Alleviating Exposure Bias in Diffusion Models through Sampling with Shifted Time Steps

Mingxiao Li*[1], Tingyu Qu*[1], Ruicong Yao[2], Wei Sun[1] and Marie-Francine Moens[1]

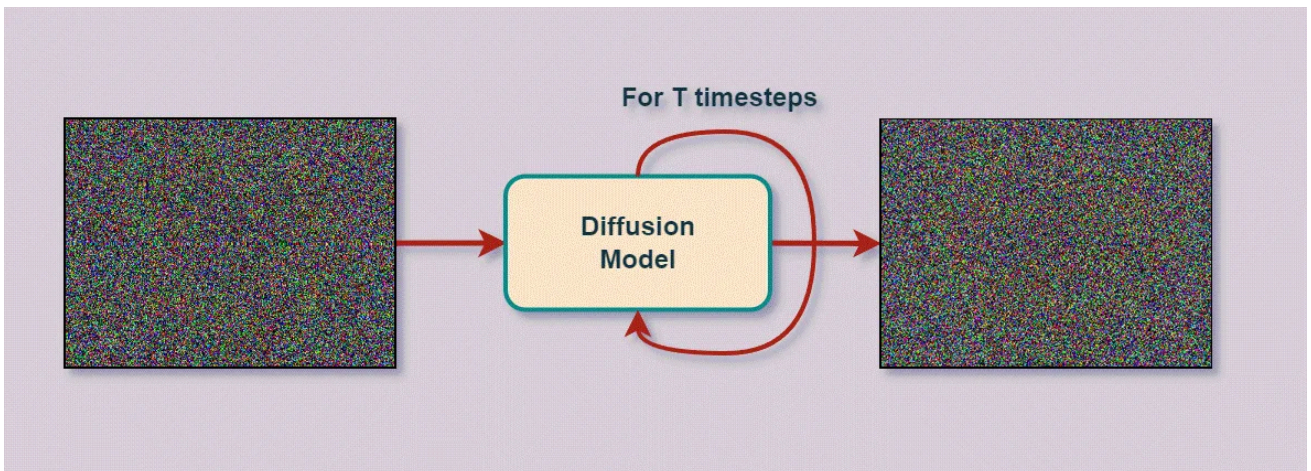
[1] Department of Computer Science, KU Leuven
[2] Department of Mathematics, KU Leuven

ICLR 2024

About Diffusion Models?

Diffusion Probabilistic Models (DPMs) have gained significant attention in recent years for various generative tasks due to their **Training Stability, Superior Quality and Diversity** of sampled results.

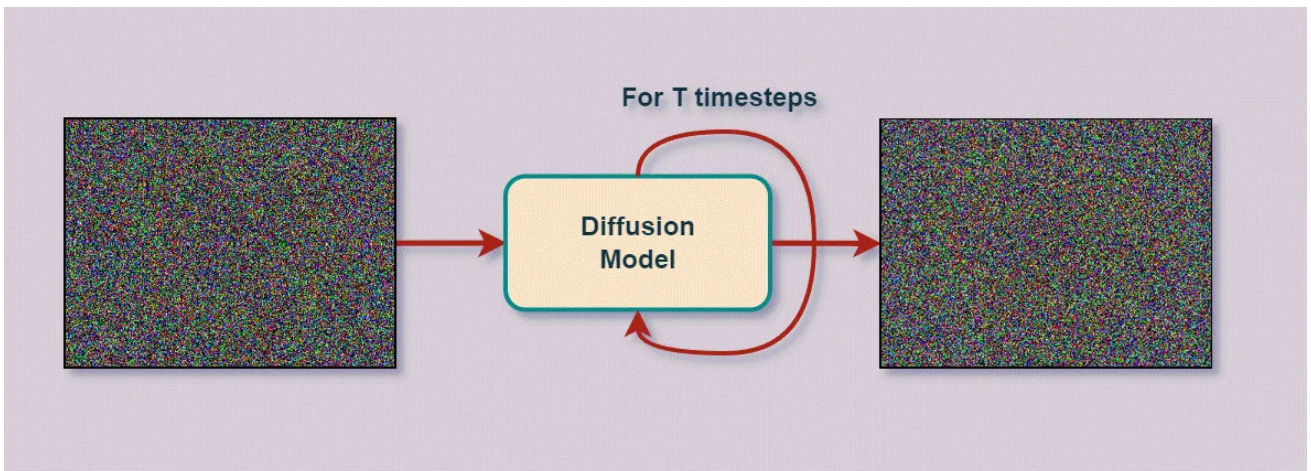
Unlike single-step generative methods such as Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs), DPMs leverage theories like Markov Chain or Score Matching to **Connect the Data Distribution with a known Distribution (e.g., Gaussian)**.



About Diffusion Models?

Unlike single-step generative methods such as Generative Adversarial Networks (GANs) and Variational AutoEncoders (VAEs), DPMs leverage theories like Markov Chain or Score Matching to **Connect the Data Distribution with a known Distribution (e.g., Gaussian)**.

