EFFICIENT DENOISING DIFFUSION VIA PROBABILISTIC MASKING

Anonymous authors
Paper under double-blind review

A DIFFUSION MODEL WITH PROBABILISTIC MASKS

The masked *forward process* is

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m}_t) &= \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t \mathbf{m}_t}\mathbf{x}_{t-1}, \beta_t \mathbf{m}_t \mathbf{I}); \\ q(\mathbf{x}_{1:T}, \mathbf{m}|\mathbf{x}_0) &= q(\mathbf{x}_{1:T}|\mathbf{x}_0, \mathbf{m}) p_{\mathbf{s}}(\mathbf{m}) \\ &= p_{\mathbf{s}}(\mathbf{m}) \Pi_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m}_t) \\ &= \Pi_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m}_t) p_{\mathbf{s}}(\mathbf{m}_t) \end{aligned}$$

The masked *reverse process* is

$$\begin{aligned} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m}_t) &= \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{m}_t, t), \mathbf{m}_t \Sigma_\theta(\mathbf{x}_t, t)) \\ p_\theta(\mathbf{x}_{0:T}, \mathbf{m}) &= p_{\mathbf{s}}(\mathbf{m}) p_\theta(\mathbf{x}_{0:T}|\mathbf{m}) \\ &= p_{\mathbf{s}}(\mathbf{m}) p(\mathbf{x}_T) \Pi_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m}_t) \\ &= p(\mathbf{x}_T) \Pi_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m}_t) p_{\mathbf{s}}(\mathbf{m}_t) \end{aligned}$$

The reduced variance variational bound for diffusion models can be written as:

$$\begin{aligned} -\log p_\theta(\mathbf{x}_0) &= -\log \int p_\theta(\mathbf{x}_{0:T}, \mathbf{m}) d\mathbf{x}_{1:T} d\mathbf{m} \\ &= -\log \int \frac{p_\theta(\mathbf{x}_{0:T}, \mathbf{m})}{q(\mathbf{x}_{1:T}, \mathbf{m}|\mathbf{x}_0)} q(\mathbf{x}_{1:T}, \mathbf{m}|\mathbf{x}_0) d\mathbf{x}_{1:T} d\mathbf{m} \\ &= -\log \mathbb{E}_q \frac{p_\theta(\mathbf{x}_{0:T}, \mathbf{m})}{q(\mathbf{x}_{1:T}, \mathbf{m}|\mathbf{x}_0)} \\ &\leq -\mathbb{E}_q \log \frac{p_\theta(\mathbf{x}_{0:T}, \mathbf{m})}{q(\mathbf{x}_{1:T}, \mathbf{m}|\mathbf{x}_0)} \\ &= -\mathbb{E}_q \log \frac{p(\mathbf{x}_T|\mathbf{m}) \Pi_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m}) p_{\mathbf{s}}(\mathbf{m}_t)}{\Pi_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m}) p_{\mathbf{s}}(\mathbf{m}_t)} \\ &= -\mathbb{E}_q \left[\log p(\mathbf{x}_T|\mathbf{m}) + \sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m})}{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m})} \right] =: \mathcal{L} \end{aligned}$$

$$\begin{aligned} q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m}) =& q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{x}_{t-1}, \mathbf{m}) \\ =& \frac{q(\mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_{t-1}, \mathbf{m})}{q(\mathbf{x}_0, \mathbf{x}_{t-1}, \mathbf{m})} \\ =& \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m}) q(\mathbf{x}_t, \mathbf{x}_0, \mathbf{m})}{q(\mathbf{x}_0, \mathbf{x}_{t-1}, \mathbf{m})} \\ =& \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m}) q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{m}) q(\mathbf{x}_0, \mathbf{m})}{q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{m}) q(\mathbf{x}_0, \mathbf{m})} \\ =& q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m}) \frac{q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{m})}{q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{m})} \end{aligned}$$

