Alleviating Exposure Bias in Diffusion Models through Time-Shift Sampling

Mingxiao Li - Tingyu Qu - Ruicong Yao - Wei Sun - Marie-Francine Moens

Department of Computer Science and Department of Mathematics, KU Leuven

Presented by: Amirhossein Alimohammadi

Diffusion Probabilistic Models

- Diffusion probabilistic models (DPM) are a powerful tool for high-quality image synthesis.
- They model the process of gradually transforming a noise vector into a realistic image.
- DPM has achieved state-of-the-art results in various image generation tasks.

An existing problem in DPMs

Exposure bias:

Exposure bias refers to the discrepancy between training and inference in diffusion probabilistic models, leading to suboptimal image generation.

Exposure Bias

This bias occurs because the sampling process originates from white noise, which lacks information about the target sample distribution.

As a result, the model gradually shapes the predicted distribution into the target distribution, but errors accumulate at each step, leading to larger errors towards the end of the sampling process.

Exposure Bias

Introducing $\mathcal{C}(\tilde{x}_t, t)$ -referred to as the input couple for a trained DPM–to describe this discrepancy, which can be expressed as:

$$\mathcal{C}(\tilde{x}_t, t) = e^{-dis(\tilde{x}_t, x_t)}$$

Where $\hat{x}_{t-1}$ could be represented as:

$$\begin{aligned}\hat{x}_{t-1} &= x_{t-1} + \phi_{t-1} e_{t-1} \\ &= \sqrt{\overline{\alpha}_{t-1}} x_0 + \sqrt{1 - \overline{\alpha}_{t-1}} \epsilon_{t-1} + \phi_{t-1} e_{t-1} \\ &= \sqrt{\overline{\alpha}_{t-1}} x_0 + \lambda_{t-1} \tilde{\epsilon}_{t-1}\end{aligned}$$

What is the problem?

During the training phase, the relationship $\mathcal{C}(\tilde{x}_t, t)$ = 1 holds true for all time steps, as the network always takes ground truth xt as input.

If we could find a better coupling like $\mathcal{C}(\tilde{x}_t, t_s)$, we are able to reduces the discrepancy between training and inference, thereby alleviating exposure bias.

Previous Work

Ning et al. (2023) were the first addressing this issue,

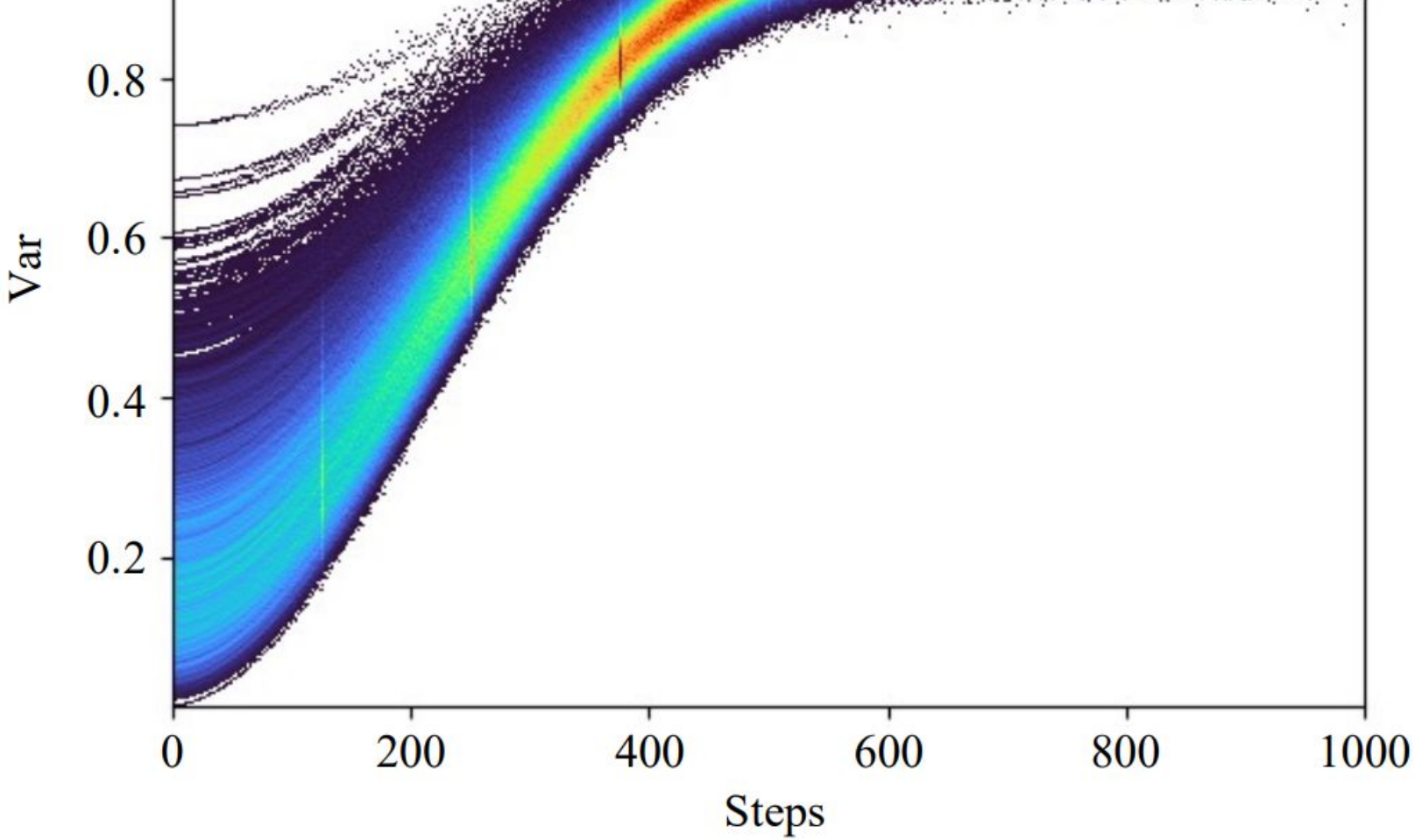
- They added perturbation to training samples to alleviate the exposure bias problem.