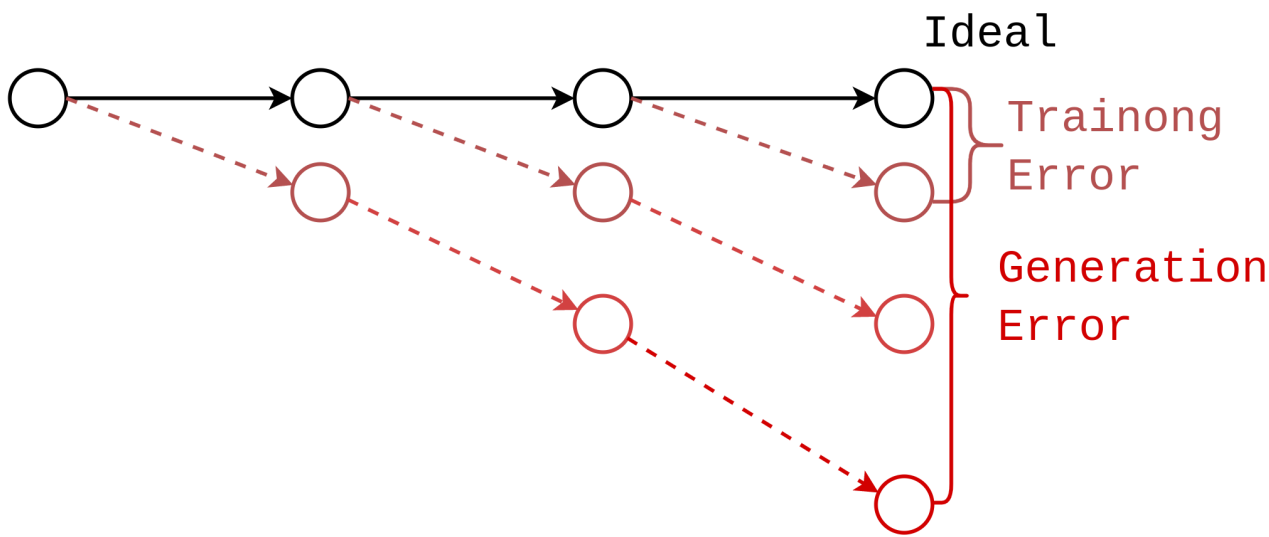
By utilizing a trained neural network, the model iteratively performs hundreds to thousands of sampling steps to transform samples drawn from the known distribution into the data distribution.



About Diffusion Models?

Disadvantages and Problems

However, as mentioned earlier, DPMs **Require Multiple Iterations to Generate Results**, making the process time-consuming. Additionally, **Estimation Errors from the Neural Network Accumulate** during these iterations, leading to a decline in the final generation quality. This is similar to the **Exposure Bias Problem** observed in Autoregressive Models.

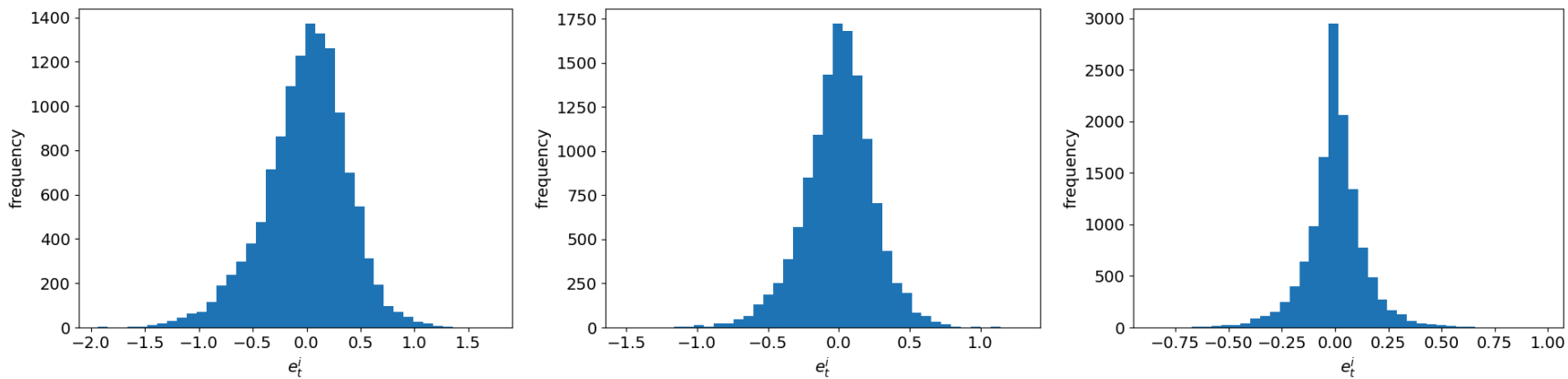


About Diffusion Models?

Exposure Bias in DPM

Previous research [DDPM-IP] has found that **Exposure Bias follows a Gaussian Distribution**.

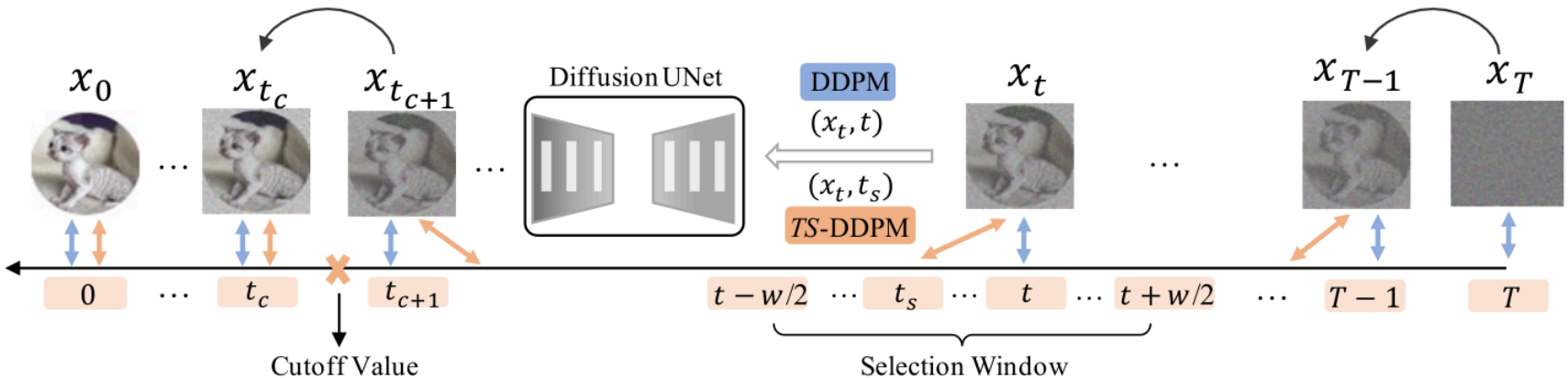
To enhance the model's robustness against bias interference during the sampling process, additional Gaussian noise is introduced during training to simulate exposure bias.



About Diffusion Models?

