFAR-10 (32x32), LSUN-tower (64x64), and FFHQ (64x64).

Baselines. We use ADM baselines[2] (Dhariwal & Nichol, 2021) for training unconditional DDPM. We incorporate flatness-enhancing approaches such as SWA (Foret et al., 2021) and SAM (Foret et al., 2021) to deliberately enhance flatness and reveal the effects of that. Since EMA may influence flatness (Klinker, 2011), we distinguish baselines with and without EMA. Additionally, we compare DDPM-IP (Ning et al., 2023c) that introduces input perturbation during training to encourage smoothness in diffusion model.

[2]https://github.com/openai/guided-diffusion

Experimental details. We trained all models for 200K steps on the CIFAR-10, LSUN Tower, and FFHQ, using the Adam optimizer with a learning rate of $1e^{-4}$, following (Dhariwal & Nichol, 2021; Ning et al., 2023a). For +IP, we used an input perturbation strength of 0.1 following (Ning et al., 2023c), and we set the cycle length c to 100 and start averaging from 180K steps for SWA. Finally, we tuned the $\rho \in [1e^{-1}, 1e^{-2}, 1e^{-3}]$ parameters for +SAM.

Evaluation metrics. We evaluate generative performance using Fréchet Inception Distance (FID) scores (Heusel et al., 2017) with subsequence of timesteps ($T' < T$). We assess Low-Pass Filter (LPF) (Bisla et al., 2022) values and loss plots under perturbation (Cha et al., 2021) to verify flat loss surface identification. To further investigate generalization, we investigate exposure bias and quantization error. For exposure bias (Ning et al., 2023b), we compare the square norm of the predicted noise, $\|\epsilon_\theta\|^2$ when models are conditioned on ground truth noisy images during training versus when they receive error-contained inputs at different sampling timesteps (Ning et al., 2023a). For quantization error, we compare the FID before and after quantization.

4.2. Generative Performance

The quantitative and qualitative results presented in Table 2 and Figure 4 demonstrate the generative performance of baselines. From Table 2, we observe that applying EMA consistently improves FID scores across all datasets and respaced timesteps, demonstrating its effectiveness in stabilizing model updates and enhancing sample quality. SWA also provides notable improvements but does not outper-

(a) Full precision (b) 8-bit precision

Figure 4. Generation results of (a) full precision (32-bit) and (b) 8-bit quantized models. We use respaced timesteps $T' = 20$ for sampling. For the 8-bit case, we selectively compare +EMA, +SAM, and their combinations, where the relatively flatter minima have been found.

LPF ↓	w/o	+EMA	+SWA
ADM	0.097	0.099	0.099
+IP	0.103	0.101	0.102
+SAM	**0.063**	**0.063**	**0.063**

↓: a lower value is preferred.

Table 3. Flatness measure on CIFAR-10. We calculate the loss with the perturbed model with Gaussian noise. Lower values indicate a flatter loss landscape.

form EMA in most cases.[3] While EMA and SWA can always be applied as auxiliary options, both standalone SAM and **+SAM+SWA** achieve comparable or better FID scores, especially under 20 timesteps, where exposure bias worsens. This suggests that explicit flatness control improves generalization, in terms of enhancing sample quality and robustness. For IP, the FID score is comparable to that of SAM at 100 steps, since the perturbation in the data distribution translates into parameter perturbations (**Theorem 1**). However, IP performs poorly at 20 steps, as it lacks a principled noise direction and tends to converge to suboptimal minima. Figure 4 further supports these findings by providing visual comparisons of generated images (see Appendix 2.2).

4.3. Flatness Measurement

In Table 3, we compare LPF flatness measures (Bisla et al., 2022), where lower values indicate a flatter loss landscape. We observe that ADM already possesses a certain level of flatness. Interestingly, we find that ensemble-like averaging, such as +SWA and +EMA, fail to induce additional flatness. In contrast, +SAM finds a flatter loss surface by explicitly perturbing model weights, as evidenced by SAM-applied models consistently achieving the lowest LPF values and significantly superior performance to other baselines.

[3]Due to limited space, we include the discussion and results of the post-hoc EMA (pEMA) (et al., 2024) in Appendix 2.3.