But, what is the cons of their idea?

- Retraining of DPM is computationally expensive.

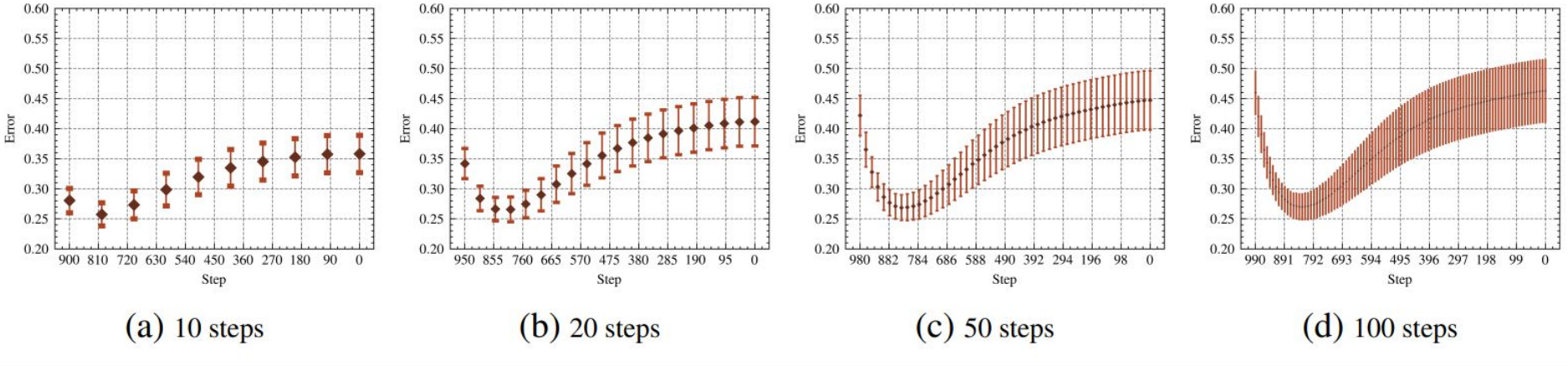
Selected approach to solve exposure bias

Given that the time step t is directly linked to the corruption level of the data samples, by adjusting the next time step t−1 during sampling according to the variance of the current generated samples, one can effectively alleviate the exposure bias.



The density distribution of the variance of 5000 samples from CIFAR10 by different time steps.

Empirical Analysis of Exposure Bias in DPM



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

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

The reasons of this phenomenon

1. In early stages, with fewer sampling steps, the error accumulation is less serious, thus the network gradually shapes the predicted distribution into the target distribution.

