Improving Probabilistic Diffusion Models With Optimal Diagonal Covariance Matching

ICLR 2025 (Oral)

Zijing Ou[1]*, Mingtian Zhang[2]*, Andi Zhang[3], Tim Z. Xiao[4], Yingzhen Li[1], David Barber[2]

[1]Imperial College London, [2]University College London, [3]University of Cambridge
[4]Max Planck Institute for Intelligent Systems, Tübingen, [5]University of Tübingen, [6]IMPRS-IS

***Equal contribution**

Denoising Diffusion Probabilistic Models

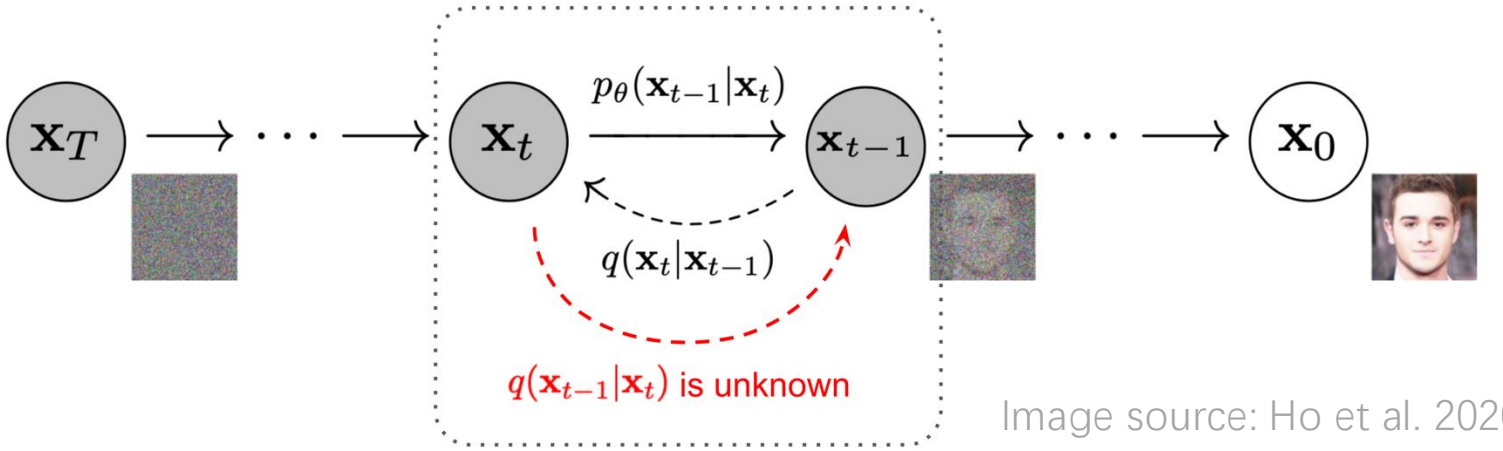


Image source: Ho et al. 2020

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.

Denoising Diffusion Probabilistic Models

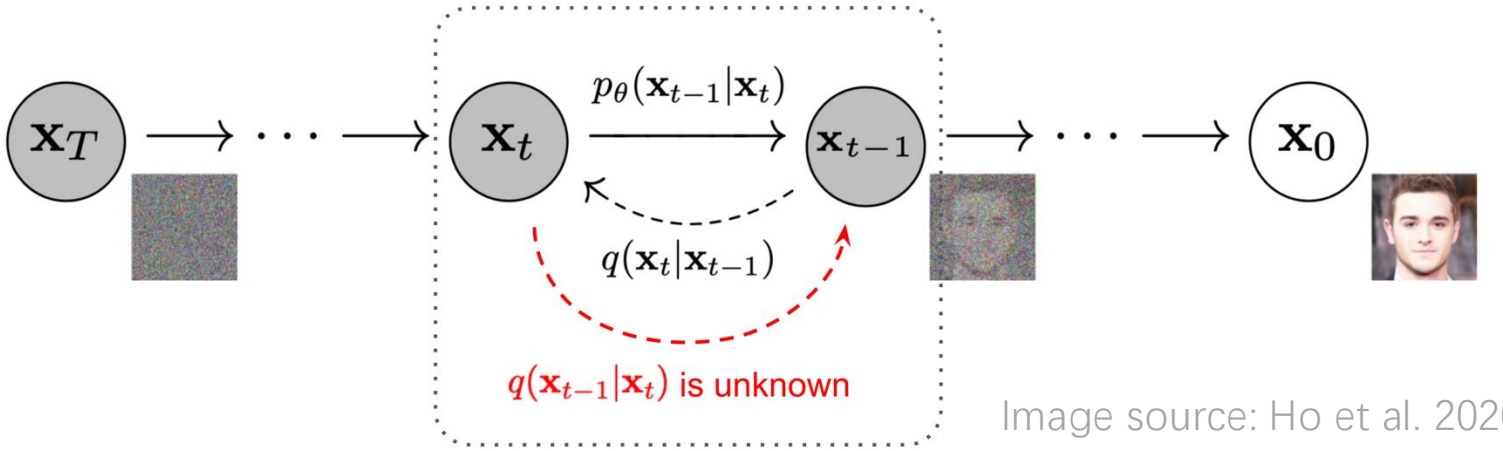


Image source: Ho et al. 2020

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \rightarrow \mathcal{N}(0, I)$

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.