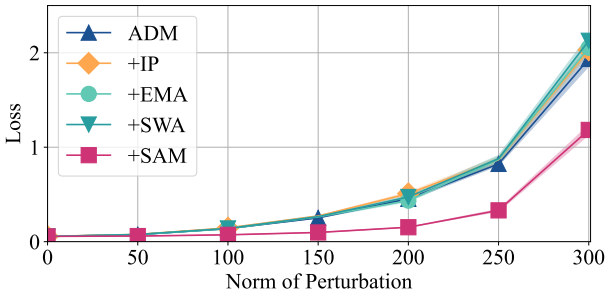


Figure 5. Loss plots under perturbation for CIFAR-10. As the norm of model perturbation increases, we plot the corresponding loss values. A smaller slope indicates a flatter loss landscape.

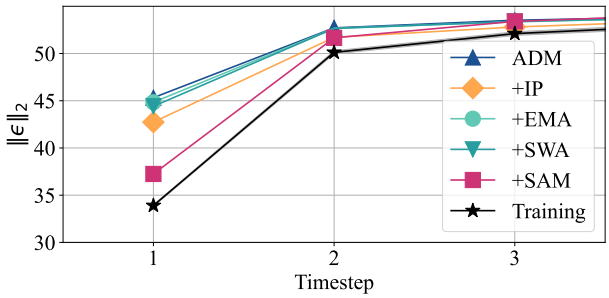


Figure 6. L2 norm of the predicted noise for CIFAR-10. As the gap with "Training" decreases, the predicted noise norm during sampling approaches the ground truth.

Also, we plot another measurement to support these findings further. Figure 5 shows how the loss value increases as the model perturbation is imposed for CIFAR-10. +IP, +EMA, and +SWA show no significant differences, coincide with the result of LPF, where these fail to find additional flatness. In this setting, +SAM shows a flatter behavior than other algorithms. It shows that diffusion models already show

32-bit → 8-bit	20 steps	100 steps
ADM	34.47 $\xrightarrow{+13.65}$ 48.02	8.80 $\xrightarrow{+3.98}$ 12.78
+EMA	10.63 $\xrightarrow{+10.02}$ 20.65	4.06 $\xrightarrow{+3.3}$ 7.36
+SAM	9.01 $\xrightarrow{-0.07}$ 8.94	3.83 $\xrightarrow{+0.19}$ 4.02
+SAM+EMA	7.00 $\xrightarrow{+0.2}$ 7.20	3.18 $\xrightarrow{-0.06}$ 3.12

Table 4. FID performance on CIFAR-10 when model parameters are quantized to 8-bit. We apply direct quantization by clipping and rounding 32-bit parameters without additional fine-tuning. The 32-bit performance results are identical to those in Table 2.

flatter loss landscapes, and ensemble-like averaging, such as +SWA and EMA, shows less impact. Moreover, finding flat minima explicitly leads to the flatness in diffusion models. We provide plots for other algorithms, including +IP+EMA, +IP+SWA, +SAM+EMA, +SAM+SWA, in the Appendix. 2.1.

4.4. Flat Diffusion Models are Robust

We show that flat diffusion models perform well under input shifts. As an instantiation, we set two cases: 1) exposure bias and 2) quantization error. Both scenarios share the common characteristic that diffusion models receive error-accumulated inputs, leading to performance degradation.

Exposure bias. In Figure 6, we compare the L2 norm of the predicted noise in diffusion models. A model that closely aligns with the training trajectory is expected to exhibit lower exposure bias, thereby mitigating distribution shifts over iterative sampling steps. We observe that SAM-applied models consistently maintain L2 norms closer to the training curve, indicating a lower degree of exposure bias. This result aligns well with their superior FID performance reported in Table 2, suggesting that flat diffusion models generalize better across sampling timesteps and effectively combat error accumulation.

Model quantization error. Table 4 presents the FID scores of quantization across baselines, when reducing precision from 32-bit to 8-bit. Here, we quantize model parameters to 8-bit values using scaling and rounding, without any additional fine-tuning or retraining. Thanks to +SAM's robustness to model perturbation, flatter diffusion models results in surprisingly better robustness under quantization, supporting that +SAM is robust to model quantization.

5. Related works

Diffusion models. Diffusion probabilistic models (DPMs) (Ho et al., 2020; Dhariwal & Nichol, 2021; Song & Ermon, 2019; 2020; Song et al., 2020b) have recently emerged as powerful generative frameworks, achieving state-of-the-art performance in image (Rombach et al., 2022; Peebles & Xie, 2023) and video (Kim et al., 2024; Blattmann et al., 2023) synthesis. Despite extensive research on improving diffusion models through architecture modifications (Nichol & Dhariwal, 2021; Karras et al., 2022; Peebles & Xie, 2023), enhanced training strategies (Wang et al., 2023), and novel noise scheduling techniques (Song et al., 2020b; 2023), relatively little attention has been given to understanding their loss landscape properties.

