Understanding Flatness in Generative Models: Its Role and Benefits

Taehwan Lee*, Kyeongkook Seo*, Jaejun Yoo**, Sung Whan Yoon**

Ulan National Institute of Science and Technology (UNIST)
South Korea

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Outline

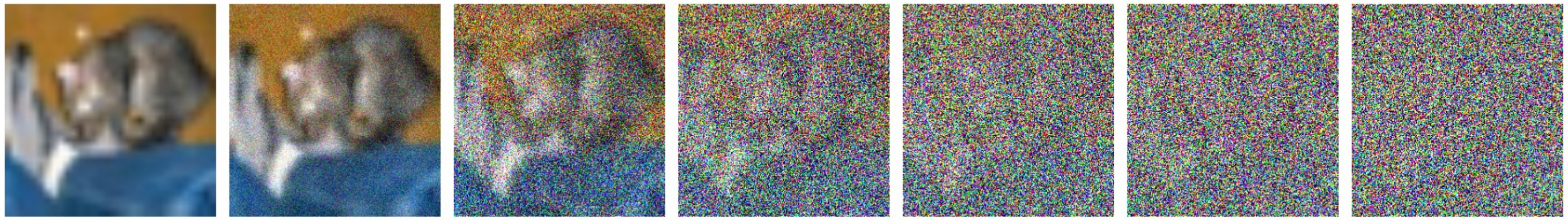
1. Preliminaries
2. Motivation
3. Theoretical Analysis
4. Experimental Results

Preliminaries – Diffusion Models

Diffusion Process

forward SDE

x_0 $d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)dw$ x_T



x_0 $d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_{\mathbf{x}}(\mathbf{x})]dt + g(t)d\bar{w}$ x_T

