Denoising Diffusion Implicit Models

ICLR 2021 Poster

Jiaming Song, Chenlin Meng, Stefano Ermon
Stanford University

马逸扬
2023/10/08

Content

- Background
- Algorithm
- Proofs
- Experiments

Content

- Background
- Algorithm
- Proofs
- Experiments

Background

DDPMs suffer from the problem of randomness.

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \ldots, 1$ **do**
3: $\quad \mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\quad \mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

Such randomness leads to two problems:

- Instability.
- Long sampling time.

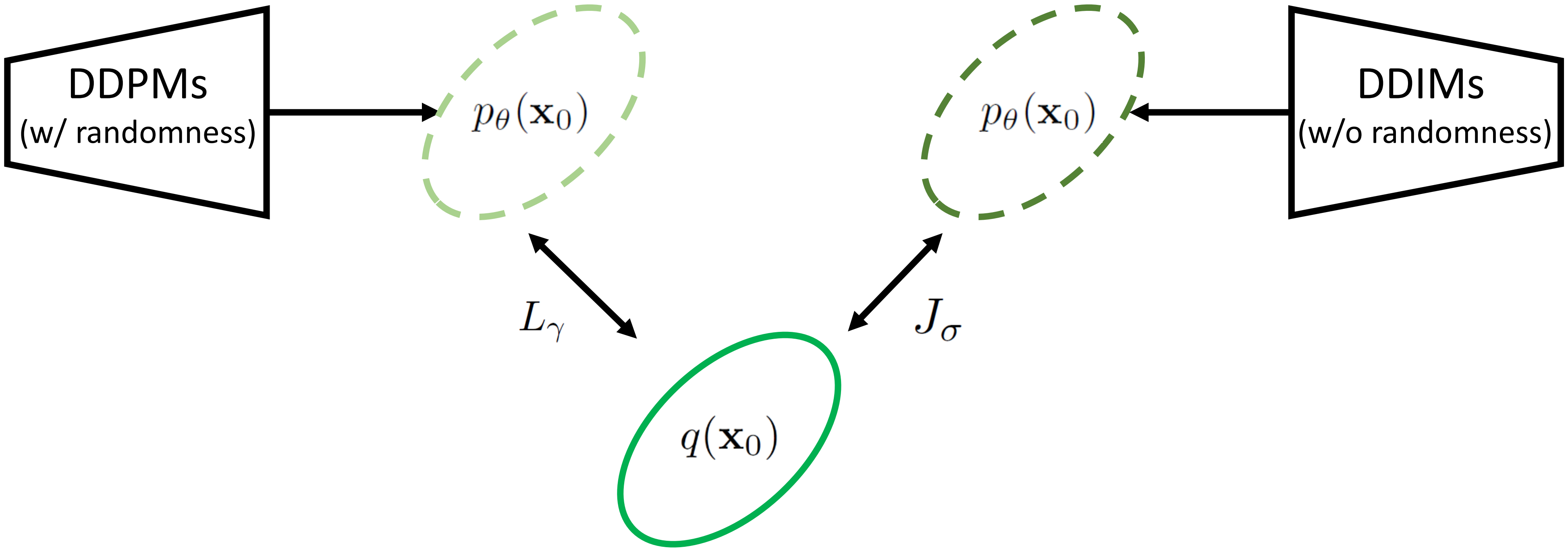
Content

- Background
- Algorithm
- Proofs
- Experiments

Algorithm

DDIMs aim to model a deterministic sampling process of diffusion models.

The core idea of DDIMs:



$$J_\sigma = L_\gamma + C$$

Algorithm

Build a set of reverse processes:

$$q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) := q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0)\prod_{t=2}^{T} q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \tag{1}$$

where

$$q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_T}\boldsymbol{x}_0, (1-\alpha_T)\boldsymbol{I}) \tag{2}$$

and

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right) \tag{3}$$

The above distribution is **CHOSEN** to ensure (the proof refers to Pg. 17):

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0, (1-\alpha_t)\boldsymbol{I}) \tag{4}$$

Algorithm

By predicting the noise in the t-th state, we can reconstruct a pseudo clean image through:

$$f_\theta^{(t)}(\boldsymbol{x}_t) := (\boldsymbol{x}_t - \sqrt{1-\alpha_t} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t} \tag{5}$$

Thus, we can build the corresponding reverse processes:

$$p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \begin{cases} \mathcal{N}(f_\theta^{(1)}(\boldsymbol{x}_1), \sigma_1^2 \boldsymbol{I}) & \text{if } t = 1 \\ q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)) & \text{otherwise} \end{cases} \tag{6}$$

The model is called DDIM.