$$
\begin{aligned}
\mathcal{L} &= -\mathbb{E}_q \left[\log p(\mathbf{x}_T|\mathbf{m}) + \sum_{t=1}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m})}{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m})} \right] \\
&= -\mathbb{E}_q \left[\log p(\mathbf{x}_T|\mathbf{m}) + \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m})}{q(\mathbf{x}_t|\mathbf{x}_{t-1}, \mathbf{m})} + \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1, \mathbf{m})}{q(\mathbf{x}_1|\mathbf{x}_0, \mathbf{m})} \right] \\
&= -\mathbb{E}_q \left[\log p(\mathbf{x}_T|\mathbf{m}) + \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m})}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m})} \cdot \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0, \mathbf{m})}{q(\mathbf{x}_t|\mathbf{x}_0, \mathbf{m})} + \log \frac{p_\theta(\mathbf{x}_0|\mathbf{x}_1, \mathbf{m})}{q(\mathbf{x}_1|\mathbf{x}_0, \mathbf{m})} \right] \\
&= -\mathbb{E}_q \left[\log \frac{p(\mathbf{x}_T|\mathbf{m})}{q(\mathbf{x}_T|\mathbf{x}_0, \mathbf{m})} + \sum_{t=2}^{T} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m})}{q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m})} + \log p_\theta(\mathbf{x}_0|\mathbf{x}_1, \mathbf{m}) \right] \\
&= \mathbb{E}_q \left[\underbrace{D_{\mathrm{KL}}\left(q(\mathbf{x}_T|\mathbf{x}_0, \mathbf{m}) \| p(\mathbf{x}_T|\mathbf{m})\right)}_{\mathcal{L}^T} + \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}\left(q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m}) \| p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m})\right)}_{\mathcal{L}^{t-1}} - \log p_\theta(\mathbf{x}_0|\mathbf{x}_1, \mathbf{m}) \right]
\end{aligned}
$$

Note that

$$
\begin{aligned}
q(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{x}_0, \mathbf{m}) &= \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\mu}(\mathbf{x}_t, \mathbf{x}_0), \tilde{\beta}_t \mathbf{I}) \\
\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\bar{\alpha}_{t-1}(\mathbf{m})}\beta_t \mathbf{m}_t}{1 - \alpha_t(\mathbf{m})} \mathbf{x}_0 + \frac{\sqrt{\alpha_t(\mathbf{m})}(1 - \bar{\alpha}_{t-1}(\mathbf{m}))}{1 - \bar{\alpha}_t(\mathbf{m})} \mathbf{x}_t, \\
\tilde{\beta}_t &= \frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \mathbf{m}_t \beta_t.
\end{aligned}
$$

where

$$
\alpha_t(\mathbf{m}) = 1 - \mathbf{m}_t \beta_t \text{ and } \bar{\alpha}_t(\mathbf{m}) = \Pi_{i=1}^{t} \alpha_i(\mathbf{m}).
$$

For the reverse process, we have

$$
p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t, \mathbf{m}) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, \mathbf{m}, t), \tilde{\sigma}_t^2(\mathbf{m}) \mathbf{I}).
$$

Therefore,

$$
\begin{aligned}
\mathcal{L}^{t-1} &= \begin{cases} 0, & \text{if } \mathbf{m}_t = 0 \\ \frac{1}{2}\left[n \frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \frac{\mathbf{m}_t \beta_t}{\tilde{\sigma}_t^2(\mathbf{m})} - n + \frac{1}{\tilde{\sigma}_t^2(\mathbf{m})} \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, \mathbf{m}, t)\|^2 + n \log\left(\frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \frac{\mathbf{m}_t \beta_t}{\tilde{\sigma}_t^2(\mathbf{m})} \right) \right], & \text{otherwise} \end{cases} \\
&= \begin{cases} 0, & \text{if } \mathbf{m}_t = 0 \\ \frac{1}{2\tilde{\sigma}_t^2(\mathbf{m})} \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, \mathbf{m}, t)\|^2 + \frac{n}{2}\left[\frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \frac{\mathbf{m}_t \beta_t}{\tilde{\sigma}_t^2(\mathbf{m})} - 1 + \log\left(\frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \frac{\mathbf{m}_t \beta_t}{\tilde{\sigma}_t^2(\mathbf{m})} \right) \right], & \text{otherwise} \end{cases} \\
&= \begin{cases} 0, & \text{if } \mathbf{m}_t = 0 \\ \frac{1}{2\tilde{\sigma}_t^2(\mathbf{m})} \|\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \mu_\theta(\mathbf{x}_t, \mathbf{m}, t)\|^2 + C(\mathbf{m}) & \text{otherwise} \end{cases}
\end{aligned}
$$