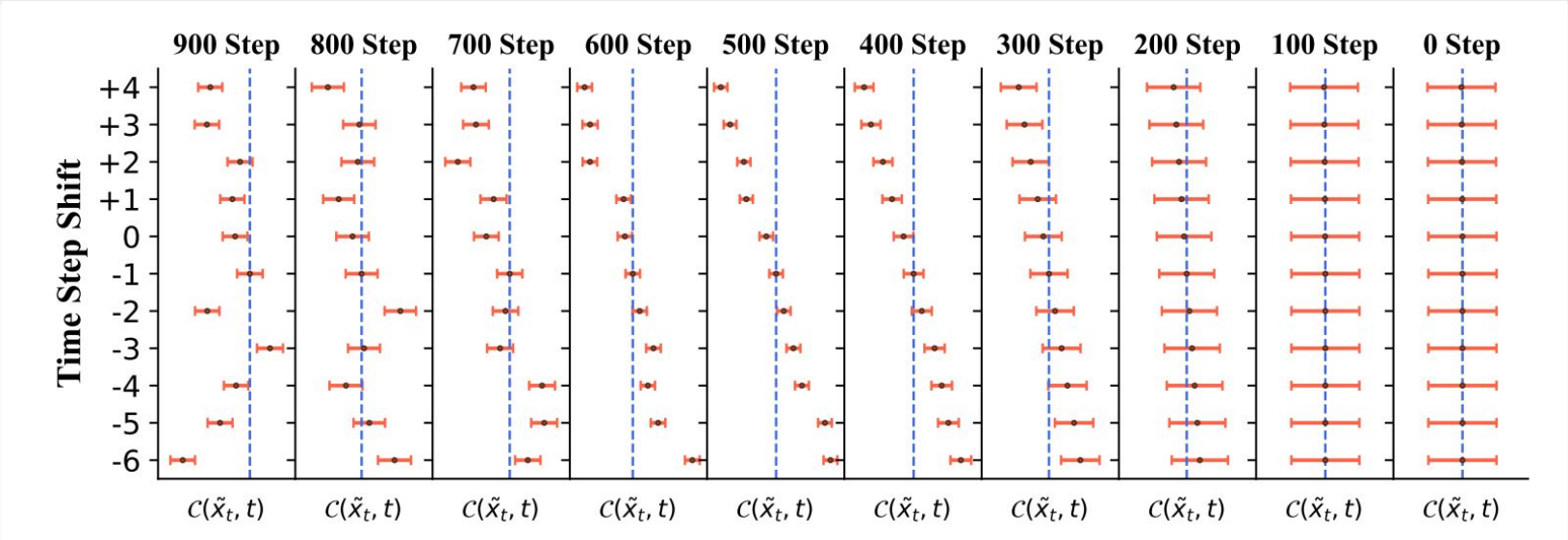
2. In the later stages, the network is more robust to the noisy inputs as discussed above. However, due to the exposure bias, the network inevitably makes errors at each step and this errors accumulate along the sequence, resulting in a slow but steady progress in error accumulation and larger errors in the end.

Assumption

During inference at time step t, the next state $\hat{x}_{t-1}$ predicted by the network, may not optimally align with time step t−1 within the context of the pretrained diffusion model. In other words, there might be an alternate time step t_s, that potentially couples better with $\hat{x}_{t-1}$:

$$\exists s \in \{1 \cdots T\}, \quad s.t. \quad \mathcal{C}(\hat{x}_{t-1}, t_s) \geq \mathcal{C}(\hat{x}_{t-1}, t-1)$$

Statistical Examination of the Assumption



The training and inference discrepancy of DDIM with 10 sampling steps on CIFAR-10. The dashed line in each column denotes the couple of prediction. Points on the right side of the dashed line mean that the corresponding time steps couple better with $\hat{x}_t$ than time step t.

Theorem

Given state $\hat{x}_t$ and assuming the predicted next state is $\hat{x}_{t-1}$, the optimal time step t_s among those time steps closely surrounding t−1, which best couples with $\hat{x}_{t-1}$, is determined to have the following variance:

$$\sigma_s \approx \sigma_{t-1} - \frac{||e||^2}{d\,(d-1)}$$

where d is the dimension of the input, e represents the network prediction error, and σ_{t-1} is the variance of the predicted $\hat{x}_{t-1}$.

Why we can choose another time step?

The Theorem could be further simplified to $\sigma_s \approx \sigma_{t-1}$, when t is large and given the assumption that the network prediction error at the current time step is minimal. This assumption has been found to hold well in practice.