Exposure Bias in DPM

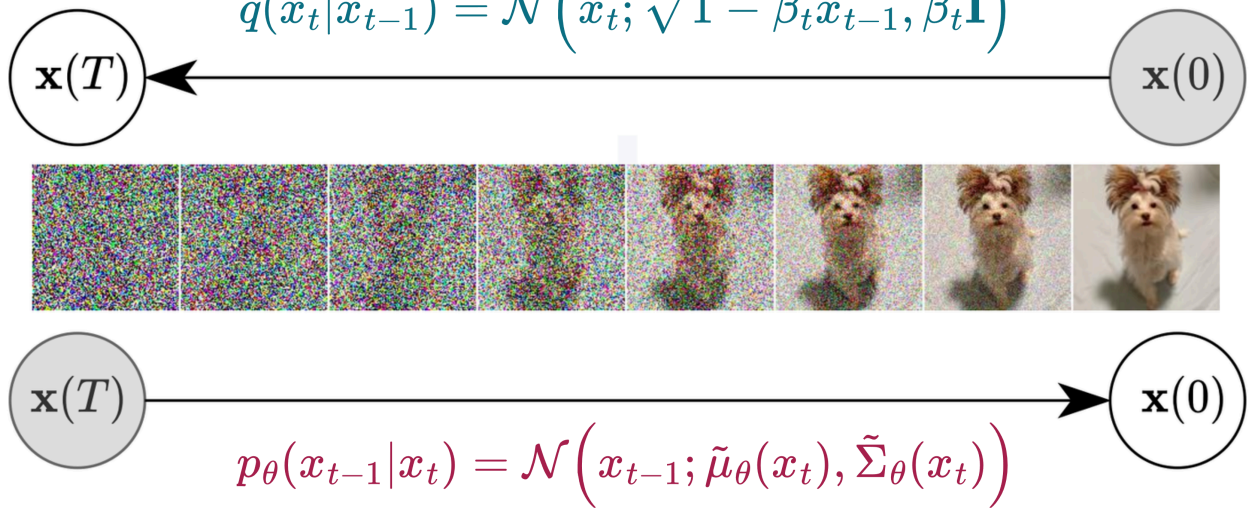
However, retraining the model is costly. To mitigate exposure bias without the need for retraining, this study proposes a technique called **Time Step Shifting**. This method dynamically adjusts the time step (**Equivalent to the Standard Deviation** of Gaussian noise) during the sampling process, improving the sampling quality **Without Requiring Model Retraining**.



Investigating Exposure Bias in DPMs

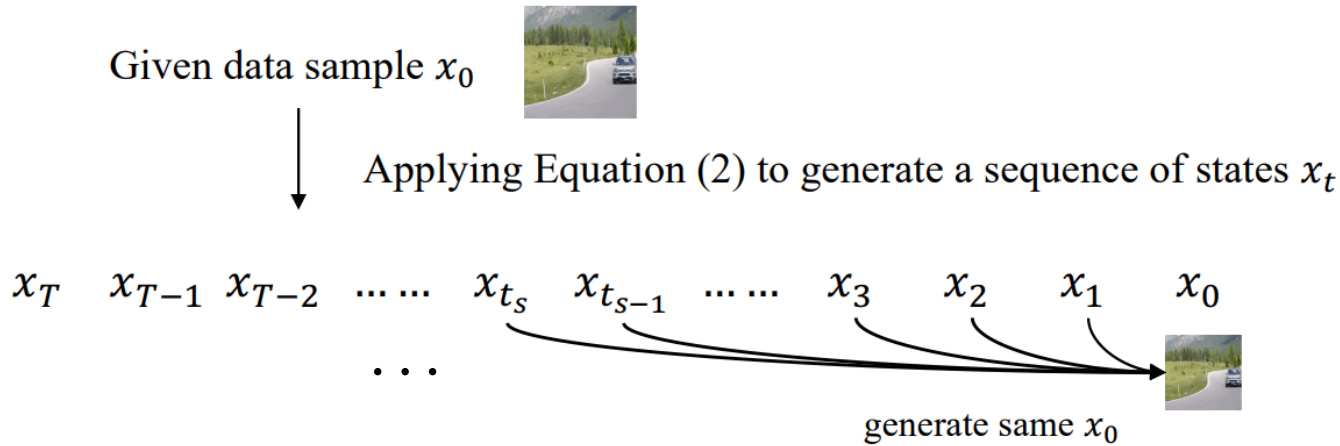
Background: Diffusion Probabilistic Models

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathbf{I}\right)$$



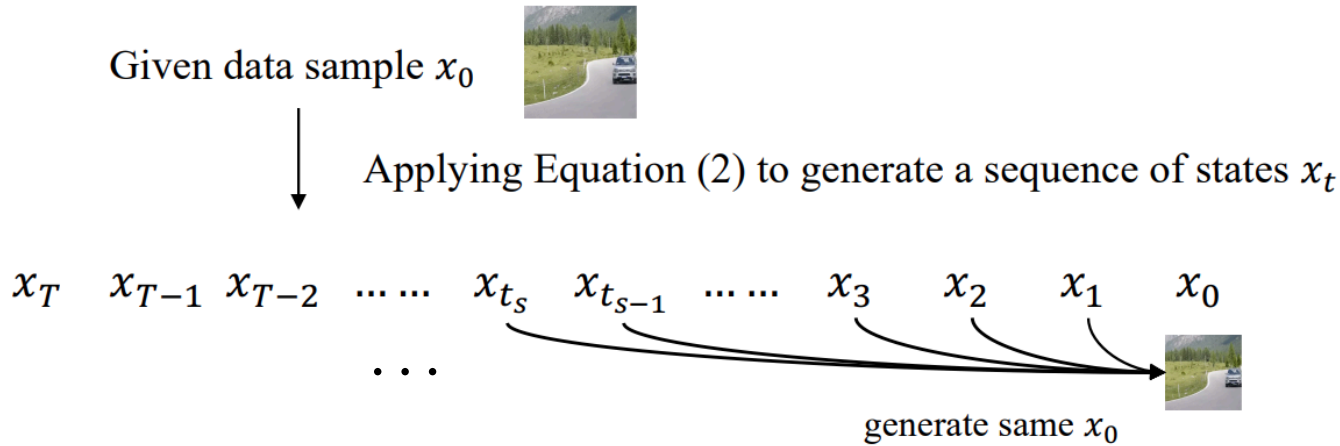
$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \tilde{\mu}_\theta(x_t), \tilde{\Sigma}_\theta(x_t)\right)$$

Investigating Exposure Bias in DPMs



To investigate the severity of exposure bias at different steps in DPM, the authors analyzed the Mean Squared Error (MSE) between generated samples and their corresponding real samples.