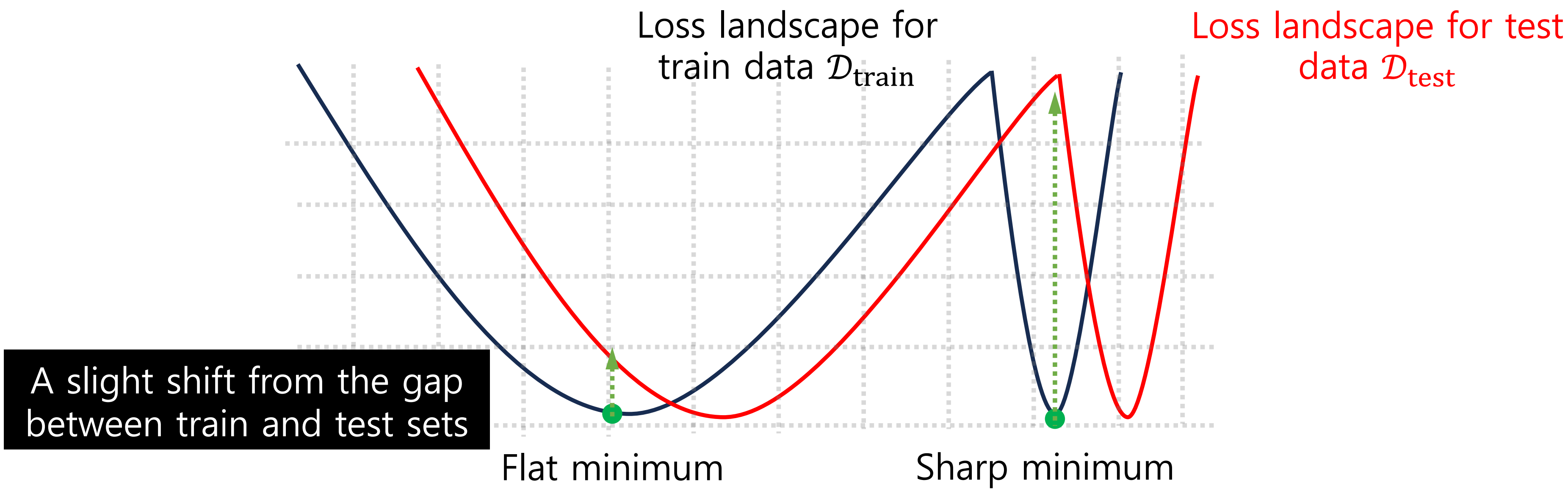
backward SDE

Score Matching Objective

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

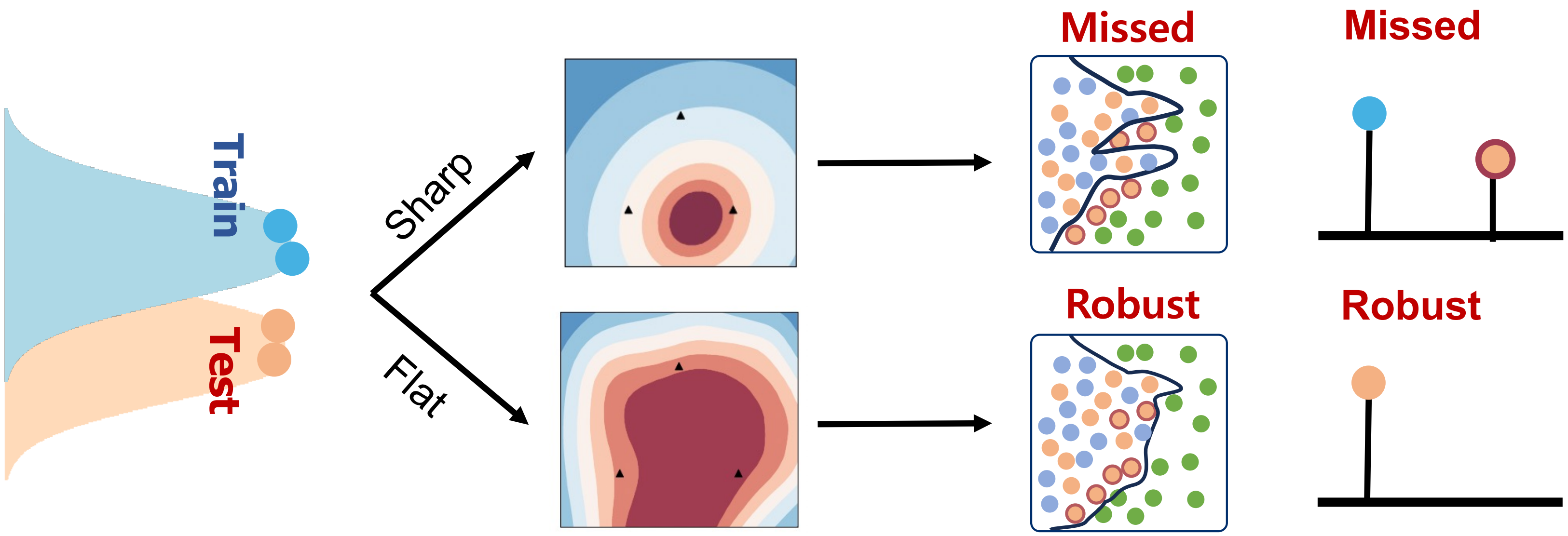
Preliminaries – Flat Minima Searching

Loss values at flat minima change
more **smoothly** than sharp ones.