Flat minima in various tasks. Flat minima have garnered significant attention in classification and domain generalization tasks with the view of loss landscape (Izmailov et al., 2018; Garipov et al., 2018; Li et al., 2018) and objective function (Foret et al., 2021; Kwon et al., 2021). Other tasks also adopt the flatness for their performance. In federated learning, for example, where client devices collect and train on heterogeneous datasets, ensuring robust generalization across diverse data distributions is a core challenge. Recent works have employed flat minima strategies to mitigate this data heterogeneity issue by improving robustness to distribution shift (Qu et al., 2022; Caldarola et al., 2022; Lee & Yoon, 2024) and enhance robustness of compressed neural networks (Na et al., 2022; Zhou et al., 2023).

Flat minima on diffusion models. Recent findings suggest that diffusion models inherently exhibit a flatter loss landscape (Xu et al.), yet without explaining why or how flatness relates to generalization and robustness. While prior works (Arjovsky et al., 2017; Chen et al., 2020; Ning et al., 2023c) leverage smoothness via Lipschitz continuity or gradient regularization, they do not directly examine flatness. To our knowledge, we are the first to systematically investigate how flat minima influence diffusion models, generative performance, and robustness–motivating our theoretical and empirical analysis.

6. Conclusion

We presented an in-depth investigation of flat loss surfaces in generative models, which have remained unknown despite the great success of deep generative models, analyzing how flatness influences generative modeling and demonstrating its possible impact on robustness. First, our theoretical analysis reveals that the loss flatness of the generative models enhances the robustness against the disruptions of the target data distributions or the perturbations of model parameters. These insights naturally link to the reduced exposure bias problem and the model quantization error. Based on the evaluations on CIFAR-10, LSUN Tower, and FFHQ datasets, we demonstrated that 1) SAM is suitable to seek flatter minima of diffusion models rather than SWA and EMA, 2) flat minima reduce exposure bias with a minimal gap between testing and training, and 3) a flat model keeps its generative performance even with a strong model quantization; coinciding with our theory.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Bisla, D., Wang, J., and Choromanska, A. Low-pass filtering sgd for recovering flat optima in the deep learning optimization landscape. In *International Conference on Artificial Intelligence and Statistics*, pp. 8299–8339. PMLR, 2022.

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Caldarola, D., Caputo, B., and Ciccone, M. Improving generalization in federated learning by seeking flat minima. In *European Conference on Computer Vision (ECCV)*, pp. 654–672. Springer, 2022.

Cha, J., Chun, S., Lee, K., Cho, H.-C., Park, S., Lee, Y., and Park, S. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Chen, D., Orekondy, T., and Fritz, M. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. *Advances in Neural Information Processing Systems*, 33:12673–12684, 2020.

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

et al., K. Analyzing and improving the training dynamics of diffusion models. In *CVPR*, 2024.

Foret, P., Kleiner, A., Mobahi, H., and Neyshabur, B. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2021.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D. P., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

Granziol, D., Wan, X., Albanie, S., and Roberts, S. Iterative averaging in the quest for best test error. *arXiv preprint arXiv:2003.01247*, 2020.

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. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35: 26565–26577, 2022.

Kim, K., Lee, H., Park, J., Kim, S., Lee, K., Kim, S., and Yoo, J. Hybrid video diffusion models with 2d triplane and 3d wavelet representation. In *European Conference on Computer Vision*, pp. 148–165. Springer, 2024.

Klinker, F. Exponential moving average versus moving exponential average. *Mathematische Semesterberichte*, 58:97–107, 2011.

Kwon, J., Kim, J., Park, H., and Choi, I. K. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pp. 5905–5914. PMLR, 2021.

Lee, T. and Yoon, S. W. Rethinking the flat minima searching in federated learning. In *Forty-first International Conference on Machine Learning*, 2024.

Li, H., Xu, Z., Taylor, G., Studer, C., and Goldstein, T. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018.

Li, M., Qu, T., Yao, R., Sun, W., and Moens, M.-F. Alleviating exposure bias in diffusion models through sampling with shifted time steps. *arXiv preprint arXiv:2305.15583*, 2023a.