where

$$
C(\mathbf{m}) = \frac{n}{2}\left[\frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \frac{\mathbf{m}_t \beta_t}{\tilde{\sigma}_t^2(\mathbf{m})} - 1 + \log\left(\frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \frac{\mathbf{m}_t \beta_t}{\tilde{\sigma}_t^2(\mathbf{m})} \right) \right].
$$

In this paper, following DDPM, we choose

$$
\tilde{\sigma}_t^2(\mathbf{m}) = \frac{1 - \bar{\alpha}_{t-1}(\mathbf{m})}{1 - \bar{\alpha}_t(\mathbf{m})} \mathbf{m}_t \beta_t.
$$

In this case,

$$
C(\mathbf{m}) = 0.
$$

For $\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0)$, since

$$
\mathbf{x}_t(\mathbf{x}_0, \epsilon) = \sqrt{\bar{\alpha}_t(\mathbf{m})} \mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t(\mathbf{m})} \epsilon \text{ with } \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),
$$

we have

$$\begin{aligned}\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) &= \frac{\sqrt{\bar{\alpha}_{t-1}(\mathbf{m})}\mathbf{m}_t\beta_t}{1-\bar{\alpha}_t(\mathbf{m})}\mathbf{x}_0 + \frac{\sqrt{\alpha_t(\mathbf{m})}(1-\bar{\alpha}_{t-1}(\mathbf{m}))}{1-\bar{\alpha}_t(\mathbf{m})}\mathbf{x}_t(\mathbf{x}_0,\epsilon) \\ &= \frac{\sqrt{\bar{\alpha}_{t-1}(\mathbf{m})}\mathbf{m}_t\beta_t}{1-\bar{\alpha}_t(\mathbf{m})}\frac{1}{\sqrt{\bar{\alpha}_t(\mathbf{m})}}\left(\mathbf{x}_t(\mathbf{x}_0,\epsilon) - \sqrt{1-\bar{\alpha}_t(\mathbf{m})}\epsilon\right) + \frac{\sqrt{\alpha_t(\mathbf{m})}(1-\bar{\alpha}_{t-1}(\mathbf{m}))}{1-\bar{\alpha}_t(\mathbf{m})}\mathbf{x}_t(\mathbf{x}_0,\epsilon) \\ &= \frac{1}{\sqrt{\alpha_t(\mathbf{m})}}\left(\mathbf{x}_t(\mathbf{x}_0,\epsilon) - \frac{\mathbf{m}_t\beta_t}{\sqrt{1-\bar{\alpha}_t(\mathbf{m})}}\epsilon\right).\end{aligned}$$

Hence, we define

$$\mu(\mathbf{x}_t, \mathbf{m}, t) = \frac{1}{\sqrt{\alpha_t(\mathbf{m})}}\left(\mathbf{x}_t - \frac{\mathbf{m}_t\beta_t}{\sqrt{1-\bar{\alpha}_t(\mathbf{m})}}\epsilon_\theta(\mathbf{x}_t, t)\right).$$

Then, we have

$$\begin{aligned}&\frac{1}{2\tilde{\sigma}_t^2(\mathbf{m})}\left\|\tilde{\mu}_t(\mathbf{x}_t,\mathbf{x}_0) - \mu_\theta(\mathbf{x}_t,\mathbf{m},t)\right\|^2 \\ =&\frac{\mathbf{m}_t\beta_t^2}{2\tilde{\sigma}_t^2(\mathbf{m})\alpha_t(\mathbf{m})(1-\bar{\alpha}_t(\mathbf{m}))}\left\|\epsilon - \epsilon_\theta\left(\mathbf{x}_t, t\right)\right\|^2 \\ =&\frac{\mathbf{m}_t\beta_t^2}{2\tilde{\sigma}_t^2(\mathbf{m})\alpha_t(\mathbf{m})(1-\bar{\alpha}_t(\mathbf{m}))}\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t(\mathbf{m})}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t(\mathbf{m})}\epsilon, t\right)\right\|^2\end{aligned}$$

