Utrecht University

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

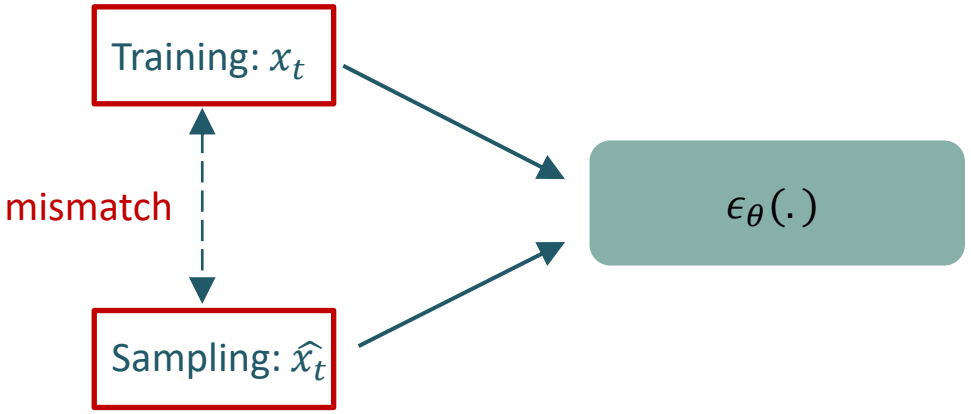
Input Perturbation Reduces Exposure Bias in Diffusion Models

Mang Ning, Enver Sangineto, Angelo Porrello, Simone Calderara, Rita Cucchiara

1. Introduction

Exposure Bias

Training: $x_t \rightarrow \epsilon_\theta(.) \rightarrow x_{t-1}$



Sampling: $\widehat{x_t} \rightarrow \epsilon_\theta(.) \rightarrow \widehat{x_{t-1}} \ldots \rightarrow \epsilon_\theta(.) \rightarrow \widehat{x_0}$

the network alwasy see x_t duiring training

the network alwasy see $\widehat{x_t}$ duiring inference

2. Exposure Bias in DDPM

Compare the difference between $q(x_0)$ and $p(\hat{x}_0)$ using FID on ADM [1]

For t in range(1, 1000):

repeat:

$$x_0 \sim q(x_0) \rightarrow x_t \sim q(x_t|x_0) \rightarrow p_\theta(x_{t-1}|x_t) \rightarrow \hat{x}_0$$

Until 50k $\hat{x}_0$

Table 1. An empirical estimate of the exposure bias on ImageNet 32×32.

Model	Number of reverse diffusion steps				
	100	300	500	700	1,000
ADM	0.983	1.808	2.587	3.105	3.544
ADM-IP (ours)	0.972	1.594	2.198	2.539	2.742

[1] Dhariwal, Prafulla, and Alexander Nichol. "Diffusion models beat gans on image synthesis." *NeuIPS* (2021)

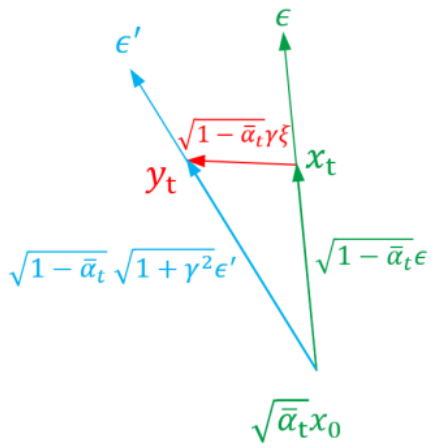
3. Method – Input Perturbation

DDPM [2]

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \qquad (4)$$

Algorithm 1 DDPM Standard Training

1: **repeat**
2: $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), t \sim \mathbb{U}(\{1, ..., T\}), \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
3: compute $\boldsymbol{x}_t$ using Eq. 4
4: take a gradient descent step on $\nabla_{\boldsymbol{\theta}}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)||^2$
5: **until** converged



	Input	Target
DDPM	x_t	ϵ
DDPM-IP	y_t	ϵ
DDPM-y	y_t	ϵ'

where $\epsilon \sim N(0, I)$ and $\epsilon' \sim N(0, I)$

Figure 1. The inputs and the prediction targets are different in vanilla DDPM, DDPM-IP and DDPM-y.