Denoising Diffusion Probabilistic Models

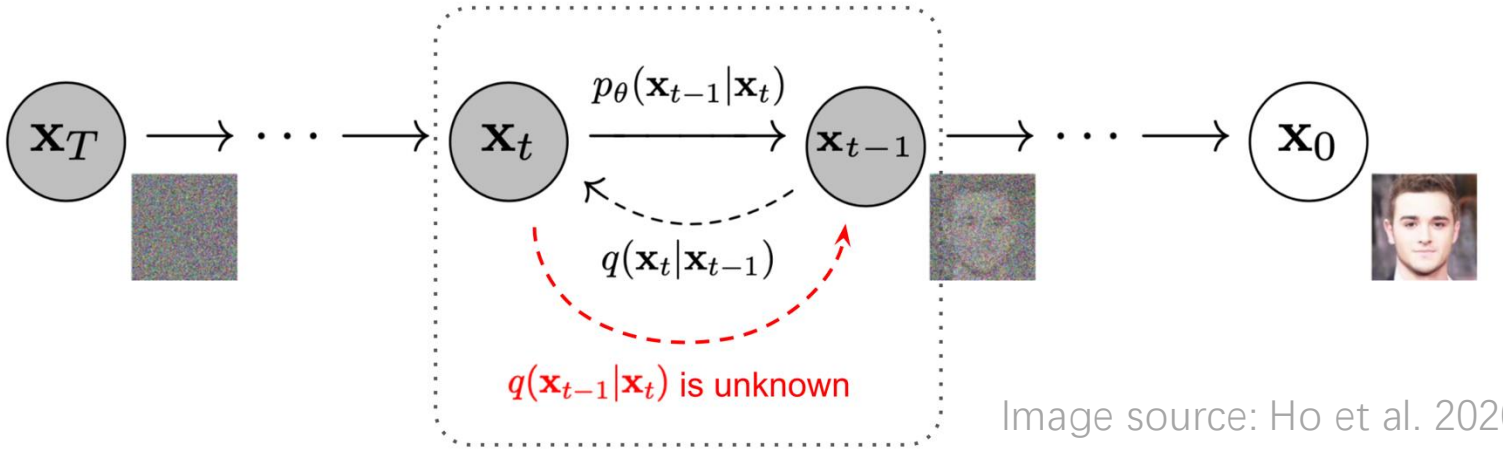


Image source: Ho et al. 2020

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

Gaussian denoising model assumption

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.

Mean of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

How to learn the mean function μ ?

Mean of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

1st-Order Tweedie's Formula

$$\mathbb{E}_q[x_{t-1}|x_t] = \frac{x_t + \sigma_t^2 \nabla_{x_t} \log q(x_t)}{\alpha_t}$$

Mean of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

1st-Order Tweedie's Formula

$$\mathbb{E}_q[x_{t-1}|x_t] = \frac{x_t + \sigma_t^2 \nabla_{x_t} \log q(x_t)}{\alpha_t}$$

Score Parameterization

$$\mu_{\theta_1}(x_t) = \frac{x_t + \sigma_t^2 s_{\theta_1}(x_t)}{\alpha_t}, \qquad s_{\theta_1}(x_t) \approx \nabla_{x_t} \log q(x_t)$$

Mean of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

Denoising Score Identity

$$\nabla_{x_t} \log q(x_t) = \mathbb{E}_{q(x_0|x_t)}[\nabla_{x_t} \log q(x_t|x_0)]$$

Denoising Score Matching

$$\underset{\theta_1}{\operatorname{argmin}}\, \mathbb{E}_{t,q(x_0,x_t)} \parallel s_{\theta_1}(x_t) - \nabla_{x_t} \log q(x_t|x_0) \parallel_2^2$$

Covariance of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

How to choose the covariance Σ ?

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.

Covariance of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

Heuristic choices of Σ gives good results when $T \to \infty$ 😄

$$\Sigma(\mathrm{x_t}) = \mathrm{Cov}[q(x_t|x_{t-1})] \text{ or } \Sigma(\mathrm{x_t}) = \mathrm{Cov}[q(x_{t-1}|x_t, x_0)]$$

Figure 1 in Ho et al. 2020

Ho et al. Denoising Diffusion Probabilistic Models. NeurIPS, 2020.