Finally, we get the loss as follows:

$$\mathcal{L}^{t-1} = \begin{cases}0, & \text{if } \mathbf{m}_t = 0 \\ \frac{\mathbf{m}_t\beta_t^2}{2\tilde{\sigma}_t^2(\mathbf{m})\alpha_t(\mathbf{m})(1-\bar{\alpha}_t(\mathbf{m}))}\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t(\mathbf{m})}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t(\mathbf{m})}\epsilon, t\right)\right\|^2, & \text{otherwise}\end{cases}$$

Thus, we get the objective function in the maintext.

B BASICS

B.1 MUTUAL SKIPPING OF SAMPLING STEPS

To improve the efficiency of sample generation process, previous methods Song et al. (2020); Bao et al. (2022a;b) always manually select the denoising steps through uniform skipping and quadratic skipping. The mathematical expression of the above skipping approaches can be written as:

$$\mathbb{T} = \{1, 1+S, ..., 1+iS, ..., L\}, \text{ with } S = \begin{cases}\frac{T}{L} & , \text{ uniform skipping,} \\ \left(\frac{0.8T}{L}\right)^2 & , \text{ quadratic skipping.}\end{cases} \tag{1}$$

where $i = 1, \ldots, L$. T and L are the number of diffusion steps and number of denoising steps in the training and testing state respectively. S is the skipping step. The difference of T and L results in decoupled forward and reverse processes, which makes a suboptimal performance. Instead, our proposed probabilistic masking method can identify and keep the most informative steps during training.

B.2 MULTIVARIATE TIME SERIES IMPUTATION

Let us denote each time series as $\mathbf{X} \in R^{K\times P}$, where K is the number of features and P is the length of time series. Probabilistic time series imputation is to estimate the missing values of X by exploiting the observed values of X. The diffusion model is used to estimate the true conditional data distribution $q(x_0^t|x_0^c)$, where x_0^t and x_0^c are the imputation targets and conditional observations respectively.

C EXPERIMENTS

C.1 TIME SERIES DATASETS

Healthcare dataset Silva et al. (2012) consists of 4000 clinical time series with 35 variables for 48 hours from intensive care unit (ICU), and it contains around 80% missing values. Following previous study Tashiro et al. (2021), we randomly choose 10/50/90% of observed values as ground-truth on the test data for imputation.

Air-quality dataset is composed of air quality data from 36 stations in Beijing from 2014/05/01 to 2015/04/30, and it has around 13% missing values. We set 36 consecutive time steps as one time series. To build missing values in the time series, we follow the empirical settings of the baseline Tashiro et al. (2021), we adopt the random strategy for the healthcare dataset and the mix of the random and historical strategy for the air quality dataset.

C.2 IMPLEMENT DETAILS

All the experiments are implemented by Pytorch 1.7.0 on a virtual workstation with 8 11G memory Nvidia GeForce RTX 2080Ti GPUs.

Time series. As for model hyper-parameters, we set the batch size as 16 and the number of epochs as 200. We used Adam Kingma & Ba (2014) optimizer with learning rate 0.001 that is decayed to 0.0001 and 0.00001 at 75% and 90% of the total epochs, respectively. For the diffusion model, we follow the CSDI Tashiro et al. (2021) architecture to set the number of residual layers as 4, residual channels as 64, and attention heads as 8. The denoising step T is set to 50 as our baseline.

Image data. Following Nichol & Dhariwal (2021), we use the U-Net model architecture, train 500K iterations with a batch size of 128, use a learning rate of 0.0001 with the Adam Kingma & Ba (2014) optimizer and use an exponential moving average (EMA) with a rate of 0.9999. The denoising step T is set to 1000 and the linear forward noise schedule is used as our baseline.