Time-Shift Sampler Algorithm

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: $\Sigma = \{1 - \overline{\alpha}_{t-w/2}, 1 - \overline{\alpha}_{t-w/2+1}, \cdots, 1 - \overline{\alpha}_{t+w/2}\}$

    $t_s = \arg\min_\tau ||var(x_{t-1}) - \sigma_\tau||$, for $\sigma_\tau \in \Sigma$ and $\tau \in [t - w/2, t + w/2]$

  else

    $t_s = -1$

  end if

end for

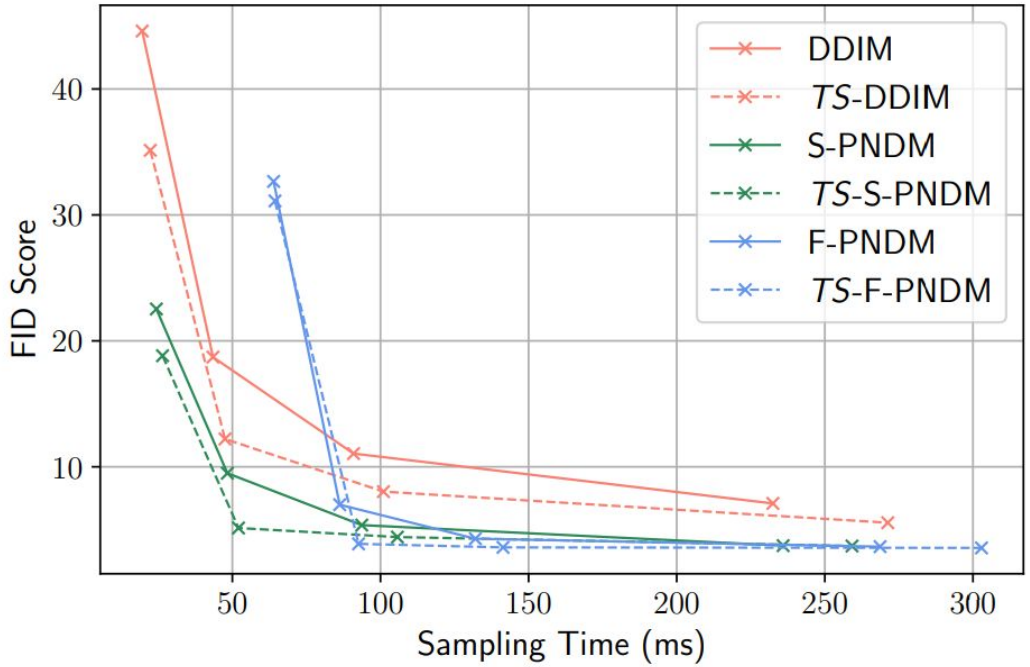
return x_0

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%)**

Quality of the image generation measured with FID ↓ on CIFAR-10 (32×32) and CelebA (64×64) with varying time steps for different sampling algorithms.

Efficiency & Performance



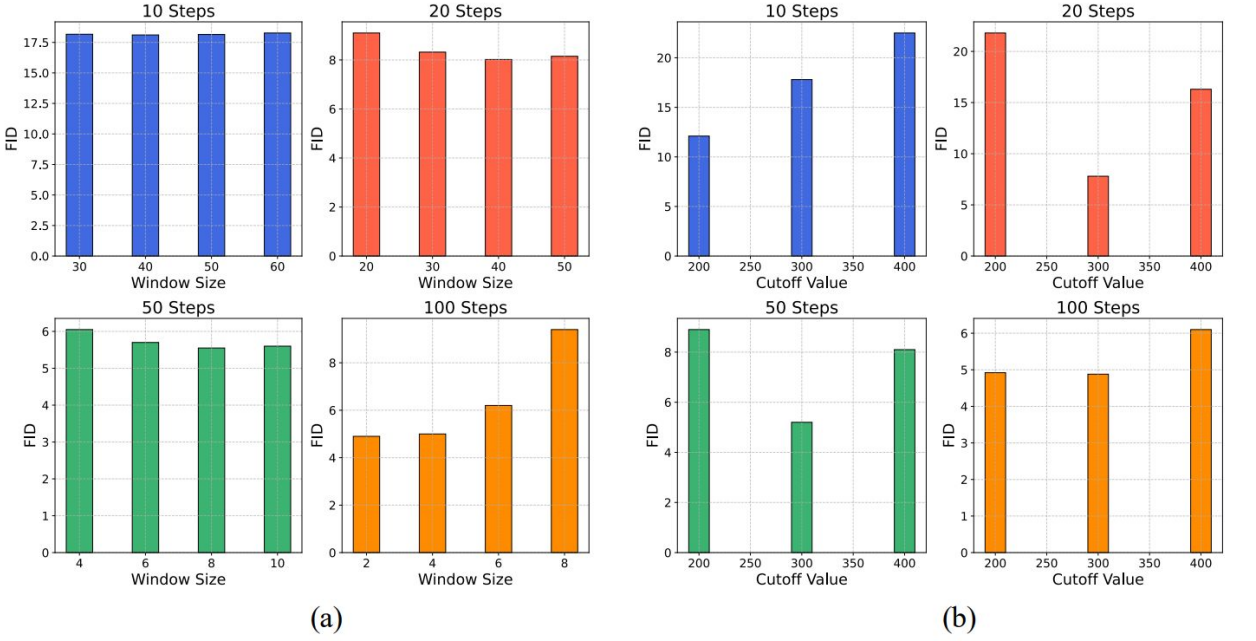
Sampling time VS FID on CIFAR-10 using DDPM as backbone with various sampling methods. We report the results of {5,10,20,50} sampling steps from left to right for each sampler, denoted with ”×” symbol.