Covariance of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

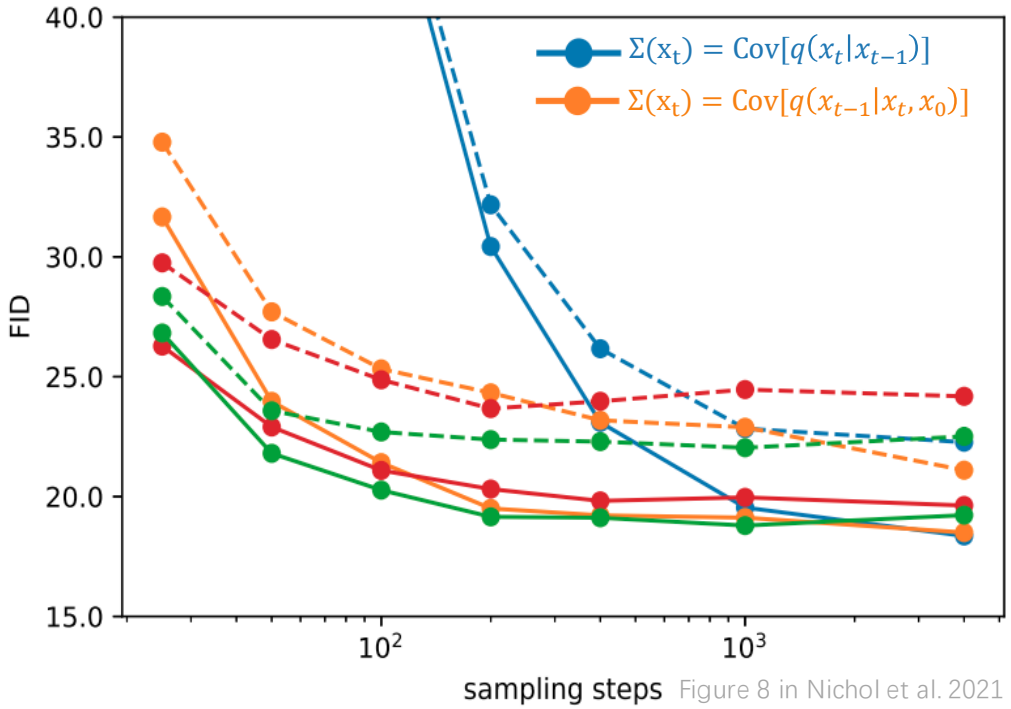


Figure 8 in Nichol et al. 2021

However, generation performance drops significantly with fewer denoising steps.

Nichol et al. Improved Denoising Diffusion Probabilistic Models. ICML, 2021.

Covariance of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

Do we have a better choice of Σ ? 🤔

Covariance of the Denoising Distribution

Forward Diffusion Process:

- $q(x_{0:T}) = q(x_0)\prod_{t=1}^{T} q(x_t|x_{t-1})$
- $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \alpha_t x_{t-1}, \sigma_t^2 I)$
- $q(x_T) \to \mathcal{N}(0, I)$

Backward Generation Process:

- $p_\theta(x_{0:T}) = q(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)$
- $p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \Sigma_{\theta_2}(x_t)\right)$
- $p_\theta(x_{t-1}|x_t) \approx q(x_{t-1}|x_t)$

Do we have a better choice of Σ ? 🤔

YES, WE DO! 😎

Optimal Covariance of the Denoising Distribution

Optimal Covariance: Given $q(x, x_t) = q(x)q(x_t|x)$ with $q(x_t|x) = \mathcal{N}(\alpha_t x, \sigma_t^2 I)$, the covariance of the posterior $q(x|x_t) \propto q(x)q(x_t|x)$ has a closed form

$$\Sigma(x_t) = (\sigma_t^4 \nabla_{x_t}^2 \log q(x_t) + \sigma_t^2 I)/\alpha_t^2$$

This is also known as the **2nd-Order Tweedie's Lemma**

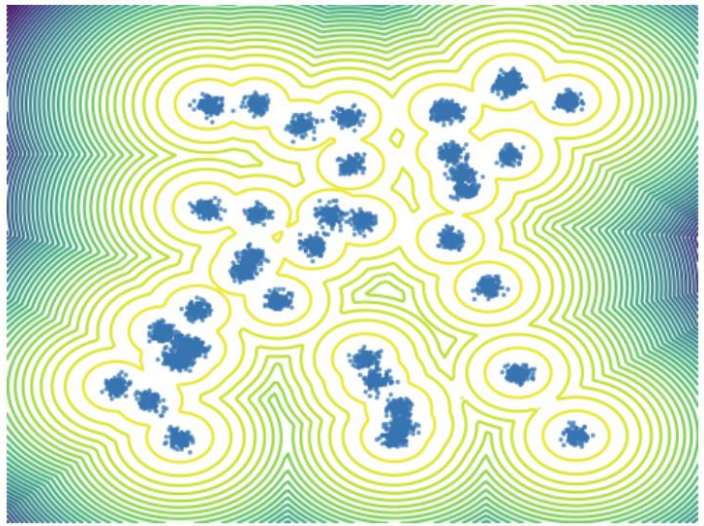
Bradley Efron. Tweedie's Formula and Selection Bias. Journal of the American Statistical Association, 2011
Zhang et al. Moment Matching Denoising Gibbs Sampling. NeurIPS, 2023.