C.3 EVALUATION METRIC

The detailed formulations of three metrics for time series task are:

$$MAE(x, \hat{x}) = \frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|, \tag{2}$$

$$RMSE(x, \hat{x}) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2}, \tag{3}$$

$$CRPS(F, \hat{F}) = \int_{-\infty}^{\infty} \left[F(z) - \hat{F}(z)\right]^2 dz, \tag{4}$$

where x denotes the ground truth of the missed time series, $\hat{x}$ represents the predicted values. F is the cumulative distribution function of observations.

C.4 MAIN RESULTS

As shown in Table 1, our proposed EDDPM can achieve better results than the baselines with 100% steps (blue text) even if 60% $\sim$ 75% denoising steps are masked. These results are consistent with the conclusion of the main paper.

C.5 VISUALIZATION RESULTS

From the results illustrated in Figure 1, we can conclude that our proposed EDDPM can generate more accurate probabilistic imputation results by only using the original 20% $\sim$ 50% steps.

For CIFAR-10 image generation, Figure 2 and 4 show that our proposed EDDPM can generate more high-quality image samples than DDIM Song et al. (2020) when using 10 denoising steps. Figure 3

Table 1: Comparising sampling acceleration methods in terms of **CRPS** results on variable denoising steps. † indicate that the sampling is accelerated by quadratic skipping during inference, the others utilize uniform skipping. We highlight the best results that surpass the baselines in red color, which means our method generates high-quality time series with fewer denoising steps. The **bold** results show that our proposed EDDPM achieves better performance than other sampling acceleration methods.

Dataset	Missing	Method	Denoising steps				Baselines
			10%	25%	40%	50%	
Healthcare	10%	DDPM†	0.688	0.501	0.382	0.326	0.238
		DDPM	0.640	0.431	0.344	0.276	
		DDIM	0.641	0.495	0.564	0.840	
		AnalyticDPM	0.615	0.536	0.516	0.501	
		SN-DDPM	0.769	0.757	0.762	0.769	
		NPR-DDIM	0.573	0.502	0.504	0.516	
		Ours	**0.267**	0.237	0.235	0.231	
	50%	DDPM†	0.699	0.582	0.490	0.439	0.331
		DDPM	0.675	0.516	0.437	0.372	
		DDIM	0.675	0.562	0.601	0.810	
		AnalyticDPM	0.698	0.586	0.572	0.579	
		SN-DDPM	0.761	0.752	0.759	0.772	
		NPR-DDIM	0.612	0.546	0.547	0.561	
		Ours	**0.357**	**0.337**	0.321	0.330	
	90%	DDPM †	0.731	0.690	0.648	0.622	0.522
		DDPM	0.737	0.654	0.594	0.557	
		DDIM	0.737	0.695	0.715	0.856	
		AnalyticDPM	0.715	0.685	0.672	0.668	
		SN-DDPM	0.840	0.810	0.808	0.810	
		NPR-DDIM	0.704	0.647	0.643	0.644	
		Ours	**0.572**	0.517	0.516	0.513	
			4%	10%	20%	40%	
Air-quality	13%	DDPM†	0.568	0.482	0.374	0.217	0.109
		DDPM	0.536	0.453	0.344	0.209	
		DDIM†	0.569	0.507	0.464	0.605	
		DDIM	0.537	0.485	0.619	1.553	
		AnalyticDPM	0.489	0.453	0.429	0.382	
		SN-DDPM	0.557	0.507	0.482	0.481	
		SN-DDIM	0.653	0.568	0.558	0.582	
		NPR-DDPM	0.359	0.355	0.377	0.395	
		NPR-DDIM	0.362	0.344	0.305	0.271	
		Ours	**0.170**	**0.133**	**0.112**	0.104	

and 5 show the sample pairs generated by our EDDPM with 5, 10 and 100 denoising steps, from these results we can conclude that our method generate high-quality CIFAR-10 images using 5 steps.

