Input Perturbation Reduces Exposure Bias in Diffusion Models

Mang Ning[1 2], Enver Sangineto[2], Angelo Porrello[2], Simone Calderara[2], Rita Cucchiara[2]

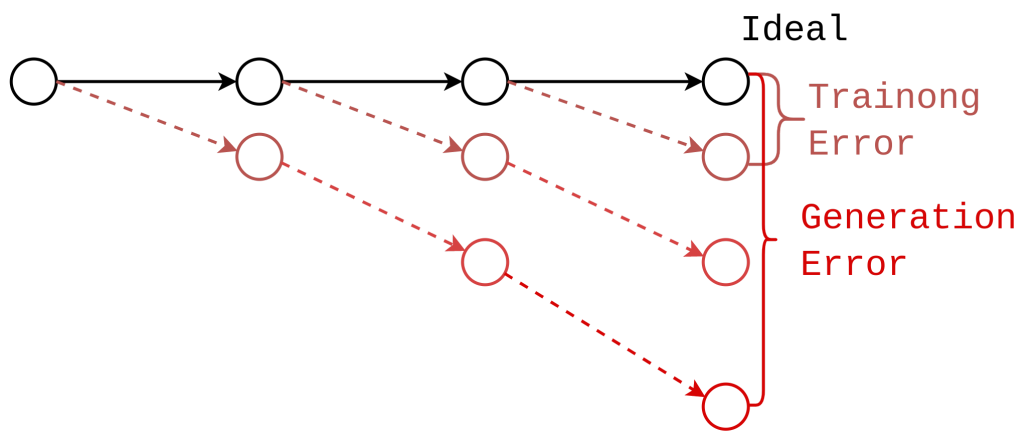
[1]Department of Information and Computing Science, Utrecht University, the Netherlands.
[2]Department of Engineering (DIEF), University of Modena and Reggio Emilia, Italy.

ICML 2023

Exposure Bias

Exposure bias, also known as the **Teacher Forcing Problem**, is a common issue in training autoregressive models. This issue occurs when there is a **Mismatch between how the model is Trained and how it is used during Generation**.

The problem arises because the model may not perform as well during inference as it did during training. It can make mistakes that accumulate over time, especially when generating longer sequences. These errors can result from slight inaccuracies in earlier predictions, leading to a divergence from the true sequence.



Exposure Bias in DPM

Diffusion Probabilistic Models (DPM) have achieved excellent performance in various generation tasks, but their **High Training and Generation Time** have been criticized.

Similar to autoregressive models, the **Step-by-Step** generation approach leads to **Exposure Bias** in DPM. To achieve high-quality generation results, a significant amount of training time is required to reduce errors at each step.

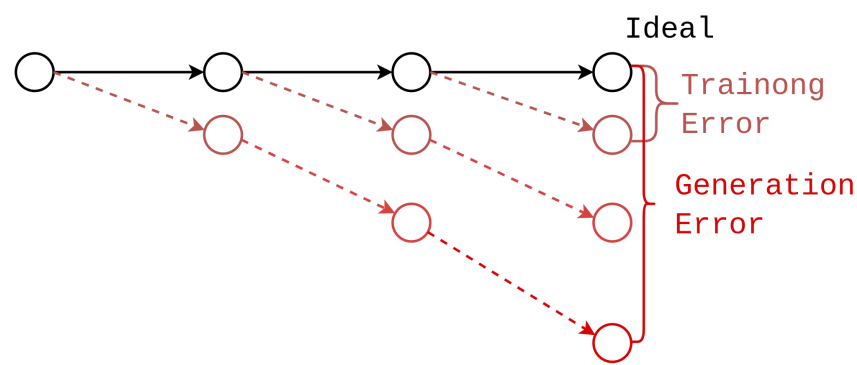
To alleviate this problem, the authors propose a surprisingly simple yet very effective method, which consists in explicitly modelling the prediction error during training.