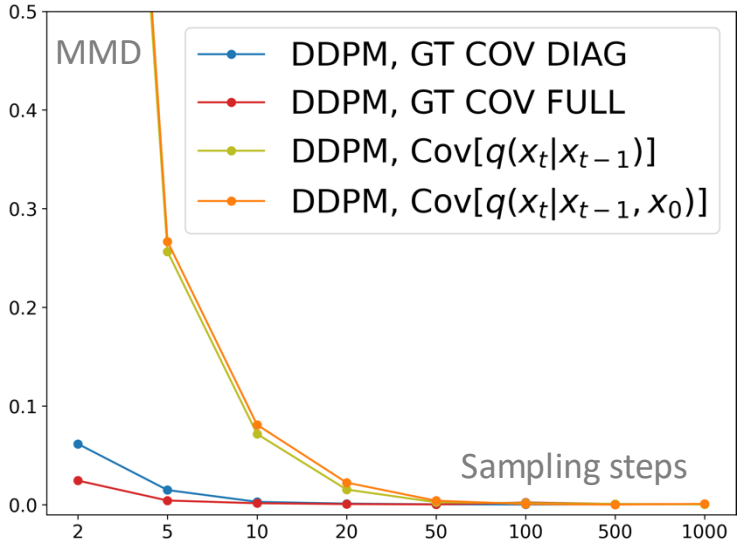
Optimal Covariance of the Denoising Distribution

Optimal Covariance: Given $q(x, x_t) = q(x)q(x_t|x)$ with $q(x_t|x) = \mathcal{N}(\alpha_t x, \sigma_t^2 I)$, the covariance of the posterior $q(x|x_t) \propto q(x)q(x_t|x)$ has a closed form

$$\Sigma(x_t) = (\sigma_t^4 \nabla_{x_t}^2 \log q(x_t) + \sigma_t^2 I)/\alpha_t^2$$

This is also known as the **2nd-Order Tweedie's Lemma**





Covariance structure of the denoising distribution matters!

Optimal Covariance of the Denoising Distribution

Optimal Covariance: Given $q(x, x_t) = q(x)q(x_t|x)$ with $q(x_t|x) = \mathcal{N}(\alpha_t x, \sigma_t^2 I)$, the covariance of the posterior $q(x|x_t) \propto q(x)q(x_t|x)$ has a closed form

$$\Sigma(x_t) = (\sigma_t^4 \nabla_{x_t}^2 \log q(x_t) + \sigma_t^2 I)/\alpha_t^2$$

This is also known as the **2nd-Order Tweedie's Lemma**

Two challenges in applying optimal covariance for D-dimensional data

1. Memory inefficiency: store Hessian requires $O(D^2)$
2. Computation inefficiency: compute Hessian requires $O(D)$ network evaluations

Optimal Diagonal Covariance Estimation

- Estimate the diagonal of the optimal covariance

$$\mathrm{diag}(\Sigma(x_t)) = (\sigma_t^4 \mathrm{diag}(\nabla_{x_t}^2 \log q(x_t)) + \sigma_t^2 I)/\alpha_t^2$$

Optimal Diagonal Covariance Estimation

- Estimate the diagonal of the optimal covariance

$$\mathrm{diag}(\Sigma(x_t)) = (\sigma_t^4 \mathrm{diag}(\nabla_{x_t}^2 \log q(x_t)) + \sigma_t^2 I)/\alpha_t^2$$

- Hutchinson diagonal Hessian estimation

$$\mathrm{diag}\big(H(x_t)\big) \approx \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t)v_m$$

Rademacher RVs

Jacobian Vector Product

$$H(x_t) = \nabla_{x_t}^2 \log q(x_t) \approx \nabla_{x_t} s_\theta(x_t)$$

Optimal Diagonal Covariance Estimation

- Estimate the diagonal of the optimal covariance

$$\mathrm{diag}(\Sigma(x_t)) = (\sigma_t^4 \mathrm{diag}(\nabla_{x_t}^2 \log q(x_t)) + \sigma_t^2 I)/\alpha_t^2$$

- Hutchinson diagonal Hessian estimation

$$\mathrm{diag}\big(H(x_t)\big) \approx \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t)v_m$$

Rademacher RVs

Jacobian Vector Product

$$H(x_t) = \nabla_{x_t}^2 \log q(x_t) \approx \nabla_{x_t} s_\theta(x_t)$$

(a) OCM-DDPM ($M=5$) (b) OCM-DDPM ($M=10$) (c) OCM-DDPM ($M=15$) (d) OCM-DDPM ($M=20$)

Larger Monte-Carlo sample number leads to better generation quality!

Optimal Diagonal Covariance Estimation

- Estimate the diagonal of the optimal covariance

$$\mathrm{diag}(\Sigma(x_t)) = (\sigma_t^4 \mathrm{diag}(\nabla_{x_t}^2 \log q(x_t)) + \sigma_t^2 I)/\alpha_t^2$$

- Hutchinson diagonal Hessian estimation

$$\mathrm{diag}\big(H(x_t)\big) \approx \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m$$

Rademacher RVs

Jacobian Vector Product

$$H(x_t) = \nabla_{x_t}^2 \log q(x_t) \approx \nabla_{x_t} s_\theta(x_t)$$

One challenge in applying diagonal covariance estimation for D-dimensional data

~~1. Memory inefficiency: store Hessian requires $O(D^2)$~~

2. Computation inefficiency: Hutchinson estimator requires $O(M)$ network evaluations

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

$H(x_t) \approx \nabla_{x_t} s_{\theta_1}(x_t)$

$h: \mathcal{X} \mapsto \mathbb{R}^{|x|}$

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

$h: \mathcal{X} \mapsto \mathbb{R}^{|x|}$

$H(x_t) \approx \nabla_{x_t} s_{\theta_1}(x_t)$

consistent if $M \to \infty$ 😁

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

h_{θ_2}: $h: \mathcal{X} \mapsto \mathbb{R}^{|x|}$

$H(x_t) \approx \nabla_{x_t} s_{\theta_1}(x_t)$

consistent if $M \to \infty$ 😄

biased for small M 😅

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

$h: \mathcal{X} \mapsto \mathbb{R}^{|x|}$

$H(x_t) \approx \nabla_{x_t} s_{\theta_1}(x_t)$

consistent if $M \to \infty$ 😄

biased for small M 😅

- Unbiased optimal diagonal covariance matching (OCM) 😎

$$L_{OCM}(\theta_2) = \mathbb{E}_{t,q(x_t),p(v)} \| h_{\theta_2}(x_t) - v \odot H(x_t) v \|_2^2$$