Motivation – Flatness in Classification

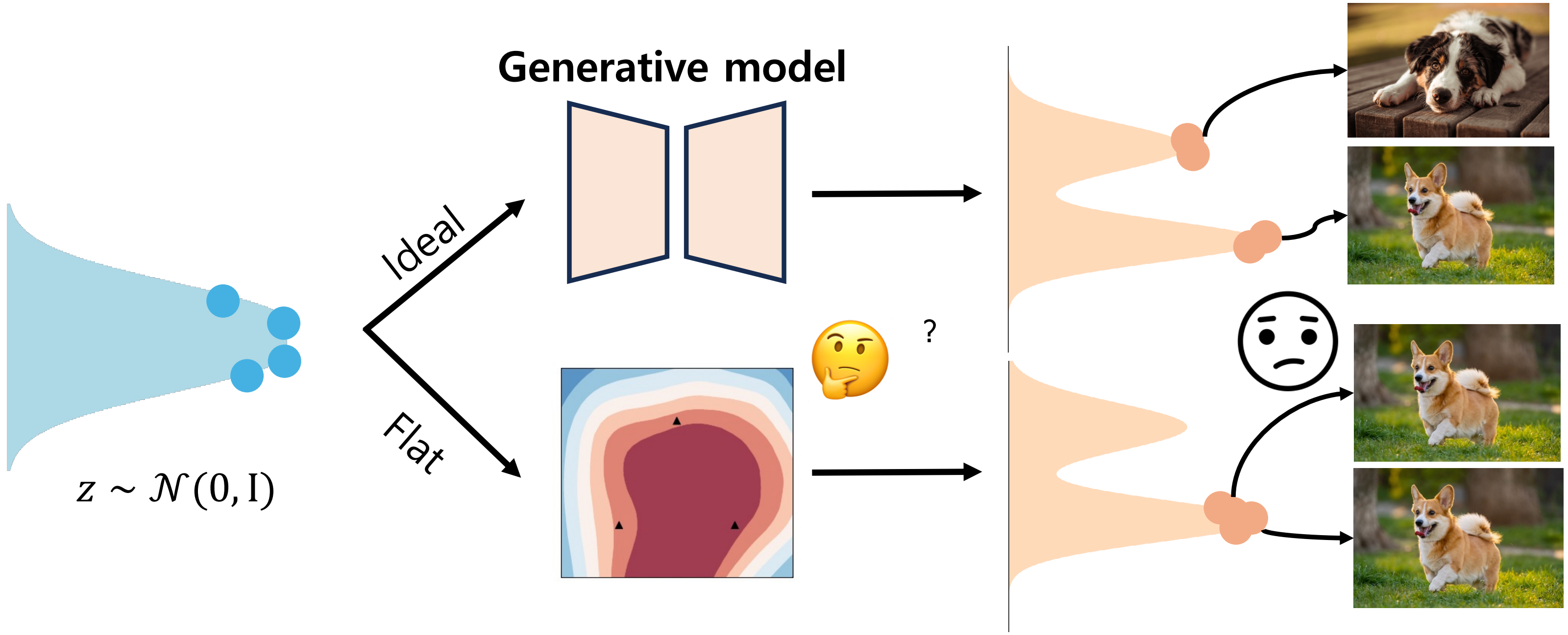
Sharp minima are prone to unseen input distribution.
Flat minima remain consistent under distribution shifts.



Motivation – Flatness in Generative Model

Flat minima in the classification task show robustness to distribution shift.