Exposure Bias in DPM

$$x_t \sim \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, 1-\bar{\alpha}_t)$$

$$\hat{x}_t \sim \mathcal{N}(\hat{x}_t; \mu_\theta(\hat{x}_{t+1}), \sigma_t)$$

$$\hat{x}_t = x_t + \underbrace{\xi_t}_{\text{Exposure Bias}}$$



Exposure Bias in DPM

Supporting Evidence in Improved-DDPM

Ideally, the more steps the better.

However, when the number of steps increases to more than 100~300,

the quality of the generated samples begins to decline.

This is due to excessive error accumulation due to exposure bias.

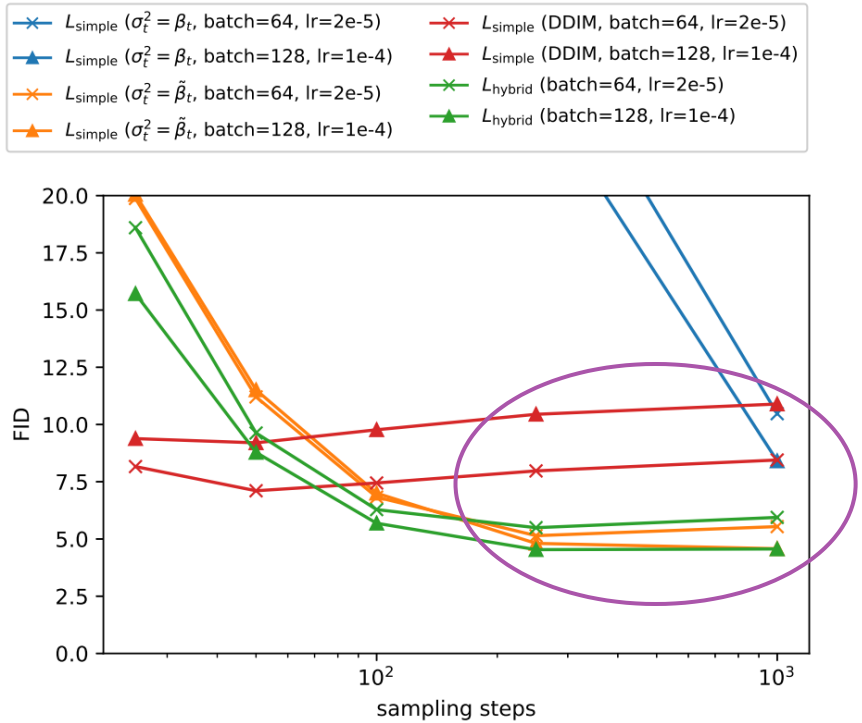


Figure 11. FID vs. number of sampling steps from an LSUN 256×256 bedroom model.

cite: Improved Denoising Diffusion Probabilistic Models.

Solutions

Smoother prediction function $\mu(\cdot)$

Regularization based on Lipschitz Continuous Functions

- Gradient penalty. $\mathcal{L}_{\text{GP}} = \mathcal{L}_{\text{DPM}} + \lambda_{\text{GP}} \left\| \frac{\partial \epsilon_\theta(x_t, t)}{\partial x} \right\|^2$
- Weight decay. $\mathcal{L}_{\text{WD}} = \mathcal{L}_{\text{DPM}} + \lambda_{\text{WD}} \|\theta\|^2$