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

$h: \mathcal{X} \mapsto \mathbb{R}^{|x|}$

$H(x_t) \approx \nabla_{x_t} s_{\theta_1}(x_t)$

consistent if $M \to \infty$ 😄

biased for small M 😅

- Unbiased optimal diagonal covariance matching (OCM) 😎

$$L_{OCM}(\theta_2) = \mathbb{E}_{t,q(x_t),p(v)} \| h_{\theta_2}(x_t) - v \odot H(x_t) v \|_2^2$$

Same gradient with the naive obective with $M \to \infty$

Optimal Diagonal Covariance Matching

- Naive optimal diagonal covariance matching

$$\min_{\theta_2} \mathbb{E}_{t,q(x_t)} \| h_{\theta_2}(x_t) - \frac{1}{M}\sum_{m=1}^{M} v_m \odot H(x_t) v_m \|_2^2$$

$h: \mathcal{X} \mapsto \mathbb{R}^{|x|}$

$H(x_t) \approx \nabla_{x_t} s_{\theta_1}(x_t)$

consistent if $M \to \infty$ 😄

biased for small M 😅

- Unbiased optimal diagonal covariance matching (OCM) 😎

$$L_{OCM}(\theta_2) = \mathbb{E}_{t,q(x_t),p(v)} \| h_{\theta_2}(x_t) - v \odot H(x_t) v \|_2^2$$

Same gradient with the naive obective with $M \to \infty$

- Final state-dependent diagonal covariance estimation

$$\mathrm{diag}(\Sigma_{\theta_2}(x_t)) = (\sigma_t^4 h_{\theta_2}(x_t) + \sigma_t^2 I)/\alpha_t^2$$

Improved DDPM (I-DDPM)

- I-DDPM learns a state-dependent diagonal covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \mathrm{diag}(\Sigma_{\theta_2}(x_t))\right)$$

Nichol et al. Improved Denoising Diffusion Probabilistic Models. ICML, 2021.

Improved DDPM (I-DDPM)

- I-DDPM learns a state-dependent diagonal covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \mathrm{diag}(\Sigma_{\theta_2}(x_t))\right)$$

- Maximize ELBO to learn diagonal covariance $\mathrm{diag}(\Sigma_{\theta_2})$

$$\underset{\theta_2}{\mathrm{argmin}}\, \mathbb{E}_{t,q(x_0,x_t)} \mathrm{KL}(q(x_{t-1}|x_t, x_0) \| p_{\theta_1,\theta_2}(x_{t-1}|x_t))$$

Nichol et al. Improved Denoising Diffusion Probabilistic Models. ICML, 2021.

Improved DDPM (I-DDPM)

- I-DDPM learns a state-dependent diagonal covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \mathrm{diag}(\Sigma_{\theta_2}(x_t))\right)$$

- Maximize ELBO to learn diagonal covariance $\mathrm{diag}(\Sigma_{\theta_2})$

$$\underset{\theta_2}{\mathrm{argmin}}\, \mathbb{E}_{t,q(x_0,x_t)} \mathrm{KL}(q(x_{t-1}|x_t, x_0) \| p_{\theta_1,\theta_2}(x_{t-1}|x_t))$$

😄 A general solution applicable to all forward processes

😅 Does not leverage the optimal covariance structure of Gaussian diffusion

Nichol et al. Improved Denoising Diffusion Probabilistic Models. ICML, 2021.

Analytical DDPM (A-DDPM)

- A-DDPM assumes a state-independent isotropic covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \tau_t^2 I\right)$$

Bao et al. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. ICLR, 2022.

Analytical DDPM (A-DDPM)

- A-DDPM assumes a state-independent isotropic covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \tau_t^2 I\right)$$

- τ_t has an analytical form under the optimal ELBO

$$\tau_t = (\sigma_t^2 - \frac{\sigma_t^4}{d}\mathbb{E}_{q(x_t)}[s_\theta(x_t)])/\alpha_t$$

empirical estimation from training data

Bao et al. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. ICLR, 2022.

Analytical DDPM (A-DDPM)

- A-DDPM assumes a state-independent isotropic covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \tau_t^2 I\right)$$

- τ_t has an analytical form under the optimal ELBO

$$\tau_t = (\sigma_t^2 - \frac{\sigma_t^4}{d}\mathbb{E}_{q(x_t)}[s_\theta(x_t)])/\alpha_t$$

empirical estimation from training data

😄 Training-free estimation estimation with the pre-trained score

😅 State-independent and isotropic covariance has a limited flexibility

Bao et al. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. ICLR, 2022.