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

Li, S., Liu, Z., Tian, J., Wang, G., Wang, Z., Jin, W., Wu, D., Tan, C., Lin, T., Liu, Y., et al. Switch ema: A free lunch for better flatness and sharpness. *arXiv preprint arXiv:2402.09240*, 2024a.

Li, T., Zhou, P., He, Z., Cheng, X., and Huang, X. Friendly sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5631–5640, June 2024b.

Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., and Keutzer, K. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17535–17545, 2023c.

Li, Y., Gong, R., Tan, X., Yang, Y., Hu, P., Zhang, Q., Yu, F., Wang, W., and Gu, S. Brecq: Pushing the limit of post-training quantization by block reconstruction. *arXiv preprint arXiv:2102.05426*, 2021.

Liu, Y., Mai, S., Cheng, M., Chen, X., Hsieh, C.-J., and You, Y. Random sharpness-aware minimization. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24543–24556. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9b79416c0dc4b09feaa169ed5cdd63d4-Paper-Conference.pdf.

Na, C., Mehta, S. V., and Strubell, E. Train flat, then compress: Sharpness-aware minimization learns more compressible models. *arXiv preprint arXiv:2205.12694*, 2022.

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

Ning, M., Li, M., Su, J., Salah, A. A., and Ertugrul, I. O. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023a.

Ning, M., Li, M., Su, J., Salah, A. A., and Ertugrul, I. O. Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321*, 2023b.

Ning, M., Sangineto, E., Porrello, A., Calderara, S., and Cucchiara, R. Input perturbation reduces exposure bias in diffusion models. *arXiv preprint arXiv:2301.11706*, 2023c.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Qu, Z., Li, X., Duan, R., Liu, Y., Tang, B., and Lu, Z. Generalized federated learning via sharpness aware minimization. In *International Conference on Machine Learning (ICML)*, pp. 18250–18280. PMLR, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. 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.

SHI, G., CHEN, J., Zhang, W., Zhan, L.-M., and Wu, X.-M. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 6747–6761. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/357cfba15668cc2e1e73111e09d54383-Paper.pdf.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y. and Ermon, S. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020.

Song, Y., Garg, S., Shi, J., and Ermon, S. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in artificial intelligence*, pp. 574–584. PMLR, 2020a.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. 2023.

Wang, Z., Jiang, Y., Zheng, H., Wang, P., He, P., Wang, Z., Chen, W., Zhou, M., et al. Patch diffusion: Faster and more data-efficient training of diffusion models. *Advances in neural information processing systems*, 36: 72137–72154, 2023.

Wu, J., Wang, H., Shang, Y., Shah, M., and Yan, Y. Ptq4dit: Post-training quantization for diffusion transformers. *Advances in Neural Information Processing Systems*, 37: 62732–62755, 2025.

Xu, T., Mi, P., Wang, R., and Chen, Y. Why diffusion models are stable and how to make them faster: An empirical investigation and optimization.

Zhou, Y., Yang, Y., Chang, A., and Mahoney, M. W. A three-regime model of network pruning. In *International Conference on Machine Learning*, pp. 42790–42809. PMLR, 2023.

A. Supplementary material

B. Mathematical claims and proofs

For the main claims, we follow $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, t, p_t) := ||s_{\boldsymbol{\theta}}(\mathbf{x}, t) - \nabla_{\mathbf{x}} \log p_t(\mathbf{x})||_2^2$, while dropping the timestep t without loss of generality. Our mathematical claims are valid for all timesteps.

Definition 1. *(Δ-flat minima) Let us consider a SGM with loss function $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, p)$. A minimum $\boldsymbol{\theta}^*$ is Δ-flat minima when the following constraints are hold:*

$$\forall\, \delta \in \mathbb{R}^{d\times m} \text{ s.t. } \|\delta\|_2 \leq \Delta,\;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^* + \delta, p) = l^*$$
$$\exists\, \delta \in \mathbb{R}^{d\times m} \text{ s.t. } \|\delta\|_2 > \Delta,\;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^* + \delta, p) > l^*,$$

where $l^* := \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, p)$ and $\Delta \in \mathbb{R}^+$.[4]

Definition 2. *($\mathcal{E}$-distribution gap robustness) A minimum $\boldsymbol{\theta}^*$ is $\mathcal{E}$-distribution gap robust when the following constraints are hold:*

$$\forall\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p||\hat{p}) \leq \mathcal{E},\;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, \hat{p}) = l^*$$
$$\exists\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p||\hat{p}) > \mathcal{E},\;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, \hat{p}) > l^*,$$

where $D(\cdot||\cdot)$ is the divergence between two probability density functions, $\hat{p}$ is the perturbed prior distribution of $\mathbf{x}$, and $\mathcal{E}$ is a positive real number.