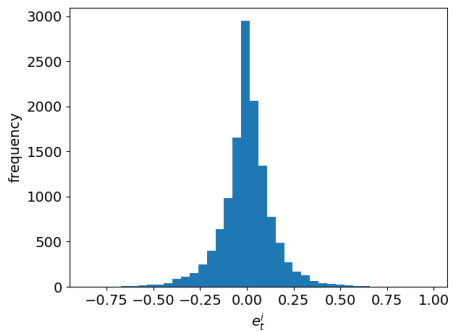
Regularization with Input Perturbation

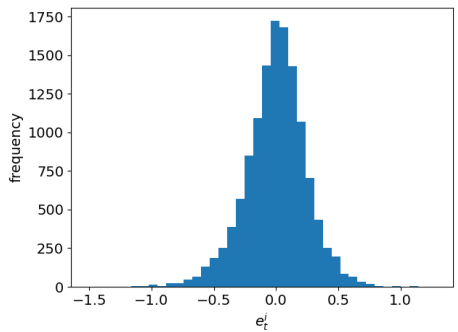
Assume that the **Exposure Bias** follows a **Normal Distribution**
$\Rightarrow$ 1. Two spatially close points a and b should lead to similar predictions $\mu(a, t)$ and $\mu(b, t)$.

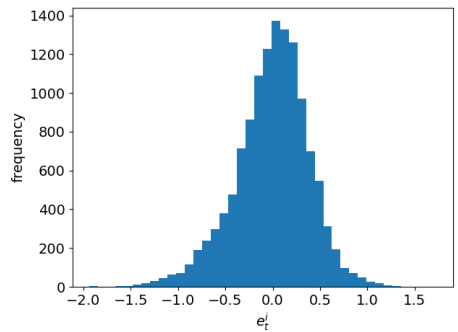
Regularization with Input Perturbation

Assume that the **Exposure Bias** follows a **Normal Distribution**
$\Rightarrow$ 1. Two spatially close points a and b should lead to similar predictions $\mu(a, t)$ and $\mu(b, t)$.

$\Rightarrow$ 2. Gaussian Prediction Error



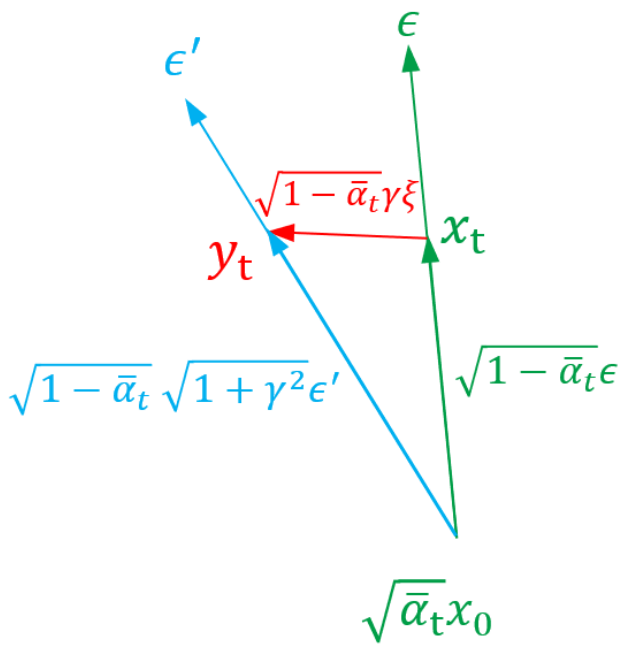




The empirical distribution of $e_t^i = x_0 - \hat{x}_0,\ \hat{x}_0 \sim p_\theta(x_{0:t-1}|x_t)$, and i is the input dimension.

- The Shapiro–Wilk test shows that they follow a standard normal distribution.

DDPM-IP



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

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

DDPM-IP

DDPM Standard Training

repeat

$\mathbf{x}_0 \sim q(\mathbf{x}_0), t \sim \mathbb{U}(\{1, \dots, T\}), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$

update θ by $\nabla_\theta \|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2$

until converged

DDPM-IP: Training with input perturbation

repeat

$\mathbf{x}_0 \sim q(\mathbf{x}_0), t \sim \mathbb{U}(\{1, \dots, T\}), \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

$\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

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

update θ by $\nabla_\theta \|\epsilon - \epsilon_\theta(\mathbf{y}_t, t)\|^2$

until converged

Comparison of Different Regularization Method

- GP costs more computation (second differential), but is no better than Input Perturbation.
- Attempting to remove the assumed bias (DDPM-y) actually makes it worse.