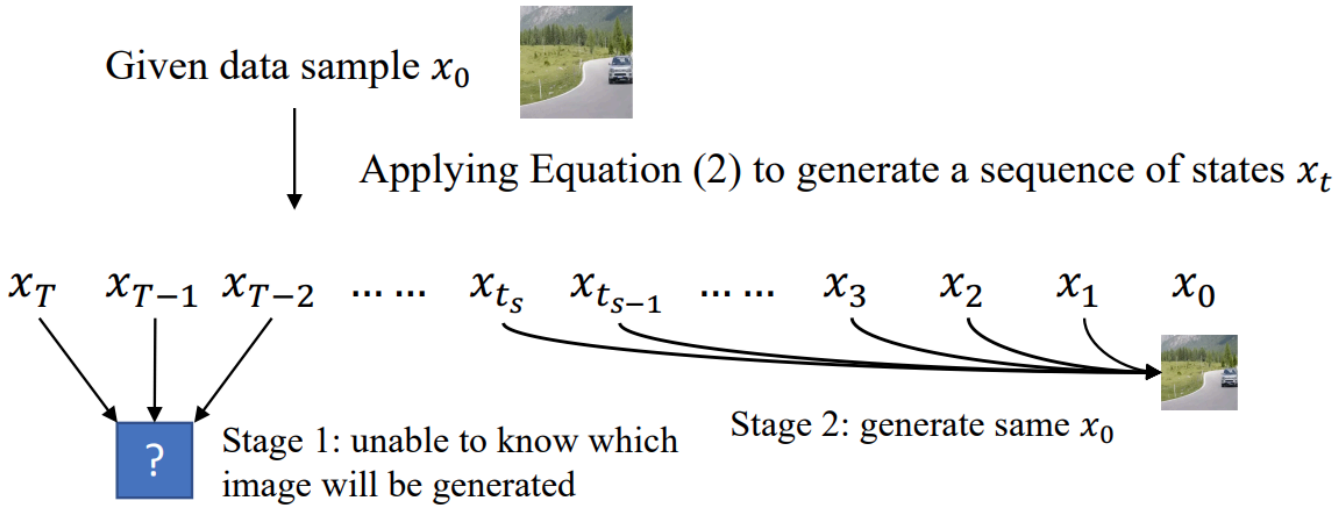
Investigating Exposure Bias in DPMs



To investigate the severity of exposure bias at different steps in DPM, the authors analyzed the Mean Squared Error (MSE) between generated samples and their corresponding real samples.

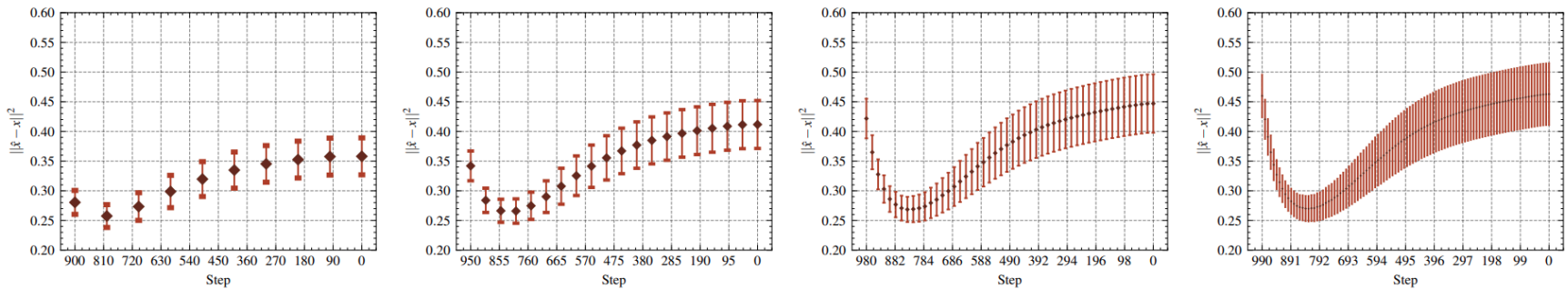
Additionally, they found that **If the Initial Sampling Step t_s is too Close to T**, the generated samples become unrelated to the real samples.

Investigating Exposure Bias in DPMs



- **Stage 1** ($t > t_s$)
  - $\text{MSE}_t = \dfrac{\|\hat{x}_t - x_t\|_2^2}{n}$, $\hat{x}_{T-1:t_s+1} \sim p_\theta(x_{T-1:t_s+1}|x_T)$
- **Stage 2** ($t \leq t_s$)
  - $\text{MSE}_t = \dfrac{\|\hat{x}_0^t - x_0\|_2^2}{n}$, $\hat{x}_0^t = x_\theta(\hat{x}_{t+1})$, $\hat{x}_{t_s-1:1} \sim p_\theta(x_{t_s-1:1}|x_{t_s})$