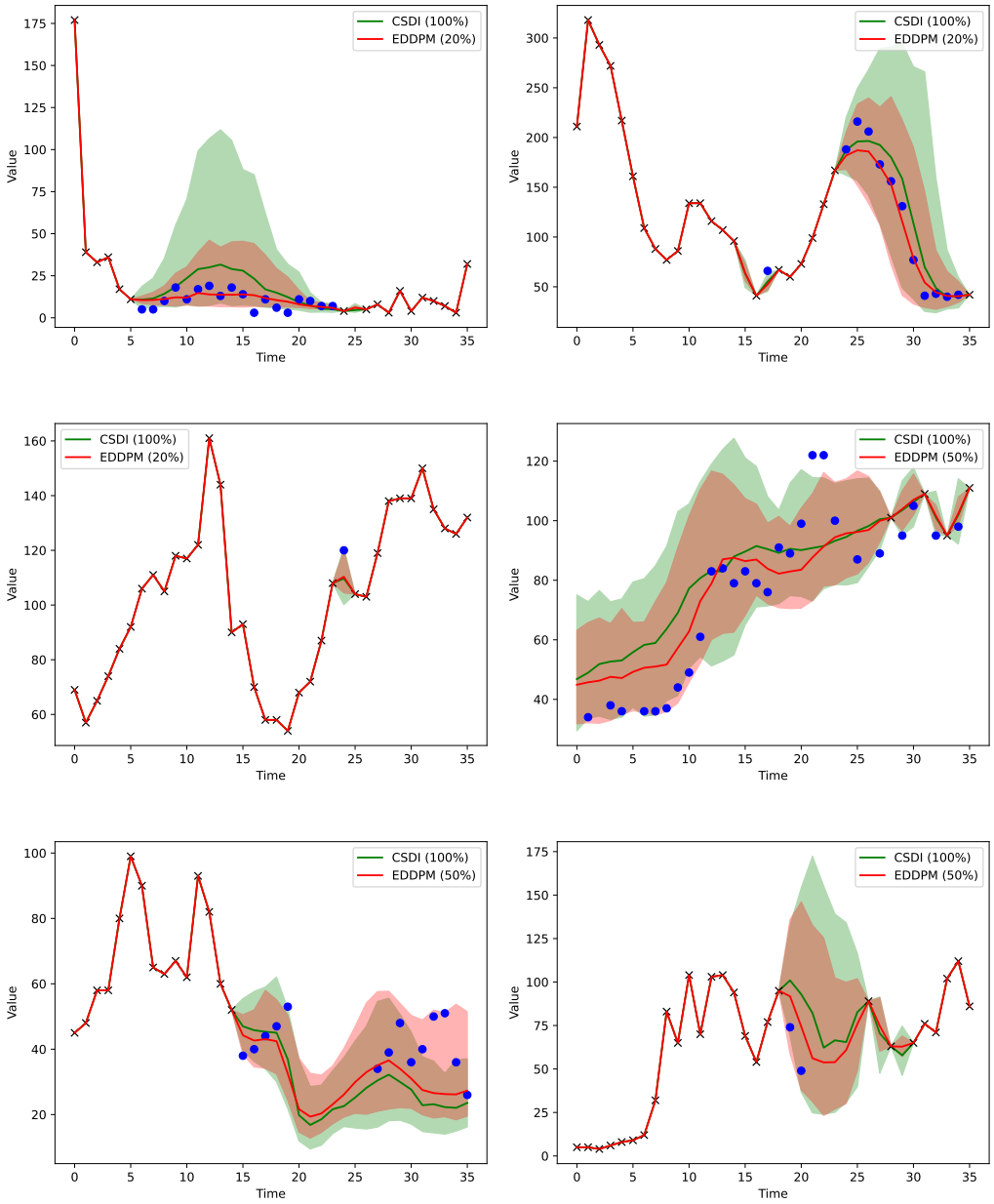
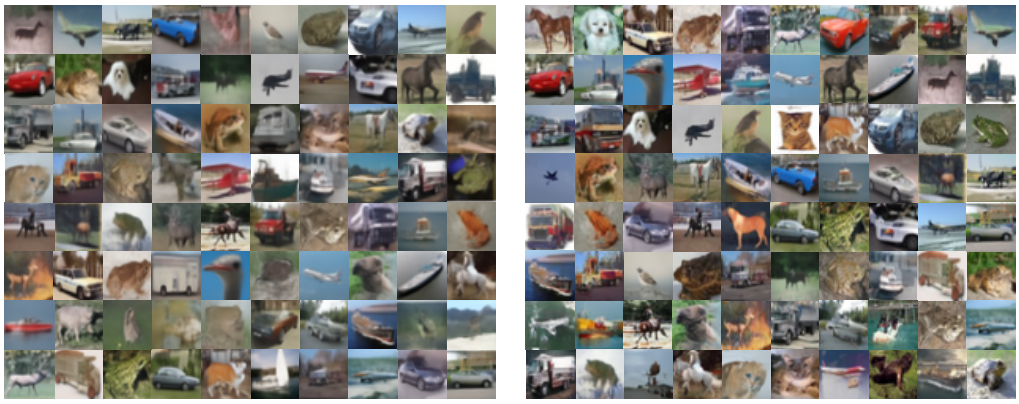
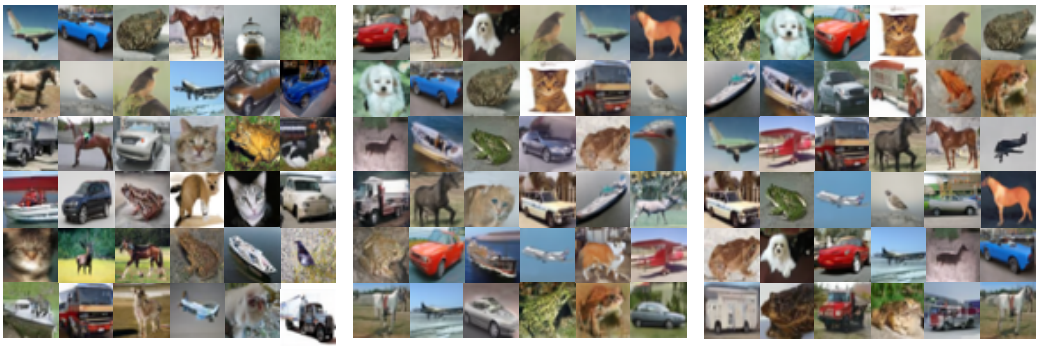