Square Noise DDPM (SN-DDPM)

- SN-DDPM learns a state-dependent diagonal covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \mathrm{diag}(\Sigma_{\theta_2}(x_t))\right)$$

Bao et al. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. ICLR, 2022.

Square Noise DDPM (SN-DDPM)

- SN-DDPM learns a state-dependent diagonal covariance

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_{\theta_1}(x_t), \mathrm{diag}(\Sigma_{\theta_2}(x_t))\right)$$

- $\mathrm{diag}(\Sigma_{\theta_2}(x_t))$ has an analytical form under the optimal ELBO

$$\mathrm{diag}(\Sigma(x_t)) = \frac{\sigma_t^2}{\alpha_t}\left(\frac{1}{\beta_t}\mathbb{E}_{q(x|x_t)}[\epsilon^2] - \nabla_{x_t}\log q(x_t)^2\right)$$

Bao et al. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. ICLR, 2022.

Square Noise DDPM (SN-DDPM)

- SN-DDPM parametrizes the state-dependent diagonal covariance as

$$\mathrm{diag}(\Sigma(x_t)) = \frac{\sigma_t^2}{\alpha_t}\left(\frac{1}{\beta_t} g_{\theta_2}(x_t) - s_{\theta_1}^2(x_t)\right)$$

Bao et al. Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models. ICML, 2022.

Square Noise DDPM (SN-DDPM)

- SN-DDPM parametrizes the state-dependent diagonal covariance as

$$\mathrm{diag}(\Sigma(x_t)) = \frac{\sigma_t^2}{\alpha_t}\left(\frac{1}{\beta_t} g_{\theta_2}(x_t) - s_{\theta_1}^2(x_t)\right)$$

- $g_{\theta_2}: \mathcal{X} \to \mathbb{R}^{|\mathcal{X}|}$ can be learned via square noise prediction

$$\min_{\theta_2} \mathbb{E}_{t, q(x_0, x_t)} \| \epsilon_t^2 - g_{\theta_2}(x_t) \|_2^2$$

$$\epsilon_t = (x_t - \alpha_t x_{t-1})/\sigma_t$$

Bao et al. Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models. ICML, 2022.

Square Noise DDPM (SN-DDPM)

- SN-DDPM parametrizes the state-dependent diagonal covariance as

$$\mathrm{diag}(\Sigma(x_t)) = \frac{\sigma_t^2}{\alpha_t}\left(\frac{1}{\beta_t} g_{\theta_2}(x_t) - s_{\theta_1}^2(x_t)\right)$$

- $g_{\theta_2}: \mathcal{X} \to \mathbb{R}^{|\mathcal{X}|}$ can be learned via square noise prediction

$$\min_{\theta_2} \mathbb{E}_{t,q(x_0,x_t)} \| \epsilon_t^2 - g_{\theta_2}(x_t) \|_2^2$$

$$\epsilon_t = (x_t - \alpha_t x_{t-1})/\sigma_t$$

😄 State-dependent covariance increases flexibility

😅 Quadratic term ϵ_t^2 amplifies the estimation error when $t \to 0$

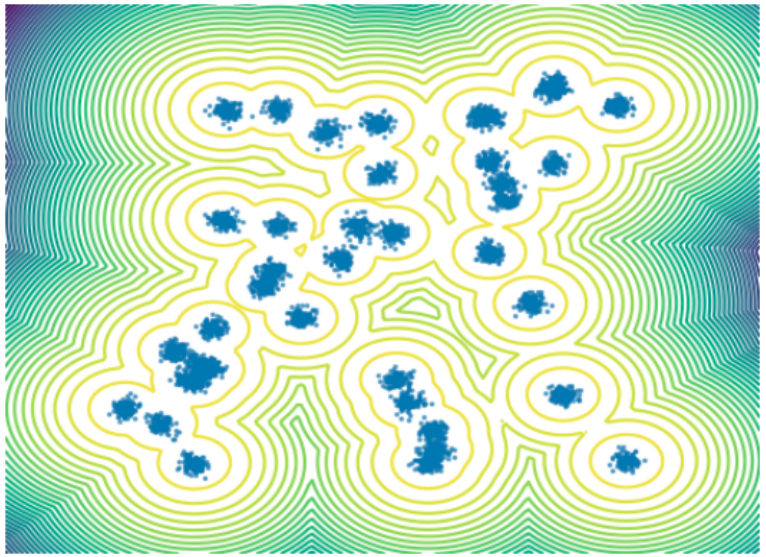
Bao et al. Estimating the Optimal Covariance with Imperfect Mean in Diffusion Probabilistic Models. ICML, 2022.