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**

Table 7. CIFAR10: comparing DDPM, DDPM-y and DDPM-IP using different numbers of revers diffusion steps.

Model	Input	Target	80 steps		100 steps		300 steps		1000 steps	
			FID	sFID	FID	sFID	FID	sFID	FID	sFID
DDPM	$\boldsymbol{x}_t$	$\boldsymbol{\epsilon}$	3.63	5.97	3.37	5.66	2.95	4.95	2.99	4.76
DDPM-y	$\boldsymbol{y}_t$	$\boldsymbol{\epsilon}'$	4.24	6.51	3.90	6.23	3.21	5.39	3.25	5.04
DDPM-IP	$\boldsymbol{y}_t$	$\boldsymbol{\epsilon}$	**2.93**	**4.69**	**2.70**	**4.51**	**2.67**	**4.14**	**2.76**	**4.05**

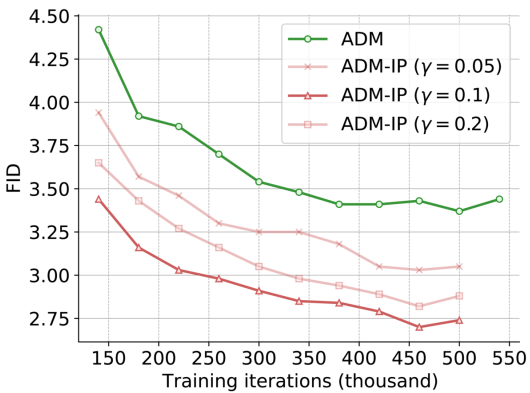
The Power of Input Perturbation

Table 3. Comparison between ADM and ADM-IP using models trained with $T = 1,000$ sampling steps and tested with $T' \leq T$ steps.

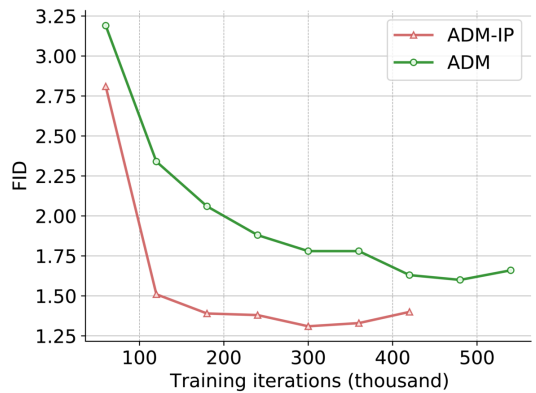
Sampling steps (T')	Model	CIFAR10		ImageNet 32		LSUN tower 64		CelebA 64		FFHQ 128	
		FID	sFID	FID	sFID	FID	sFID	FID	sFID	FID	sFID
1,000	ADM (baseline)	2.99	4.76	3.60	3.30	3.39	7.96	1.60	3.80	9.65	12.53
	ADM-IP (ours)	**2.76**	**4.05**	**2.87**	**2.39**	**2.68**	**6.04**	**1.31**	**3.38**	**2.98**	**5.59**
300	ADM	2.95	4.95	3.58	3.48	3.31	8.39	1.82	4.25	9.55	12.6
	ADM-IP	**2.67**	**4.14**	**2.74**	**2.58**	**2.60**	**5.98**	**1.43**	**3.36**	**3.74**	**5.97**
100	ADM	3.37	5.66	4.26	4.48	3.50	11.10	3.02	5.76	14.52	16.02
	ADM-IP	**2.70**	**4.51**	**3.24**	**3.13**	**2.79**	**6.56**	**2.21**	**4.33**	**5.94**	**7.90**
80	ADM	3.63	5.97	4.61	4.76	4.17	12.60	3.75	6.80	17.00	18.02
	ADM-IP	**2.93**	**4.69**	**3.57**	**3.33**	**2.95**	**6.93**	**2.67**	**4.69**	**6.89**	**8.79**