Comparison with (Ning et al. 2023)

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%)

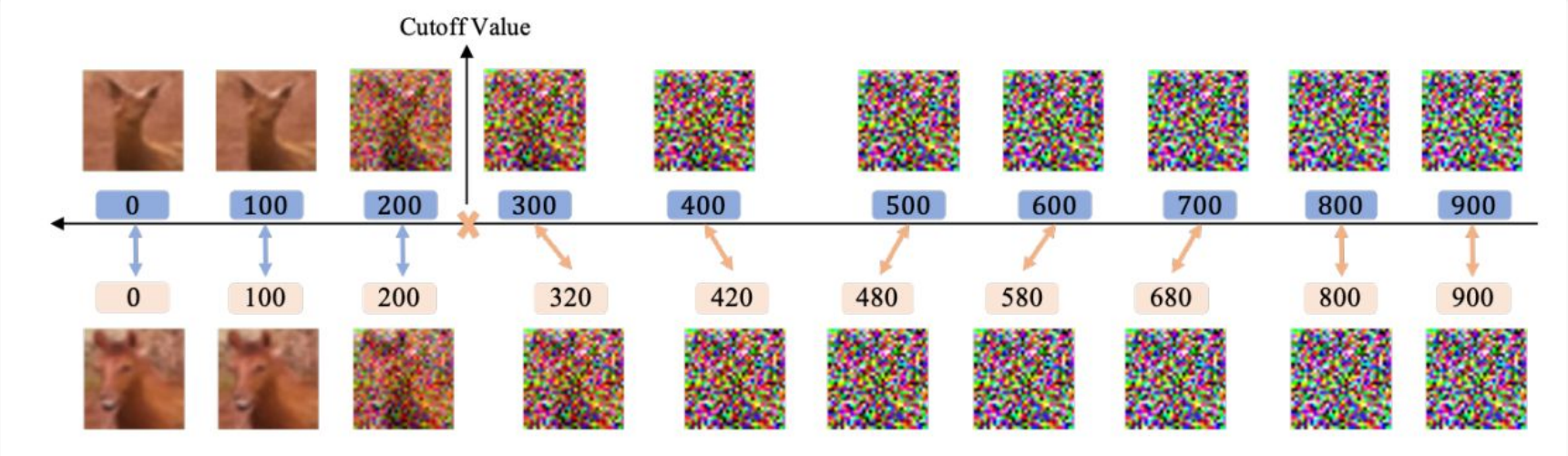
Performance comparison on CIFAR-10 with ADM and ADM-IP as the backbone models.

Influence of Window Sizes and Cutoff Values



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.

Case Study



Example of generation process of TS-DDIM and DDIM on CIFAR-10 using the horizontal black arrow to represent the timeline, with the DDIM generation chain above it, TS-DDIM generation chain underneath it. Sample 10 steps with window [t − 20, t + 20] and cutoff value=200.

Limitation and Future Work

Sampler suffers from the limitation that it introduces two parameters, i.e., window size and cutoff value.

More advanced methods to analytically derive the optimal value of these two parameters might be possible since both the cutoff value and the window size are related to the noise level of each step

Thank you!

- Thank you for your attention!
- I appreciate your time and interest.
- If you have any questions, please feel free to ask.
- Contact information: alimohammadiamirhossein@gmail.com