Overview of Different Covariance Choices in Diffusion Models

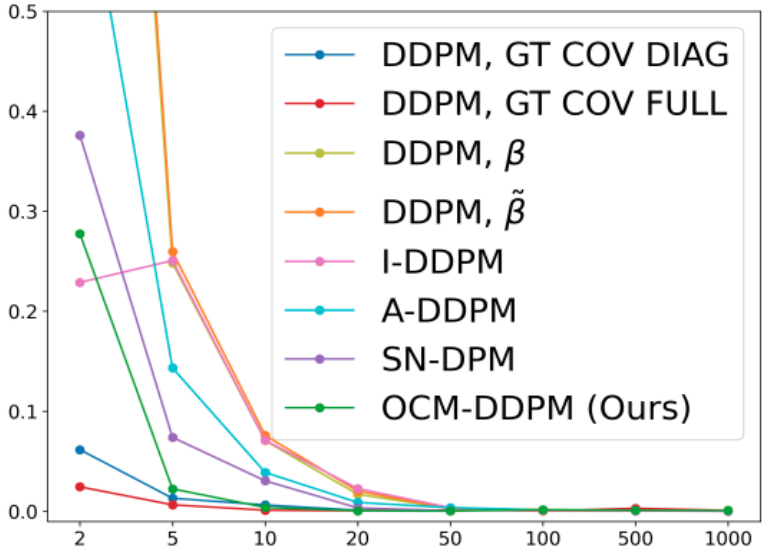
Modeling Capability ↓

Covariance Type	+#Passes	Intuition
x_t-independent Isotropic β (Ho et al., 2020)	0	Cov. of $q(x_t\|x_{t-1})$
x_t-independent Isotropic $\tilde{\beta}$ (Ho et al., 2020)	0	Cov. of $q(x_t\|x_{t-1}, x_0)$
x_t-independent Isotropic Estimation (Bao et al., 2022b)	0	Estimate from data
x_t-dependent Diagonal VLB (Nichol & Dhariwal, 2021)	1	Learn from data
x_t-dependent Diagonal NS (Bao et al., 2022a)	1	Learn from data
x_t-dependent Diagonal OCM (Ours)	1	Learn from score
x_t-dependent Diagonal Estimation (Zhang et al., 2024b)	$2M$	Estimate from score
x_t-dependent Diagonal Analytic (Zhang et al., 2024b)	D	Calculate from score
x_t-dependent Full Analytic (Zhang et al., 2024b)	D	Calculate from score

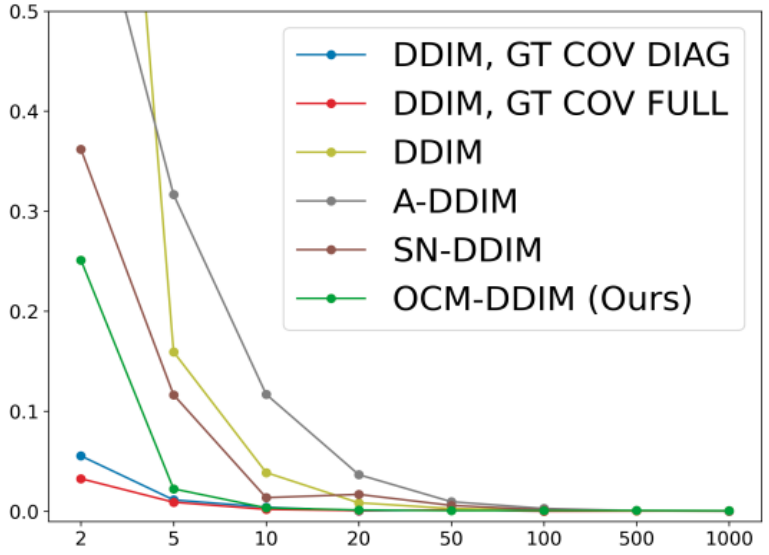
Toy Demonstration



(a) Training Data and Density



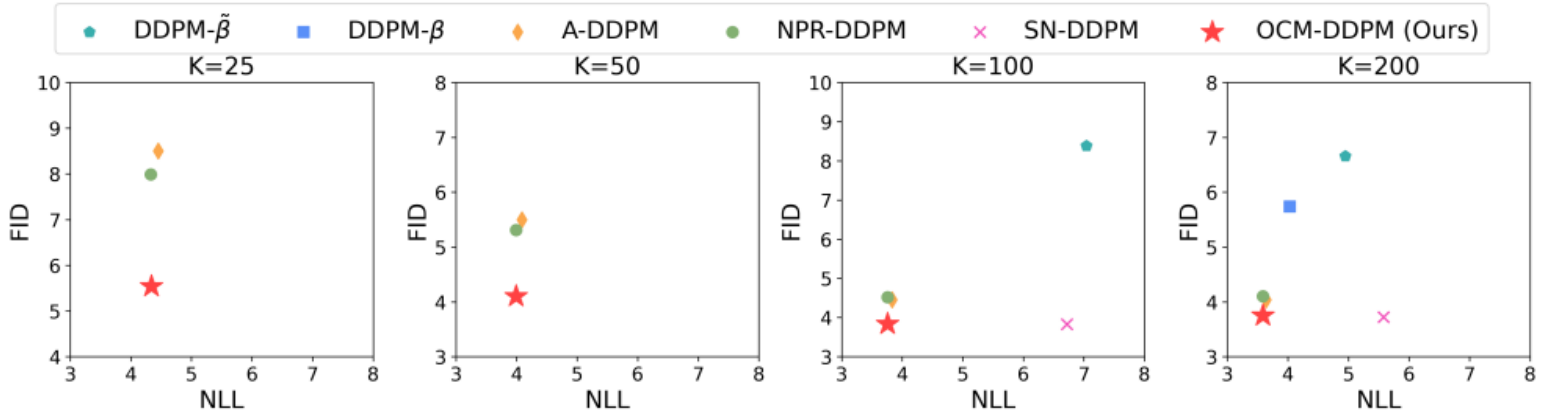
(b) DDPM MMD v.s. Steps



(c) DDIM MMD v.s. Steps

Our method outperforms baselines and achieves results closer to the optimal covariance.