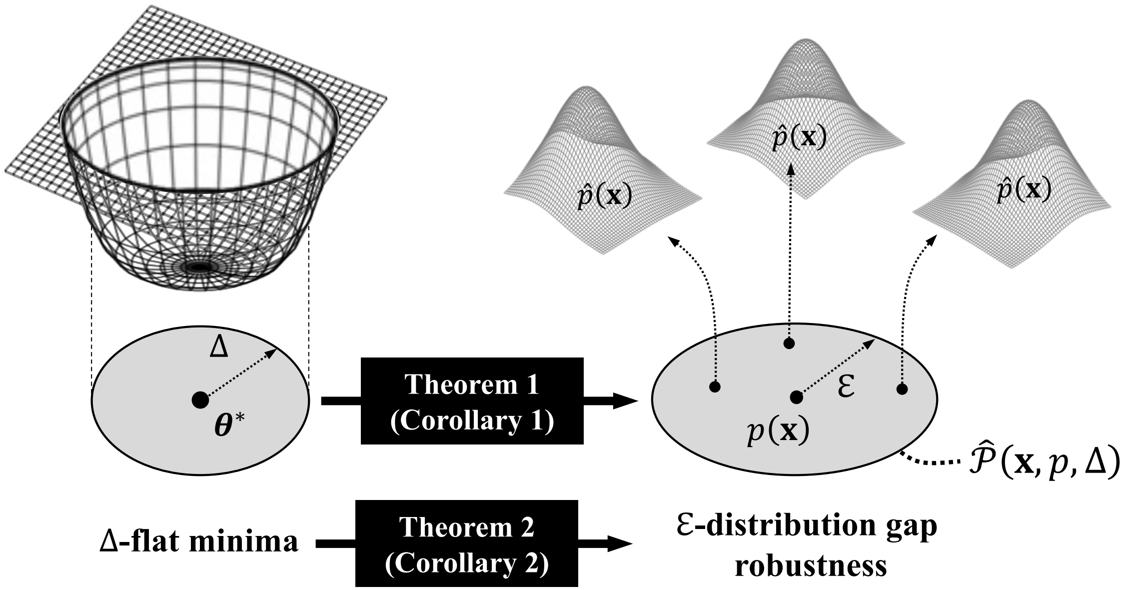
Then, what happens if the generative model is flat?



Research Question

What is the role of the flatness of the Generative model?

Theoretical Results – Overview



- **Theorem 1:** bridges parameter perturbation to the data space.
- **Theorem 2:** links flatness to robustness in distribution space.

Theoretical Results – Theorem 1

Theorem 1. A perturbed distribution

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

$$\hat{p}(\mathbf{x}) = \exp\big(-I(\mathbf{x}, \boldsymbol{\delta})\big)\, p(\mathbf{x})$$

Remark

Perturbations in $\boldsymbol{\theta}$-space translate to scaled pdfs in $\mathbf{x}$-space and flat minima enable the generative model to perform well on them.

Theoretical Results – Theorem 2

Theorem 2. Link from flatness to distribution gap

A δ-flat minimum achieves $\mathcal{E}$-distribution gap robustness, such that $\mathcal{E}$ is upper-bounded as follows:

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

Remark

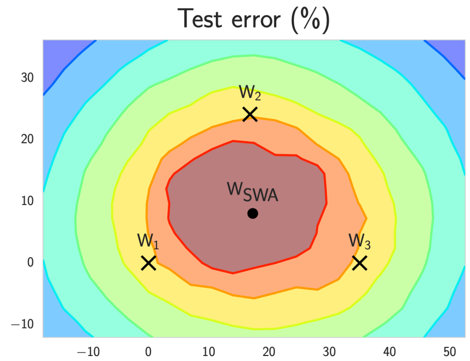
Flat generative model remains robust up to the maximum KL-divergence between p and $\hat{p}$, implying that flatter generative models achieve broader coverage.

Experimental Results - Baselines

- Explicit method [SAM'21]
  - SAM adopts the sharpness in the optimization objective [SAM'21]
  - $\left[\max_{\|\epsilon\|_2 \leq \rho} \mathcal{L}(w+\epsilon) - \mathcal{L}(w)\right] + \mathcal{L}(w) + h(\|w\|_2^2/\rho^2)$