Investigating Exposure Bias in DPMs

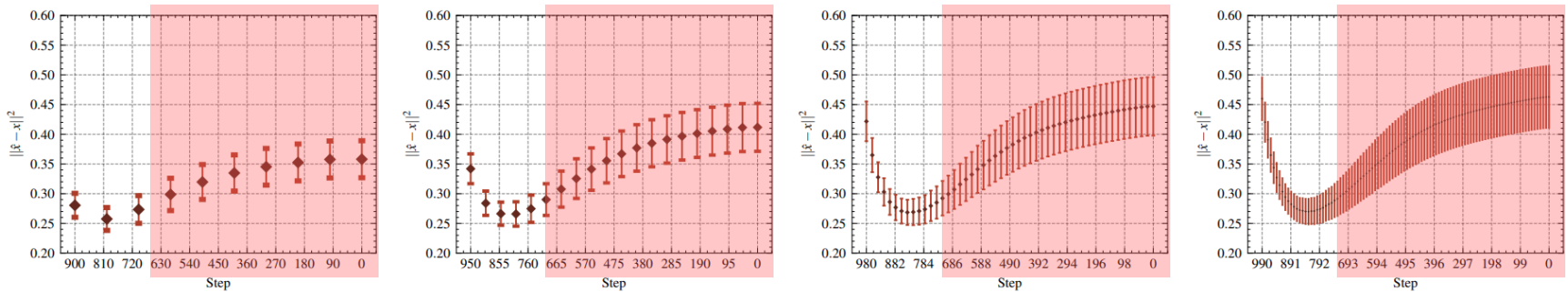


(a) 10 steps (b) 20 steps (c) 50 steps (d) 100 steps

Figure 3: CIFAR-10 prediction errors of training samples for different numbers of sampling steps.

- The generation error in the second phase did not decrease as the time steps progressed;

Investigating Exposure Bias in DPMs



(a) 10 steps (b) 20 steps (c) 50 steps (d) 100 steps

Figure 3: CIFAR-10 prediction errors of training samples for different numbers of sampling steps.

- The generation error in the second phase did not decrease as the time steps progressed;
- Instead, **It Gradually Increased.**

Alleviating Exposure Bias via Time Step Shifting

- The Exposure Bias e_t in DPMs follows Normal Distribution. [DDPM-IP]

$$\begin{aligned}\tilde{x}_t =& x_t + e_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon + e_t \\ =& \sqrt{\bar{\alpha}_t} x_0 + \lambda_t \tilde{\epsilon}_t\end{aligned}$$

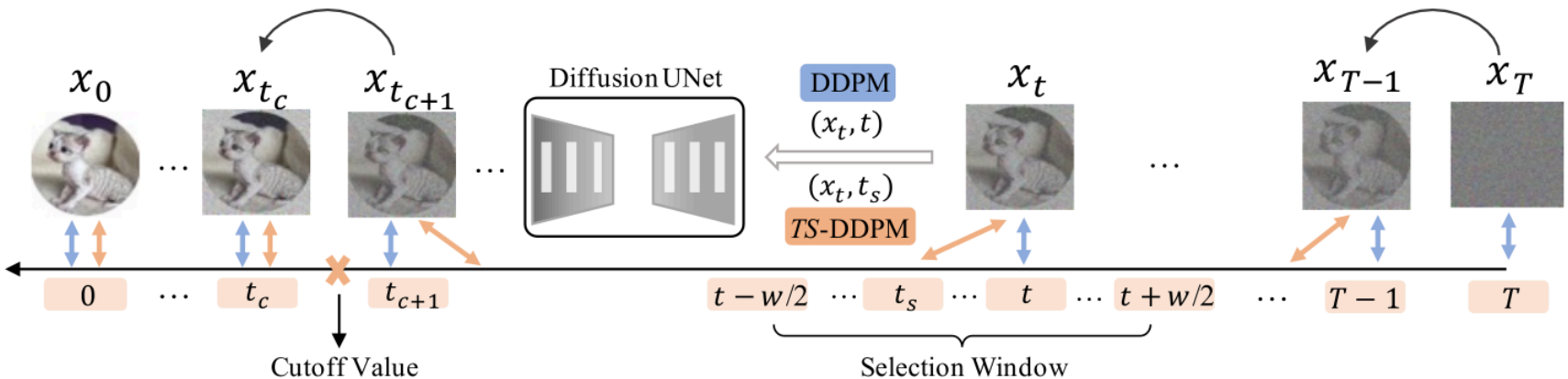
-

Alleviating Exposure Bias via Time Step Shifting

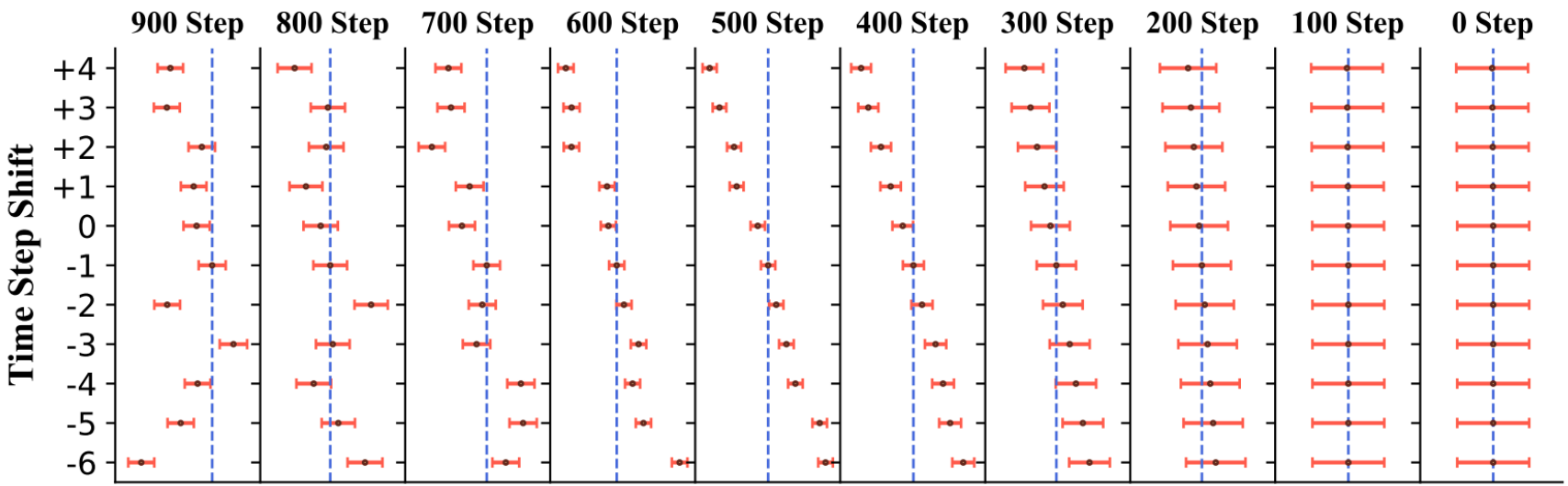
- The Exposure Bias e_t in DPMs follows Normal Distribution. [DDPM-IP]

$$\begin{aligned}\tilde{x}_t =& x_t + e_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon + e_t \\ =& \sqrt{\bar{\alpha}_t}x_0 + \lambda_t\tilde{\epsilon}_t\end{aligned}$$