Improving Image Modelling in the Pixel Space

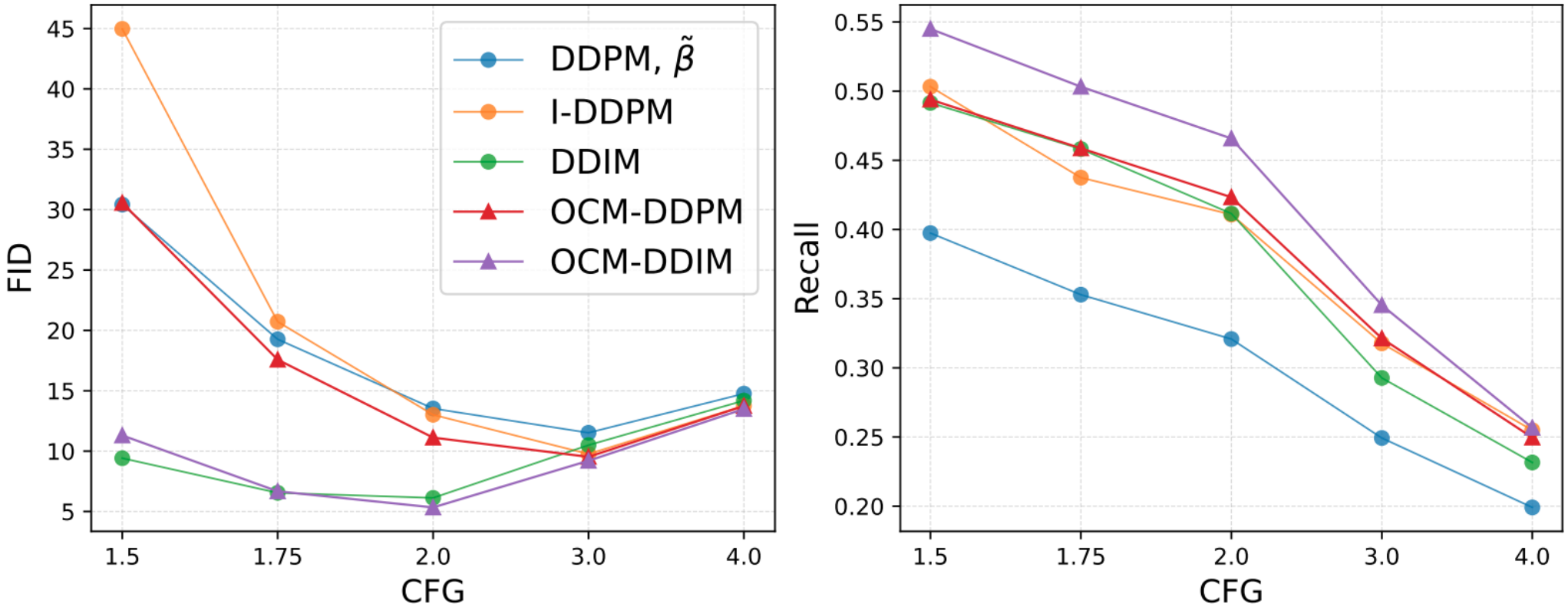


CIFAR10: Our method consistently achieves the best trade-off between FID and NLL

METHOD	CIFAR10	CELEBA 64X64	LSUN BEDROOM	IMAGENET 256X256
DDPM	90 (6.12)	> 200	130 (6.06)	21 (5.89)
DDIM	30 (5.85)	> 100	BEST FID > 6	11 (5.58)
IMPROVED DDPM	45 (5.96)	MISSING MODEL	**90** (6.02)	22 (6.08)
ANALYTIC-DPM	25 (5.81)	55 (5.98)	100 (6.05)	MISSING MODEL
NPR-DPM	1.002×23 (5.76)	1.013×50 (6.04)	$1.021\times$**90** (6.01)	MISSING MODEL
SN-DPM	1.005×17 (5.81)	1.019×22 (5.96)	1.114×92 (6.02)	MISSING MODEL
OCM-DPM (OURS)	$1.003\times$**16** (5.83)	$1.015\times$**21** (5.94)	$1.112\times$**90** (6.04)	$1.007\times$**10** (5.33)

Our method requires fewer sampling steps to achieve an FID around 6

Improving Image Modelling in the Latent Space



ImageNet256: Our method improve generation diversity while maintains high generated image quality

Thank you for watching

- **Contact:** Zijing Ou (z.ou22@imperial.ac.uk) & Mingtian Zhang (m.zhang@cs.ucl.ac.uk)
- **Code:** https://github.com/J-zin/OCM_DPM
- **Oral time:** 2025/04/24, 11:06am (Oral Session 1C)
- **Poster time:** 2025/04/24, 3:30pm-5:30pm (Poster Session 2)

IMPERIAL UCL UNIVERSITY OF CAMBRIDGE MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS EBERHARD KARLS UNIVERSITÄT TÜBINGEN