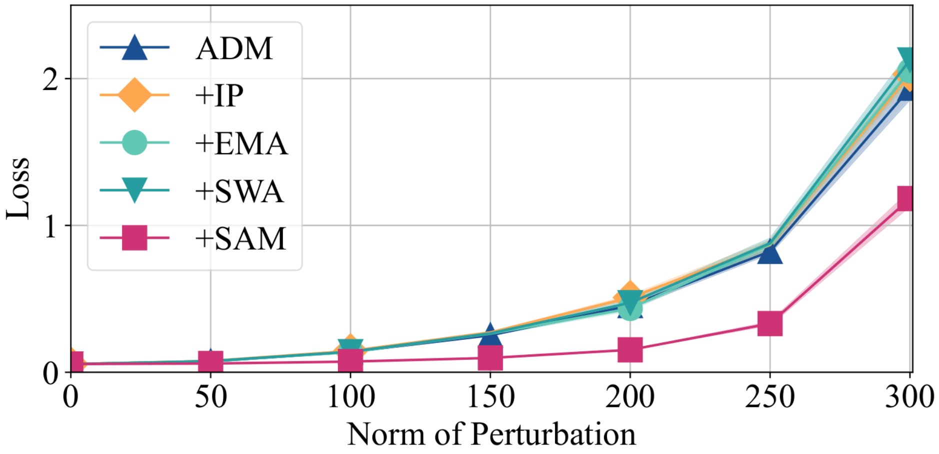
Sharpness　　Loss at minima　　L2 Reg.

- Implicit method [SWA'18, EMA'24]
  - Averaging model parameters leads flat minima
  - W_1, W_2, W_3 : trained model with SGD.
  - W_{SWA} : Averaged model of W_1, W_2, W_3.
  - Finding flat minima results in better performance.



[SAM'21] P. Foret et al., "SHARPNESS-AWARE MINIMIZATION FOR EFFICIENTLY IMPROVING GENERALIZATION," ICLR 2021.
[EMA'24] Li, Siyuan, et al. "Switch ema: A free lunch for better flatness and sharpness." arXiv, 2024.
[SWA'18] Izmailov, Pavel, et al. "Averaging weights leads to wider optima and better generalization." UAI, 2018.

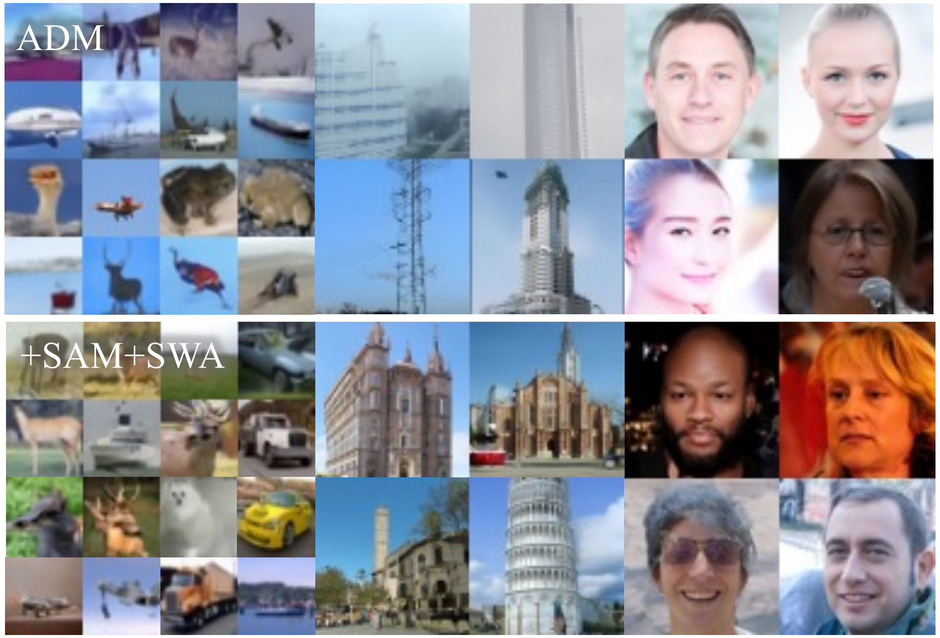
Experimental Results – Flatness



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**

While **+SAM** finds a flatter loss landscape **explicitly**, empirical methods (**+EMA, +SWA**) shows less impact.