Algorithm

Thus, we can build the ELBO of DDIMs (which has the same form of DDPMs'):

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}[\log q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) - \log p_\theta(\boldsymbol{x}_{0:T})] \tag{7}$$

which can be expanded as:

$$\mathbb{E}_{\boldsymbol{x}_{0:T} \sim q_\sigma(\boldsymbol{x}_{0:T})}\left[\log q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) - \log p_\theta(\boldsymbol{x}_T)\right] \tag{8}$$

We provide the ELBO of DDPMs here:

$$\mathbb{E}_q\left[-\log p(\mathbf{x}_T) - \sum_{t \geq 1} \log \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_t|\mathbf{x}_{t-1})}\right] =: L \tag{9}$$

It can be proven that (proof refers to Pg. 19):

$$J_\sigma = L_\gamma + C \tag{10}$$

where γ is the set of weights of items in L.

Algorithm

The standard deviations of the proposed reverse processes are not limited.

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1 - \alpha_{t-1} - \sigma_t^2} \cdot \frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1 - \alpha_t}}, \sigma_t^2 \boldsymbol{I}\right) \tag{3}$$

Thus, if we set all the $\sigma_t = 0$, we can obtain a deterministic reverse process.

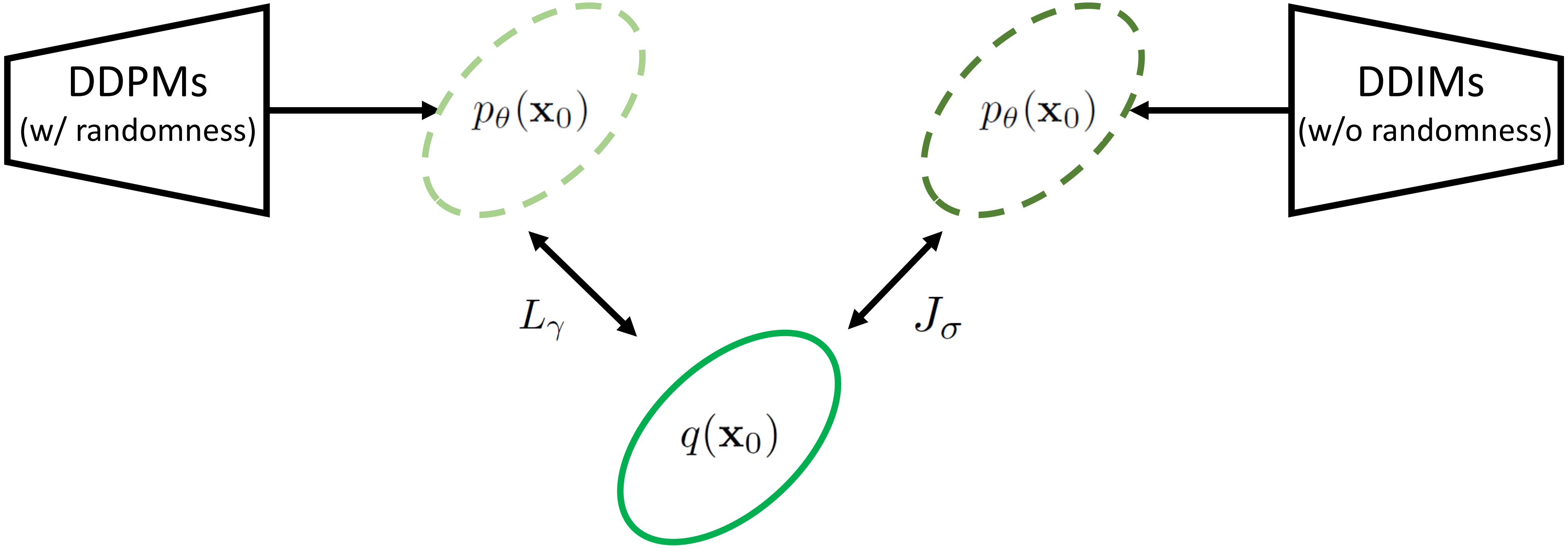
Recall that:

$$J_\sigma = L_\gamma + C \tag{10}$$

which means we can directly employ a pre-trained noise-prediction model by DDPM in the reverse process of DDIM without any additional training.