Theorem 1. *(A perturbed distribution) For a given prior distribution of $p(\mathbf{x})$ and the δ-perturbed minimum, i.e., $\boldsymbol{\theta} + \delta$, the following $\hat{p}(\mathbf{x})$ satisfies the $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta} + \delta, p) = \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \hat{p})$:*

$$\hat{p}(\mathbf{x}) = e^{-I(\mathbf{x},\delta)} p(\mathbf{x}), \tag{19}$$

where $I(\mathbf{x}, \delta) := \frac{1}{2}\mathbf{x}^T(\delta\mathbf{W}^T)\mathbf{x} + \mathbf{x}^T\delta(\mathbf{U}^T\mathbf{e}) + C$, and $C \in \mathbb{R}$ is set to satisfy $\int_{\mathbb{R}^d} e^{-I(\mathbf{x},\delta)} p(\mathbf{x}) d\mathbf{x} = 1$.

Proof. By following (Li et al., 2023b), we formulate the score model $s_{\boldsymbol{\theta}}(\cdot, \cdot)$ as a random feature model:

$$s_{\boldsymbol{\theta}}(\mathbf{x}, t) := \frac{1}{m}\boldsymbol{\theta}\sigma(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}_t) \tag{20}$$

where $\mathbf{x} \in \mathbb{R}^{d\times 1}$, $\boldsymbol{\theta} \in \mathbb{R}^{d\times m}$, $\mathbf{W} \in \mathbb{R}^{d\times m}$, $\mathbf{U} \in \mathbb{R}^{d_e\times m}$, $\mathbf{e}_t \in \mathbb{R}^{d_e\times 1}$, and d, m, d_e are positive integers.

The score matching loss objective is defined as

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, p) := ||s_{\boldsymbol{\theta}}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x})||_2^2, \tag{21}$$

For the perturbation $\delta \in \mathbb{R}^{d\times m}$ in the diffusion model parameters $\boldsymbol{\theta}$, the perturbed loss value becomes:

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\theta} + \delta, p) := ||s_{\boldsymbol{\theta}+\delta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x})||_2^2. \tag{22}$$

$$s_{\boldsymbol{\theta}+\delta}(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{23}$$
$$= \frac{1}{m}(\boldsymbol{\theta} + \delta)\sigma(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{24}$$
$$= \frac{1}{m}\boldsymbol{\theta}\sigma(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}) + \frac{1}{m}\delta\sigma(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{25}$$

[4] $\forall$ means 'for all,' $\exists$ means 'there exists,' and $\mathbb{R}^+$ indicates the set of positive real numbers

Let us focus on the second and third terms with the assumptions of the positive outputs for the activation function:

$$\frac{1}{m}\delta(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{26}$$

Here, let us define $I(\mathbf{x})$ as a function of $\mathbf{x}$, whose derivative is the first term of the previous equation:

$$\frac{\partial I(\mathbf{x})}{\partial \mathbf{x}} = \frac{1}{m}\delta(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}) \tag{27}$$

Based on it, $I(\mathbf{x}) \in \mathbb{R}$ is

$$I(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T(\delta\mathbf{W}^T)\mathbf{x} + \mathbf{x}^T\delta(\mathbf{U}^T\mathbf{e}) + C, \tag{28}$$

with the assumption $\delta\mathbf{W}^T$ is symmetric and where C is a constant real number.

$$\frac{1}{m}\delta(\mathbf{W}^T\mathbf{x} + \mathbf{U}^T\mathbf{e}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{29}$$

$$= \nabla_{\mathbf{x}} I(\mathbf{x}) - \nabla_{\mathbf{x}} \log p(\mathbf{x}) \tag{30}$$

$$= -\nabla \log \left(e^{-I(\mathbf{x})} p(\mathbf{x})\right) \tag{31}$$

When C is the real number that satisfies the following condition for the function I with C:

$$\int_{\mathbb{R}^d} e^{-I(\mathbf{x})} p(\mathbf{x}) = 1, \tag{32}$$

then we can define $\hat{p}(\mathbf{x})$ to be a perturbed PDF of inputs:

$$\hat{p}(\mathbf{x}) := e^{-I(\mathbf{x})} p(\mathbf{x}) \tag{33}$$

$$= \exp\Big\{ -\frac{1}{2m}\mathbf{x}^T(\delta\mathbf{W}^T)\mathbf{x} - \frac{1}{m}\mathbf{x}^T\delta(\mathbf{U}^T\mathbf{e}) - C^*)\Big\} p(\mathbf{x}) \tag{34}$$