- Assume that $\tilde{x}_t$ still follows a certain distribution of $x_{T:0}$



Alleviating Exposure Bias via Time Step Shifting



$$C(\tilde{x}_{t-1}, t-s) = e^{-\mathrm{dis}(\tilde{x}_{t-1}, x_{t-s})}$$

- For certain backward steps, **There are Alternate Time Steps** $t-s$ that display a stronger correlation with the predicted next state x_{t-1} compared to time step $t-1$.
- The zero time step, **All Nearby Time Steps Converge to the Same Distribution**.

Alleviating Exposure Bias via Time Step Shifting

Algorithm 3 Time-Shift Sampler

Input : Trained diffusion model ϵ_θ; Window size w; Reverse Time series $\{T, T-1, \cdots, 0\}$; Cutoff threshold t_c
Initialize: $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; $t_s = -1$
for $t = T, T-1, .., 0$ **do**
    If $t_s \neq -1$ then $t_{next} = t_s$ else $t_{next} = t$
    $\epsilon_t = \epsilon_\theta(x_t, t_{next}))$
    take a sampling step with t_{next} to get x_{t-1}
    if $t > t_c$ **then**
        Get variance for time steps within the window: $\Sigma = \{1 - \overline{\alpha}_{t-w/2}, \cdots, 1 - \overline{\alpha}_{t+w/2}\}$
        $t_s = \arg\min_\tau ||var(x_{t-1}) - \sigma_\tau||$, for $\sigma_\tau \in \Sigma$ and $\tau \in [t - w/2, t + w/2]$
    else
        $t_s = -1$
    end if
end for
return x_0

Alleviating Exposure Bias via Time Step Shifting

Algorithm 3 Time-Shift Sampler

Input : Trained diffusion model ϵ_θ; Window size w; Reverse Time series $\{T, T-1, \cdots, 0\}$; Cutoff threshold t_c
Initialize: $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$; $t_s = -1$
for $t = T, T-1, .., 0$ **do**
    If $t_s \neq -1$ then $t_{next} = t_s$ else $t_{next} = t$
    $\epsilon_t = \epsilon_\theta(x_t, t_{next}))$
    take a sampling step with t_{next} to get x_{t-1}
    if $t > t_c$ **then**
        Get variance for time steps within the window: $\Sigma = \{1 - \overline{\alpha}_{t-w/2}, \cdots, 1 - \overline{\alpha}_{t+w/2}\}$
        $t_s = \arg\min_\tau ||var(x_{t-1}) - \sigma_\tau||$, for $\sigma_\tau \in \Sigma$ and $\tau \in [t - w/2, t + w/2]$
    else
        $t_s = -1$
    end if
end for
return x_0

If t is large, $\sqrt{\overline{\alpha}_t} x_0$ will be very close to 0.