Experimental Results – Full Precision



FID Score	CIFAR10		LSUN-Tower		FFHQ	
	T=20	T=100	T=20	T=100	T=20	T=100
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
+EMA	9.10	3.46	7.66	2.43	11.72	4.00
+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
+EMA	**7.00**	3.18	6.66	2.30	**11.41**	5.04
+SWA	7.27	**2.96**	**6.50**	**2.27**	12.15	4.17

SAM (+EMA, +SWA) achieves comparable or better FID score.

Experimental Results – Low Precision

FID Score	T=20		T=100	
	32 bit	8 bit	32 bit	8 bit
ADM	34.47	48.02 (+13.65)	8.80	12.78 (+3.98)
+EMA	10.63	20.65 (+10.02)	4.06	7.36 (+3.3)
+SAM	9.01	8.94 (-0.07)	3.83	4.02 (+0.19)
+SAM+EMA	7.00	7.20 (+0.2)	3.18	3.12 (-0.06)

SAM (+EMA) shows robustness to **8-bit quantization.**
SAM raises robustness to quantization,
which is **essential for model deployment.**

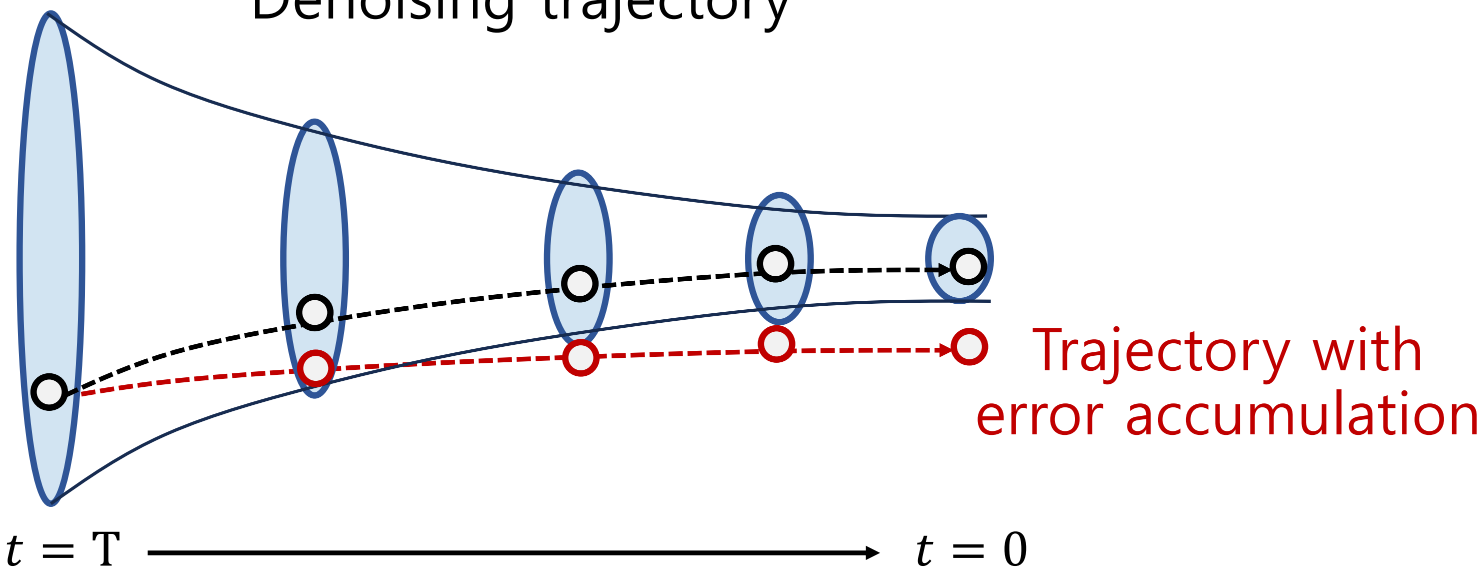
T: sampling steps

Experimental Results – Low Precision

While ADM and +IP collapse in **4-bit quantization**, **SAM** maintains the image generation performance.