□

Corollary 1. *(Diffusion version of **Theorem 1**) For a given prior Gaussian distribution of noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ and the δ-perturbed minimum, i.e., $\boldsymbol{\theta} + \delta$, the following $\hat{\epsilon}$ satisfies the $\mathcal{L}(\mathbf{x}; \boldsymbol{\theta} + \delta, p) = \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}, \hat{p})$:*

$$\hat{\epsilon} = e^{-I(\mathbf{x},\delta)}\epsilon = \mathcal{N}(\boldsymbol{\mu}_\delta, \Sigma_\delta), \tag{35}$$

where $\Sigma_\delta := \left(\mathbf{I} + \frac{\delta_w}{m}\right)^{-1}$, $\boldsymbol{\mu}_\delta := \frac{1}{m}\Sigma_\delta\delta_u$.

Proof. We provide the theoretical link that the model satisfying the $\mathcal{E}$-flat in Theorem **1** is also robust to distribution shift caused by the exposure bias.

Before that, we introduce the notations:

- $p(\mathbf{x})$: the true data distribution that is known in the training process.
- $\hat{p}(\mathbf{x})$: the perturbed distribution that caused by the ϵ model perturbation in Eq. (33).

When we train the diffusion model, we add the noise $\boldsymbol{\epsilon}$ in the forward process and want the diffusion model to predict the $\boldsymbol{\epsilon}$ in the reverse process where $\boldsymbol{\epsilon}$ follows the normal Gaussian distribution, i.e., $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. Therefore, the distribution that

the model trains is the normal Gaussian, and we can define the perturbed Gaussian distribution as follows:

$$\hat{\epsilon}_t(\mathbf{x}) := e^{-I(\mathbf{x},\delta)}\epsilon_t \tag{36}$$

$$= \exp\left(-\frac{1}{2m}\mathbf{x}^T\delta_w\mathbf{x} - \frac{1}{m}\mathbf{x}^T\delta_u - C\right)\cdot\frac{1}{\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right), \tag{37}$$

$$\text{where } \delta_w := \delta\mathbf{W}^T, \delta_u := \delta\mathbf{U}^T\mathbf{e}, \text{ and } \epsilon \sim \mathcal{N}(0,\mathbf{I})$$

$$= \frac{1}{\sqrt{(2\pi)^d}}\exp\left(-\frac{1}{2}\mathbf{x}^T(\mathbf{I}+\frac{1}{m}\delta_w)\mathbf{x} - \frac{1}{m}\mathbf{x}^T\delta_u - C\right) \tag{38}$$

$$= \frac{1}{C'}\exp\left(-\frac{1}{2}\left\{\mathbf{x}^T(\mathbf{I}+\frac{\delta_w}{m})\mathbf{x} + \frac{2}{m}\mathbf{x}^T\delta_u\right\}\right), C' := \frac{1}{(\sqrt{2\pi})^d}\exp(-C) \tag{39}$$

$$= \frac{1}{\sqrt{(2\pi)^d|\Sigma_\delta|}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_\delta)^T{\Sigma_\delta}^{-1}(\mathbf{x}-\boldsymbol{\mu}_\delta)\right), \text{ where } \Sigma_\delta := \left(\mathbf{I}+\frac{\delta_w}{m}\right)^{-1}, \text{ and } \boldsymbol{\mu}_\delta := \frac{1}{m}\Sigma_\delta\delta_u \tag{40}$$

Because the $\hat{\epsilon}(\mathbf{x})$ is also the Gaussian distribution, we present the KL Divergence between $p(\mathbf{x})$ and $p^*_\epsilon(\mathbf{x})$ as follows:

$$D_{KL}(\hat{\epsilon}\|\epsilon) = \frac{1}{2}\left[\log\frac{|\Sigma_\delta|}{|\Sigma|} - d + tr({\Sigma_\delta}^{-1}\Sigma^{-1}) + {\boldsymbol{\mu}_\delta}^T{\Sigma_\delta}^{-1}\boldsymbol{\mu}_\delta\right] \tag{41}$$

$$= \frac{1}{2}\left[\log|\Sigma_\delta| - d + tr({\Sigma_\delta}^{-1}) + {\boldsymbol{\mu}_\delta}^T{\Sigma_\delta}^{-1}\boldsymbol{\mu}_\delta\right], \because \Sigma = \mathbf{I} \tag{42}$$