Algorithm

The core idea of DDIMs:



$$J_\sigma = L_\gamma + C$$

Content

- Background
- Algorithm
- Proofs
- Experiments

Proofs

Lemma 1. With $q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0)$ and $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$ defined in $q_\sigma(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)$, we have:

$$q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0,(1-\alpha_t)\boldsymbol{I}) \tag{4}$$

Proof. Assume for all $t \leq T$, $q_\sigma(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_t}\boldsymbol{x}_0,(1-\alpha_t)\boldsymbol{I})$ holds. If:

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0, \sqrt{1-\alpha_{t-1}}\boldsymbol{I}) \tag{11}$$

then lemma is proven, since case at $t = T$ already holds:

$$q_\sigma(\boldsymbol{x}_T|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_T}\boldsymbol{x}_0,(1-\alpha_T)\boldsymbol{I}) \tag{2}$$

We have:

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\frac{\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2\boldsymbol{I}\right) \tag{3}$$

Proofs

We have:

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}}, \sigma_t^2 \boldsymbol{I}\right) \tag{3}$$

Thus, the mean and variance of $q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)$ are:

$$\begin{aligned} \mathbb{E}[q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)] &= \sqrt{\alpha_{t-1}}\boldsymbol{x}_0 + \sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \frac{\sqrt{\alpha_t}\boldsymbol{x}_0 - \sqrt{\alpha_t}\boldsymbol{x}_0}{\sqrt{1-\alpha_t}} \\ &= \sqrt{\alpha_{t-1}}\boldsymbol{x}_0 \end{aligned} \tag{12}$$

and:

$$\mathrm{Cov}[q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)] = \sigma_t^2 \boldsymbol{I} + \frac{1-\alpha_{t-1}-\sigma_t^2}{1-\alpha_t}(1-\alpha_t)\boldsymbol{I} = (1-\alpha_{t-1})\boldsymbol{I} \tag{13}$$

Thus, we have

$$q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0) = \mathcal{N}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_0, \sqrt{1-\alpha_{t-1}}\boldsymbol{I}) \tag{11}$$

□

Proofs

Lemma 2. $J_\sigma = L_\gamma + C$

Proof. The definition of J_σ is:

$$J_\sigma(\epsilon_\theta) := \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q(\boldsymbol{x}_{0:T})} \left[\log q_\sigma(\boldsymbol{x}_T | \boldsymbol{x}_0) + \sum_{t=2}^{T} \log q_\sigma(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{x}_0) - \sum_{t=1}^{T} \log p_\theta^{(t)}(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t) \right]$$

$$\equiv \mathbb{E}_{\boldsymbol{x}_{0:T} \sim q(\boldsymbol{x}_{0:T})} \left[\sum_{t=2}^{T} D_{\mathrm{KL}}(q_\theta(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t, \boldsymbol{x}_0)) \| p_\theta^{(t)}(\boldsymbol{x}_{t-1} | \boldsymbol{x}_t)) - \log p_\theta^{(1)}(\boldsymbol{x}_0 | \boldsymbol{x}_1) \right] \quad (14)$$

Proofs

For $t > 1$, we have:

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}[D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)) \| p_\theta^{(t)}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))] \\
= &\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}[D_{\mathrm{KL}}(q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)) \| q_\sigma(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, f_\theta^{(t)}(\boldsymbol{x}_t)))] \\
= &\mathbb{E}_{\boldsymbol{x}_0, \boldsymbol{x}_t \sim q(\boldsymbol{x}_0, \boldsymbol{x}_t)}\left[\frac{\|\boldsymbol{x}_0 - f_\theta^{(t)}(\boldsymbol{x}_t)\|_2^2}{2\sigma_t^2}\right] \\
= &\mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon}\left[\frac{\|(\boldsymbol{x}_t - \epsilon)/\sqrt{\alpha_t} - (\boldsymbol{x}_t - \epsilon_\theta^{(t)}(\boldsymbol{x}_t))/\sqrt{\alpha_t}\|_2^2}{2\sigma_t^2}\right] \\
= &\mathbb{E}_{\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \epsilon \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_0 + \sqrt{1-\alpha_t}\epsilon}\left[\frac{\|\epsilon - \epsilon_\theta^{(t)}(\boldsymbol{x}_t)\|_2^2}{2d\sigma_t^2\alpha_t}\right]
\end{aligned} \tag{15}
$$

where d is the dimension of $\boldsymbol{x}_0$ (which seems a typo?).