Our DDPM-IP

$$\boldsymbol{y}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}(\boldsymbol{\epsilon} + \gamma_t\boldsymbol{\xi}). \qquad (6)$$

Algorithm 3 DDPM-IP: Training with input perturbation

1: **repeat**
2: $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), t \sim \mathbb{U}(\{1, ..., T\})$
3: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$
4: compute $\boldsymbol{y}_t$ using Eq. 6
5: take a gradient descent step on $\nabla_{\boldsymbol{\theta}}||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{y}_t, t)||^2$
6: **until** converged

[2] Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *NeuIPS* (2020).

3. Method – Explicit Lipshitz Continuous

Gradient Penalty

$$L_{GP}(\boldsymbol{\theta}) = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)||^2 + \lambda_{GP} \left\| \frac{\partial \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)}{\partial \boldsymbol{x}} \right\|_F^2$$

Weight Decay

$$L_{WD}(\boldsymbol{\theta}) = ||\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)||^2 + \lambda_{WD}||\boldsymbol{\theta}||^2$$

Input Perturbation works better

Table 2. Comparison of different regularization methods. All the models are tested using $T = 1,000$ sampling steps.

Model	CIFAR10 32×32	
	FID	sFID
ADM (baseline)	2.99	4.76
ADM-GP	2.80	4.41
ADM-WD	2.82	4.61
ADM-IP	**2.76**	**4.05**

3. Main Results on DDPM (ADM)

ADM-IP:

- Significantly better FID
- Faster training

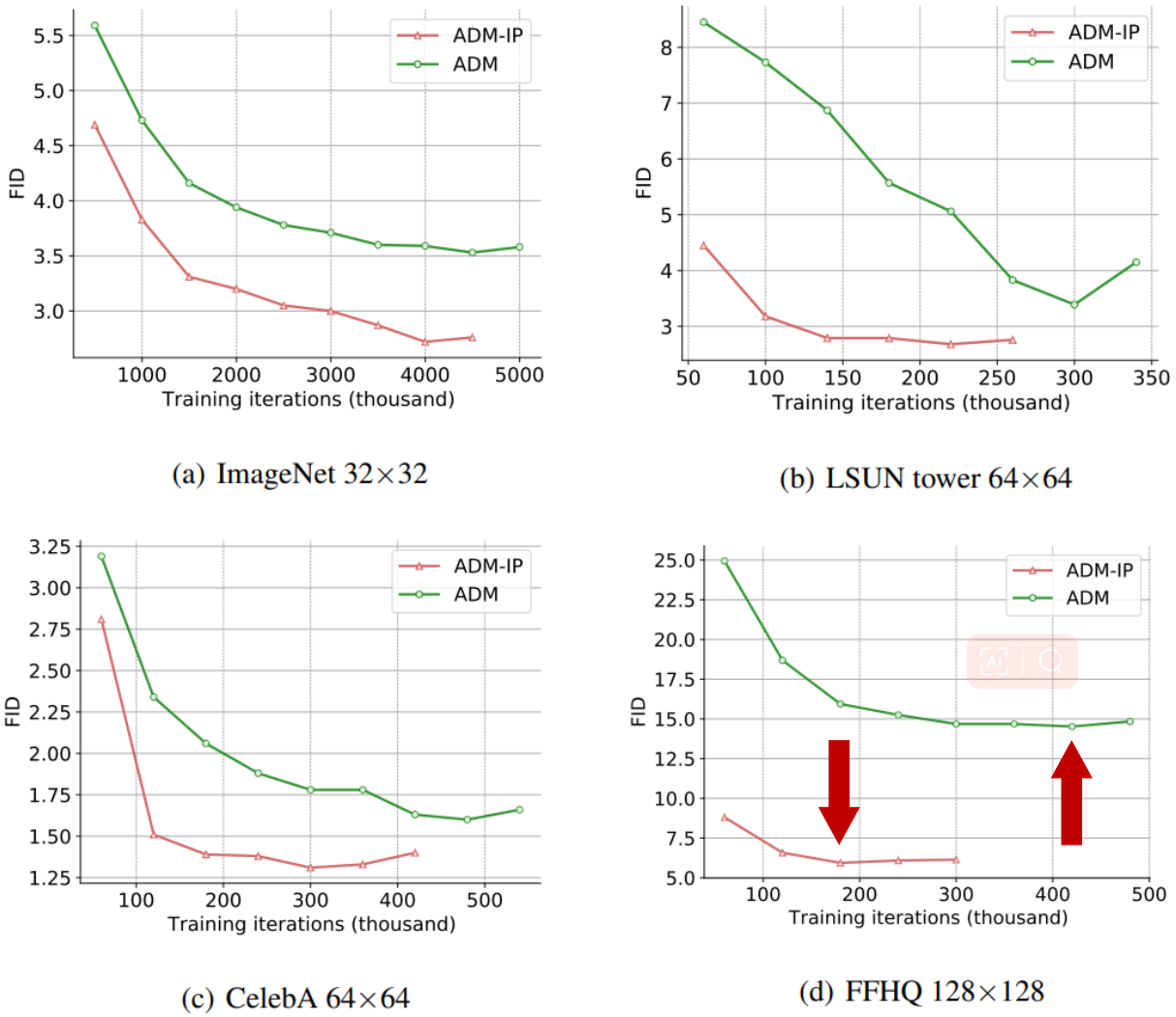


Figure 3. FID scores with respect to the number of training iterations. Each FID value is computed using $T' = 1,000$ inference sampling steps, except for the FFHQ dataset, for which we used $T' = 100$.

3. Results in ADM & DDIM

ADM-IP:

- Faster sampling

Table 4. ADM-IP training and testing acceleration. Note that, for a single training iteration, ADM and ADM-IP take exactly the same amount of time, and the same is true for a single sampling step.

Dataset	Model	Training iterations	Sampling steps	FID
CIFAR10 32×32	ADM	500K	1,000	2.99
	ADM-IP	460K	80	2.93
ImageNet 32×32	ADM	4500K	1,000	3.53
	ADM-IP	4000K	80	3.50
LSUN tower 64×64	ADM	300K	1,000	3.39
	ADM-IP	220K	60	3.31
CelabA 64×64	ADM	480K	1,000	1.60
	ADM-IP	300K	200	1.53
FFHQ 128×128	ADM	420K	1,000	9.65
	ADM-IP	180K	60	8.72

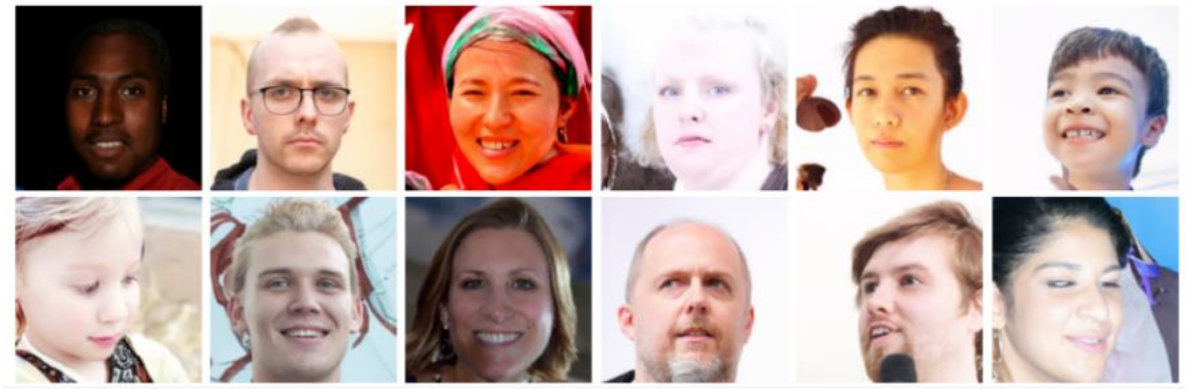
DDIM-IP:

- less sampling steps, more FID improvement

Table 5. CIFAR10: Comparison between DDIM and DDIM-IP using models trained with $T = 1,000$ sampling steps and tested with $T' \leq T$ steps.

η	Model	Sampling steps (T')				
		10	20	50	100	1,000
0	DDIM	14.21	7.50	5.17	4.66	4.29
	DDIM-IP	**10.54**	**5.70**	**4.66**	**4.52**	**4.27**
0.5	DDIM	17.24	8.87	5.59	4.88	4.45
	DDIM-IP	**10.06**	**5.53**	**3.95**	**3.66**	**3.56**

3. Qualitative Results



(a) Samples generated by ADM trained on FFHQ 128×128 (FID 9.65)

(b) Samples generated by ADM-IP trained on FFHQ 128×128 (FID 2.98)

Acknowledgement