Experimental Results – Exposure Bias

Denoising trajectory

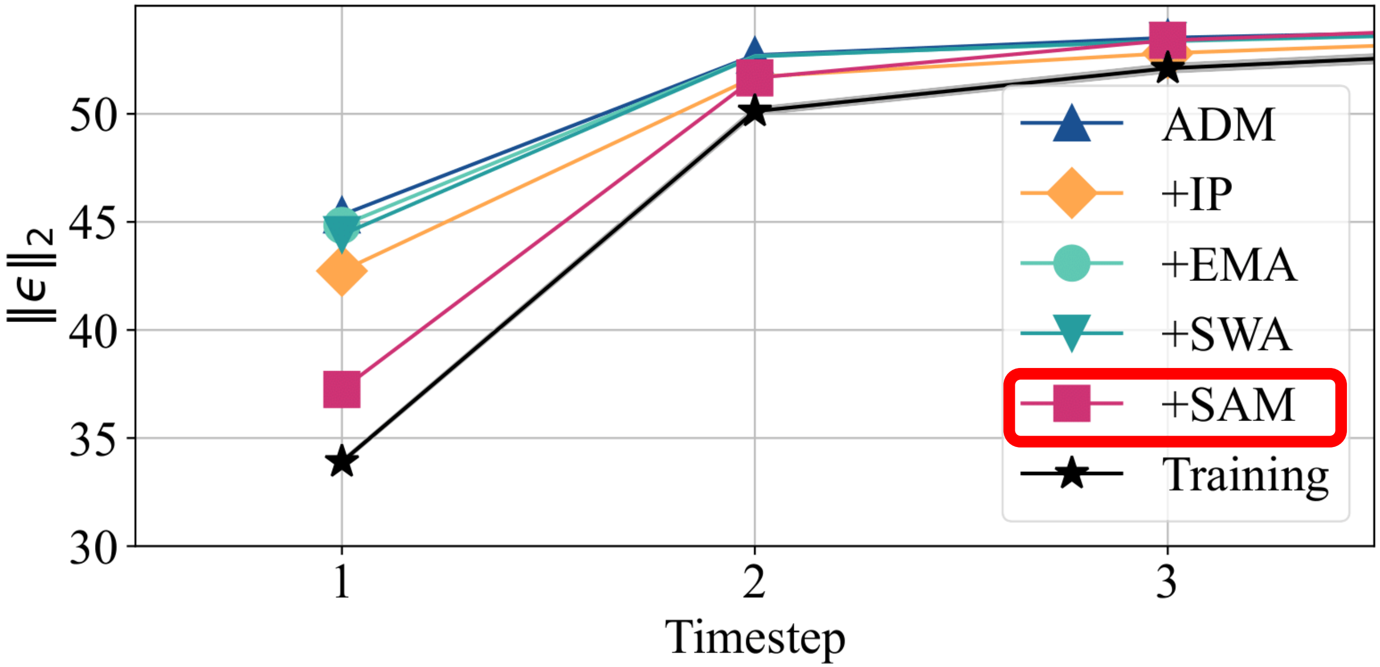


The iterative process in DMs results in error accumulation.

The accumulation of errors is referred to as exposure bias. [IP'23]

[IP'23] Ning, Mang, et al. "Input perturbation reduces exposure bias in diffusion models." ICML, 2023.

Experimental Results – Exposure Bias



Flat minima show robustness to **Exposure bias**, where **+SAM** shows closer behavior with **Training.**

THANK YOU

Our paper will be presented in the poster session at Exhibit Hall I #461

on Tuesday, Oct. 21st, at 11:45 a.m. ~ 1:45 p.m.

Please visit our poster booth and have a discussion.