Proofs

For $t = 1$, we have:

$$\mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_1\sim q(\boldsymbol{x}_0,\boldsymbol{x}_1)}\left[-\log p_\theta^{(1)}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right] \equiv \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{x}_1\sim q(\boldsymbol{x}_0,\boldsymbol{x}_1)}\left[\frac{\left\|\boldsymbol{x}_0 - f_\theta^{(t)}(\boldsymbol{x}_1)\right\|_2^2}{2\sigma_1^2}\right]$$

$$= \mathbb{E}_{\boldsymbol{x}_0\sim q(\boldsymbol{x}_0),\epsilon\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),\boldsymbol{x}_1=\sqrt{\alpha_1}\boldsymbol{x}_0+\sqrt{1-\alpha_t}\epsilon}\left[\frac{\left\|\epsilon - \epsilon_\theta^{(1)}(\boldsymbol{x}_1)\right\|_2^2}{2d\sigma_1^2\alpha_1}\right] \tag{16}$$

With $\gamma_t = 1/(2d\sigma_t^2\alpha_t)$, we have:

$$J_\sigma(\epsilon_\theta) \equiv \sum_{t=1}^{T}\frac{1}{2d\sigma_t^2\alpha_t}\mathbb{E}\left[\left\|\epsilon_\theta^{(t)}(\boldsymbol{x}_t) - \epsilon_t\right\|_2^2\right] = L_\gamma(\epsilon_\theta) \tag{17}$$

Thus, we have:

$$J_\sigma = L_\gamma + C \tag{10}$$

□

Content

- Background
- Algorithm
- Proofs
- Experiments

Results

For:

$$\sigma_{\tau_i}(\eta) = \eta\sqrt{(1-\alpha_{\tau_{i-1}})/(1-\alpha_{\tau_i})}\sqrt{1-\alpha_{\tau_i}/\alpha_{\tau_{i-1}}} \quad (18)$$

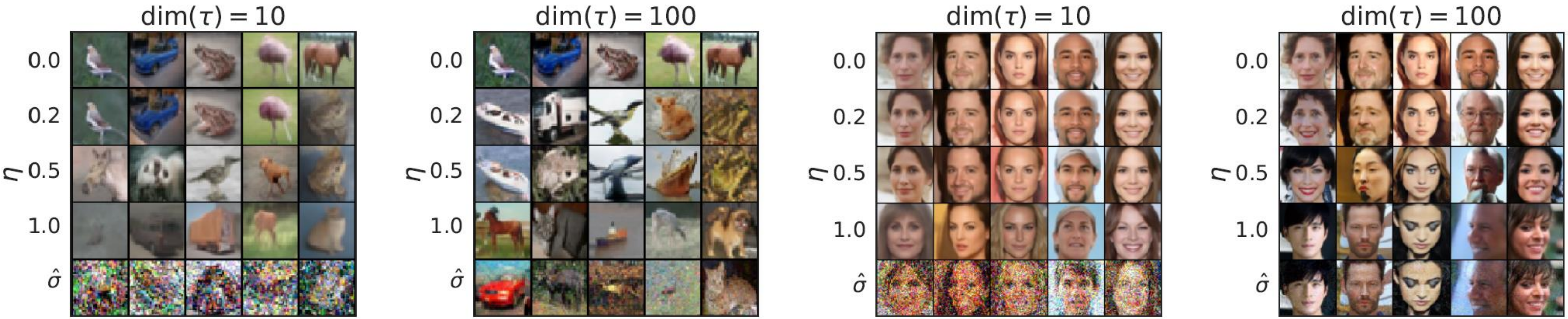
Experiment results

	S	CIFAR10 (32 × 32) 10	20	50	100	1000	CelebA (64 × 64) 10	20	50	100	1000
η	0.0	**13.36**	**6.84**	**4.67**	**4.16**	4.04	**17.33**	**13.73**	**9.17**	**6.53**	3.51
	0.2	14.04	7.11	4.77	4.25	4.09	17.66	14.11	9.51	6.79	3.64
	0.5	16.66	8.35	5.25	4.46	4.29	19.86	16.06	11.01	8.09	4.28
	1.0	41.07	18.36	8.01	5.78	4.73	33.12	26.03	18.48	13.93	5.98
$\hat{\sigma}$		367.43	133.37	32.72	9.99	**3.17**	299.71	183.83	71.71	45.20	**3.26**

(σ refers to $\sigma_t^2 = \tilde{\beta}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$ and $\hat{\sigma}$ refers to $\sigma_t^2 = \beta_t$ in DDPM)

Results

Visual results on CIFAR-10 and CelebA:



Discussion

DDIM? Diffusion ODE? DPM-Solver?

DDPM → Diffusion SDE → Diffusion ODE → DPM-Solver

Surprisingly, DDIM is the first-order solution to Diffusion ODE, *i.e.*, a special case of DPM-Solver.

Thanks for watching.

马逸扬
2023/10/08