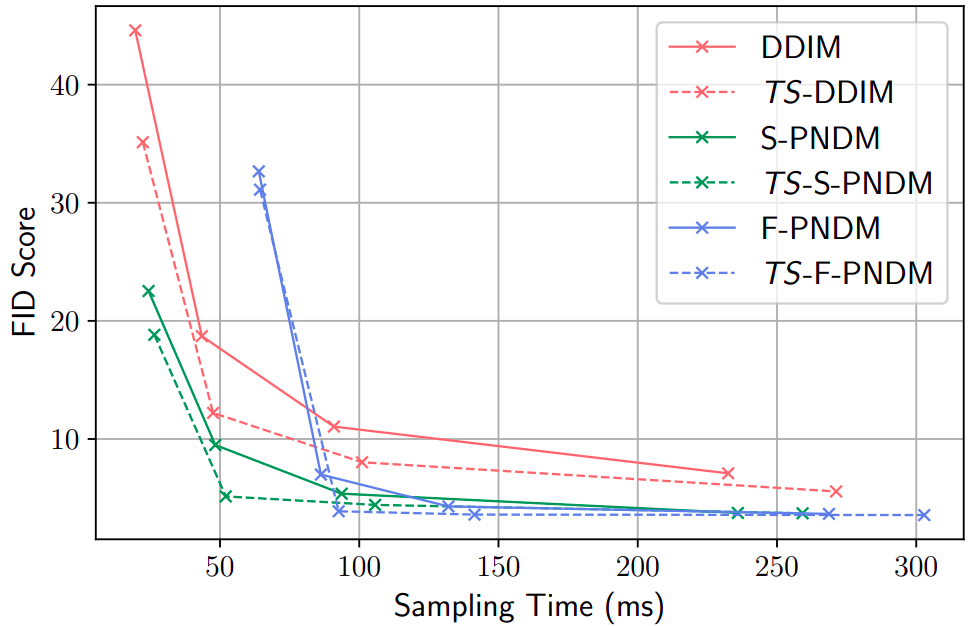
Experiment

Main Results

Dataset	Sampling Method	5 steps	10 steps	20 steps	50 steps	100 steps
CIFAR-10	DDIM (*quadratic*)	41.57	13.70	6.91	4.71	4.23
	TS-DDIM(*quadratic*)	**38.09 (+8.37%)**	**11.93 (+12.92%)**	**6.12 (+11.43%)**	**4.16 (+11.68%)**	**3.81 (+9.93%)**
	DDIM(*uniform*)	44.60	18.71	11.05	7.09	5.66
	TS-DDIM(*uniform*)	**35.13 (+21.23%)**	**12.21 (+34.74%)**	**8.03 (+27.33%)**	**5.56 (+21.58%)**	**4.56 (+19.43%)**
	DDPM (*uniform*)	83.90	42.04	24.60	14.76	10.66
	TS-DDPM (*uniform*)	**67.06 (+20.07%)**	**33.36 (+20.65%)**	**22.21 (+9.72%)**	**13.64 (+7.59%)**	**9.69 (+9.10%)**
	S-PNDM (*uniform*)	22.53	9.49	5.37	3.74	3.71
	TS-S-PNDM (*uniform*)	**18.81(+16.40%)**	**5.14 (+45.84%)**	**4.42 (+17.69%)**	**3.71 (+0.80%)**	**3.60 (+2.96%)**
	F-PNDM (*uniform*)	31.30	6.99	4.34	3.71	4.03
	TS-F-PNDM (*uniform*)	**31.11 (+4.07%)**	**3.88 (+44.49%)**	**3.60 (+17.05%)**	**3.56 (+4.04%)**	**3.86 (+4.22%)**
CelebA	DDIM (*quadratic*)	27.28	10.93	6.54	5.20	4.96
	TS-DDIM (*quadratic*)	**24.24 (+11.14%)**	**9.36 (+14.36%)**	**5.08 (+22.32%)**	**4.20 (+19.23%)**	**4.18 (+15.73%)**
	DDIM (*uniform*)	24.69	17.18	13.56	9.12	6.60
	TS-DDIM (*uniform*)	**21.32 (+13.65%)**	**10.61 (+38.24%)**	**7.01 (+48.30%)**	**5.29 (+42.00%)**	**6.50 (+1.52%)**
	DDPM (*uniform*)	42.83	34.12	26.02	18.49	13.90
	TS-DDPM (*uniform*)	**33.87 (+20.92%)**	**27.17 (+20.37%)**	**20.42 (+21.52%)**	**13.54 (+26.77%)**	**12.83 (+7.70%)**
	S-PNDM (*uniform*)	38.67	11.36	7.51	5.24	4.74
	TS-S-PNDM (*uniform*)	**29.77 (+23.02%)**	**10.50 (+7.57%)**	**7.34 (+2.26%)**	**5.03 (+4.01%)**	**4.40 (+7.17%)**
	F-PNDM (*uniform*)	94.94	9.23	5.91	4.61	4.62
	TS-F-PNDM (*uniform*)	**94.26 (+0.72%)**	**6.96 (+24.59%)**	**5.84 (+1.18%)**	**4.50 (+2.39%)**	**4.42 (+4.33%)**

Experiment

Discussion on Efficiency and Performance



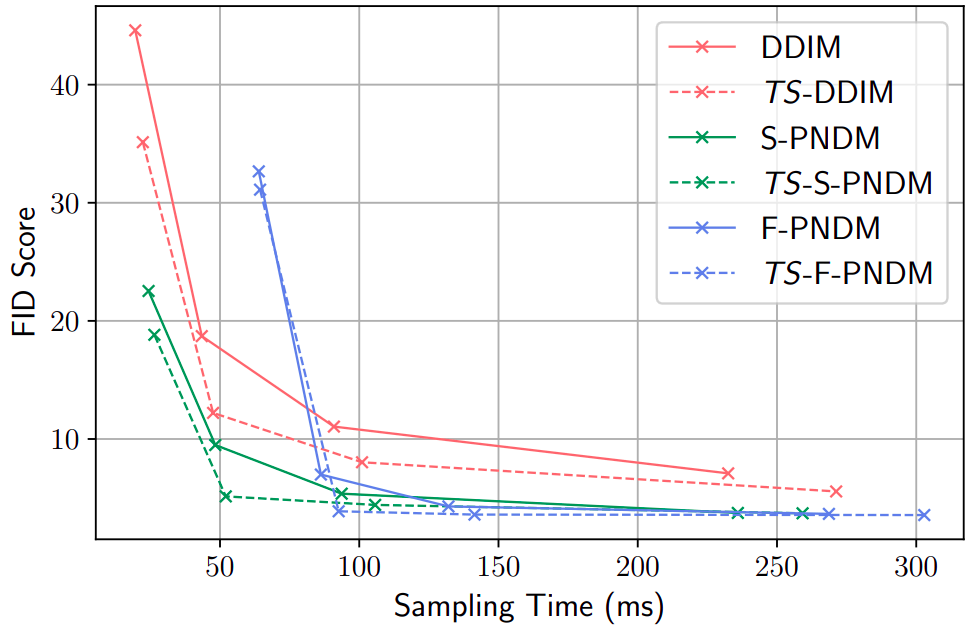
Model	Sampling Method	5 steps	10 steps	20 steps	50 steps
ADM	DDIM	28.98	12.11	7.14	4.45
ADM-IP	DDIM	50.58 (-74.53%)	20.95 (-73.00%)	7.01 (+1.82%)	**2.86 (+35.73%)**
ADM	*TS*-DDIM	**26.94 (+7.04%)**	**10.73 (+11.40%)**	**5.35 (+25.07%)**	3.52 (+20.90%)

Table 2: Performance comparison on CIFAR-10 with ADM and ADM-IP as backbone models.

Experiment

Discussion on Efficiency and Performance

The additional time required for Time Shift Sampling is **Independent of the Model Size**.



Model	Sampling Method	5 steps	10 steps	20 steps	50 steps
ADM	DDIM	28.98	12.11	7.14	4.45
ADM-IP	DDIM	50.58 (-74.53%)	20.95 (-73.00%)	7.01 (+1.82%)	**2.86 (+35.73%)**
ADM	*TS*-DDIM	**26.94 (+7.04%)**	**10.73 (+11.40%)**	**5.35 (+25.07%)**	3.52 (+20.90%)

Table 2: Performance comparison on CIFAR-10 with ADM and ADM-IP as backbone models.