□

Definition 3. *(A set of perturbed distribution) For a given distribution of $p(\mathbf{x})$, a set of distributions $\hat{\mathcal{P}}(\mathbf{x};p,\Delta)$ is defined as the set of perturbed distributions $\hat{p}(\mathbf{x})$:*

$$\hat{\mathcal{P}}(\mathbf{x};p,\Delta) := \{e^{-I(\mathbf{x},\delta)}p(\mathbf{x}) \mid \|\delta\|_2 \le \Delta\}. \tag{43}$$

Proposition 1. *(A link from Δ-flatness to $\hat{\mathcal{P}}$) A Δ-flat minimum $\boldsymbol{\theta}^*$ achieves the flat loss values for all distributions sampled from the set of perturbed distribution:*

$$\forall\, p \sim \hat{\mathcal{P}}(\mathbf{x};p,\Delta),\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,p) = l^* \tag{44}$$

$$\exists\, p \nsim \hat{\mathcal{P}}(\mathbf{x};p,\Delta),\ \ \mathcal{L}(\mathbf{x};\boldsymbol{\theta}^*,p) > l^*. \tag{45}$$

Theorem 2. *(A link from Δ-flatness to $\mathcal{E}$-gap robustness) A Δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:*

$$\mathcal{E} \le \underset{\hat{p}\sim\hat{\mathcal{P}}(\mathbf{x};p,\Delta)}{\arg\max}\ D(p||\hat{p}). \tag{46}$$

Corollary 2. *(Diffusion version of **Theorem 2**) For a diffusion model, a Δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:*

$$\mathcal{E} \le \underset{\|\delta\|_2\le\Delta}{\arg\max}\frac{1}{2}\left[\log|\Sigma_\delta| - d + tr(\Sigma_\delta^{-1}) + {\boldsymbol{\mu}_\delta}^T{\Sigma_\delta}^{-1}\boldsymbol{\mu}_\delta\right] \tag{47}$$

$$\le \frac{1}{2}\left[\sum_i^d\log\frac{m}{m+\sigma_i} + \sum_i^d\frac{\sigma_i}{m} + \frac{\sigma_d}{m^2}\|\mathbf{U}^T\mathbf{e}\|_2^2\Delta^2\right], \tag{48}$$

where σ_i is an eigenvalue of Σ_δ^{-1} with the increasing order of $\sigma_1 \le \sigma_2 \le \ldots \le \sigma_d$.

Proof. From the definition **2**, a minimum θ^* hold following:

$$\forall\, \hat{p}(\mathbf{x}) \text{ s.t. } D(p||\hat{p}) \leq \mathcal{E},\;\; \mathcal{L}(\mathbf{x}; \boldsymbol{\theta}^*, \hat{p}) = l^*.$$

Let σ_i is an eigenvalue of Σ_δ^{-1} with the increasing order $\sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_d$. Then, the Diffusion version of Theorem **2** is represented as follows:

$$\mathcal{E} \leq \underset{\|\delta\|_2 \leq \Delta}{\arg\max} \frac{1}{2}\left[\log|\Sigma_\delta| - d + tr(\Sigma_\delta^{-1}) + {\boldsymbol{\mu}_\delta}^T {\Sigma_\delta}^{-1} \boldsymbol{\mu}_\delta\right] \tag{49}$$

$$\leq \frac{1}{2}\left[\sum_i^d \log\frac{m}{m+\sigma_i} - d + \sum_i^d \left(1 + \frac{\sigma_i}{m}\right) + \frac{\sigma_d}{m^2}\|\mathbf{U}^T\mathbf{e}\|_2^2 \Delta^2\right. \tag{50}$$

$$\leq \frac{1}{2}\left[\sum_i^d \log\frac{m}{m+\sigma_i} + \sum_i^d \frac{\sigma_i}{m} + \frac{\sigma_d}{m^2}\|\mathbf{U}^T\mathbf{e}\|_2^2 \Delta^2\right], \tag{51}$$

where inequality Eq. (51) holds when $\boldsymbol{\mu}_\delta$ is the eigenvector satisfying $\Sigma_\delta^{-1}\boldsymbol{\mu}_\delta = \sigma_d \boldsymbol{\mu}_\delta$. □