Figure 1: The comparison of our EDDPM method and DDPM Ho et al. (2020) for probabilistic time series imputation on Air-quality dataset. CSDI model is trained by DDPM. The black crosses show observed values and the blue circles show ground-truth imputation targets. red and green colors correspond to our EDDPM and CSDI, respectively. For each method, median values of imputations are shown as the line and 5% and 95% quantiles are shown as the shade.



DDIM EDDPM

Figure 2: Random samples generated by DDIM Song et al. (2020) and EDDPM (ours) with 10 denoising steps on CIFAR-10 dataset. We only present the result in this extreme sparse case since the results for more denoising steps are difficult to differentiate for human beings.



$T = 100$ $T = 10$ $T = 5$

Figure 3: Random samples generated by our EDDPM with 5, 10 and 100 denoising steps on CIFAR-10 dataset.

Figure 4: Sample pair comparison based on DDIM Song et al. (2020) and EDDPM (ours) with 10 denoising steps on CIFAR-10 dataset when $T = 10$. We can see that our method can generate images with more details.

Figure 5: Random samples generated by our EDDPM with 5, 10 and 100 denoising steps on CIFAR-10 dataset.

REFERENCES

Fan Bao, Chongxuan Li, Jiacheng Sun, Jun Zhu, and Bo Zhang. Estimating the optimal covariance with imperfect mean in diffusion probabilistic models. *arXiv preprint arXiv:2206.07309*, 2022a.

Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. *arXiv preprint arXiv:2201.06503*, 2022b.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171. PMLR, 2021.

Ikaro Silva, George Moody, Daniel J Scott, Leo A Celi, and Roger G Mark. Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012. In *2012 Computing in Cardiology*, pp. 245–248. IEEE, 2012.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in Neural Information Processing Systems*, 34, 2021.