Experiment

Influence of Window Sizes and Cutoff values

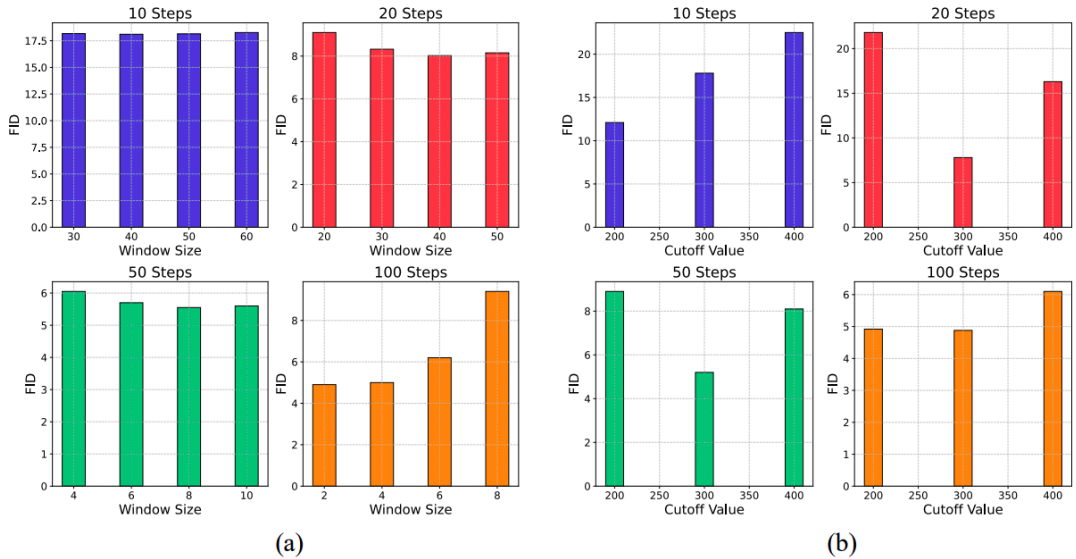
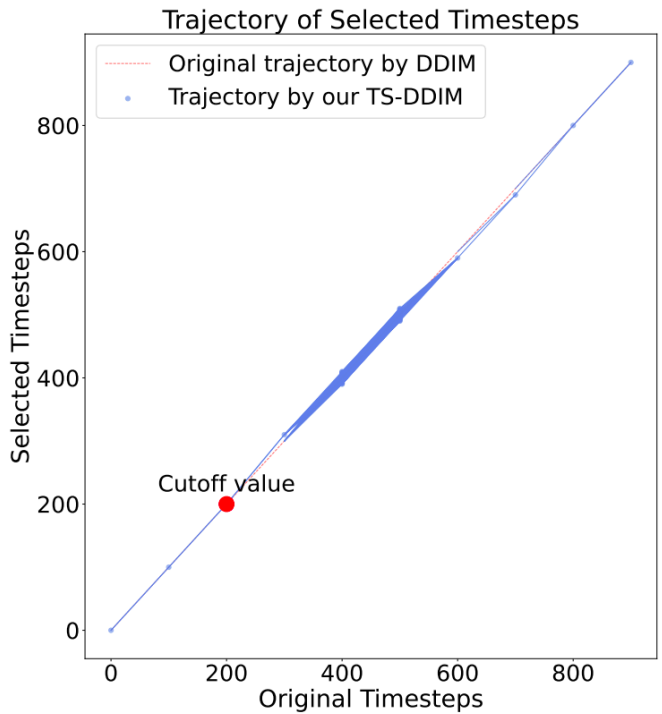
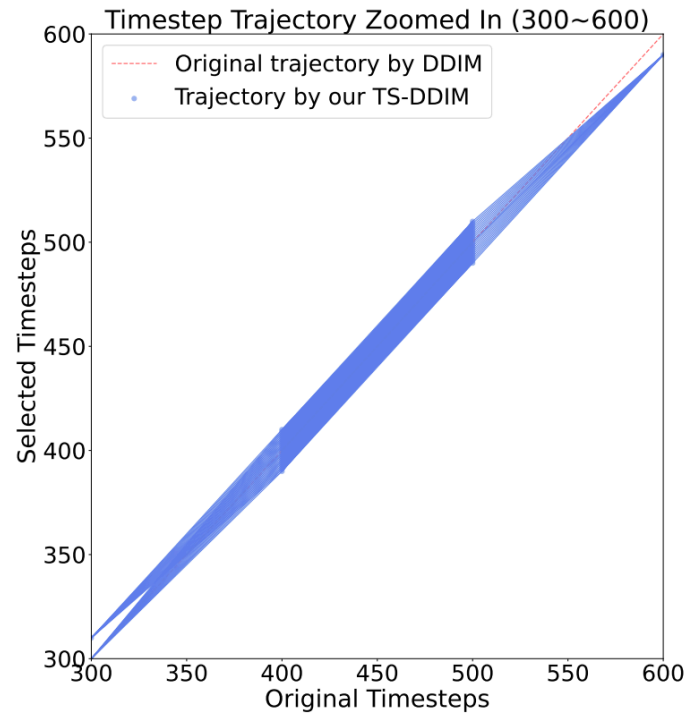


Figure 7: FID of generated CIFAR-10 images using *TS*-DDIM (*uniform*) with (a) various window sizes using cutoff value=300; (b) various cutoff values using window size= {40;30;8;2} for {10;20;50;100} steps.

Experiment

Timestep Trajectory





Conclusions

- Building on previous research, the authors analyzed and found that the $x_t + e_t$ affected by exposure bias **Still Belongs to a Distribution within the Diffusion Process**.
- Through mathematical proof, they demonstrated that as long as t **is Sufficiently Large**, the corresponding t_s **can be Identified using the Variance**.
- By employing a simple **Time Step Shifting** technique, the negative impact of exposure bias on DPM can be significantly reduced **Without the Need for Retraining the Model**.
- As the backbone size increases, **the Overhead Becomes Nearly Negligible**.

Sampling Method	5 steps	10 steps	20 steps	50 steps
ADM w/ DDIM	39.26 ms	79.30 ms	157.43 ms	393.77 ms
ADM w/ *TS*-DDIM	40.09 ms (+2.10%)	82.26 ms (+3.70%)	160.45 ms (+1.92%)	394.95 ms (+0.29%)

Table 4: Sampling time comparison for DDIM and *TS*-DDIM on CIFAR-10 with ADM as backbone.

END