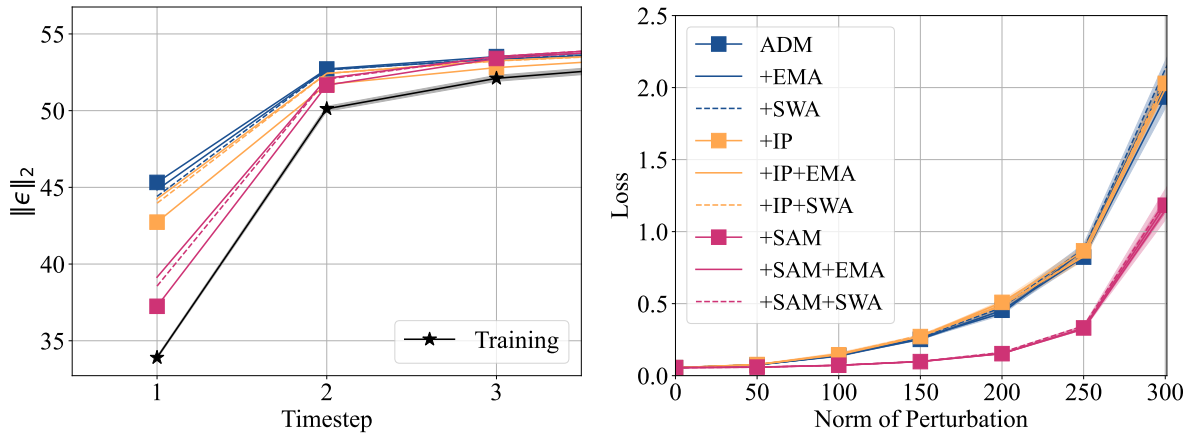


Figure 7. Additional results for CIFAR-10. We measure the L2 norm of predicted noise and loss plots under perturbation for all algorithms including `+IP+EMA`, `+IP+SWA`, `+SAM+EMA`, `+SAM+SWA`.

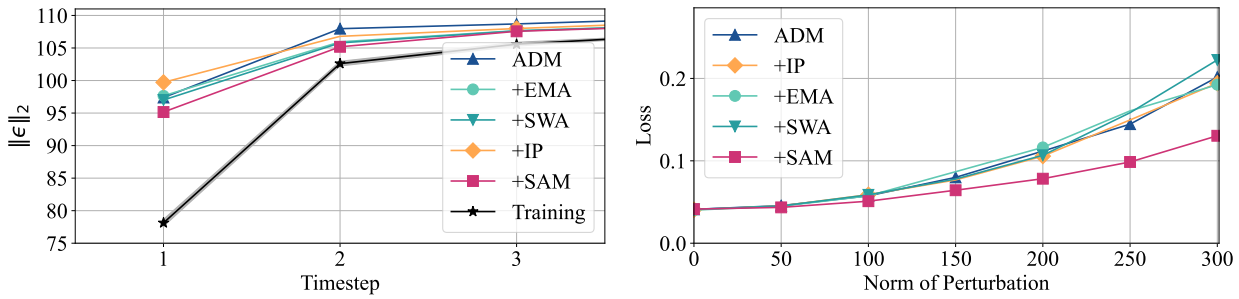


Figure 8. (Left)L2 norm of the predicted noise for LSUN Tower dataset, (Right) Loss plots under perturbation for LSUN Tower.

LPF ↓	`w/o`	`+EMA`	`+SWA`
ADM	0.091	0.090	0.092
`+IP`	0.089	0.092	0.097
`+SAM`	**0.072**	**0.070**	**0.071**

↓: a lower value is preferred.

Table 5. Flatness measure on LSUN Tower. We calculate the loss with the perturbed model with Gaussian noise. Lower values indicate a flatter loss landscape.

C. Additional experimental results

In Fig. 7, we report additional results for `+IP+EMA`, `+IP+SWA`, `+SAM+EMA`, `+SAM+SWA` for CIFAR-10. We observe that ADM already possesses a certain level of flatness supporting `+SWA` and `+EMA` fail to induce additional flatness. We also report L2 norm of predicted noise loss plots under perturbation in Fig. 8 and LPF flatness in Table. 5 for LSUN Tower dataset. It coincides with the result of CIFAR-10 that `+SAM` induces the lower exposure bias and flatter minima.