Reduce Training Time

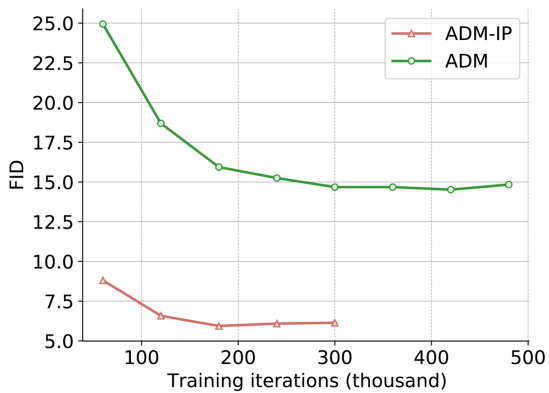
- On CelebA, ADM-IP gets FID 1.51 at **120K training iterations**, whereas ADM gets FID 1.6 at convergence (**480K iterations**). $\Rightarrow$ **4$\times$ Training Speed-up**.
- On FFHQ, ADM receives FID 14.52 at convergence (**420K iterations**), while ADM-IP achieves a FID score of 8.81 with **only 60K iterations**. $\Rightarrow$ **7$\times$ Training Speed-up**



CIFAR-10 32×32



CelebA 64×64



FFHQ 128×128

Conclusion

Contributions

This study proposes a simple method to solve Exposure Bias in DPM.

- Significantly reduce training costs.
- Achieve the same generation quality as previous methods while reducing steps.

Doubts

Assuming Bias as Gaussian Distribution conflicts with the behavior of DDPM-IP.

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

Discussions

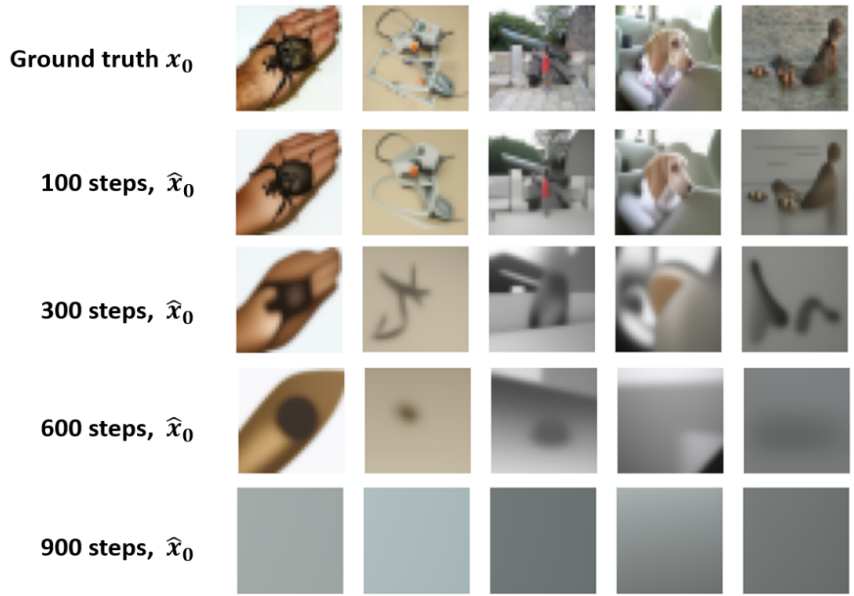


Figure 5. Visualization of the exposure bias problem with different diffusion chain lengths.
$\hat{x}_0 \sim p(x_{0:t-1}|x_t)$

Why DDPM-IP Works?

DDPM-IP actually does not match what is stated in the paper:

- "Assume bias is a normal distribution" and
- "Remove bias"

Fig. 5 shows that DPM may excessively cut high-frequency signals when t is close to T. Therefore, the reason why DDPM-IP is effective may be due to "retaining more high frequencies."