A Detailed derivations

In this section, we provide detailed derivations for the Theorem and equations shown in the main text. We follow the regularization assumptions listed in Song et al. [12].

A.1 Proof of Theorem 1

Proof. For any two timesteps $0 \leq s < t \leq T$, i.e., the transition probability from x_s to x_t is written as $q_{st}(x_t|x_s) = \mathcal{N}\left(x_t \middle| \alpha_{t|s} x_s, \sigma_{t|s}^2 \mathbf{I}\right)$, where $\alpha_{t|s} = \frac{\alpha_t}{\alpha_s}$ and $\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2$. The marginal distribution $q_t(x_t) = \int q_{st}(x_t|x_s) q_s(x_s) dx_s$ and we have

$$
\begin{aligned}
\nabla_{x_t} \log q_t(x_t) &= \frac{1}{\alpha_{t|s}} \nabla_{\alpha_{t|s}^{-1} x_t} \log \left(\frac{1}{\alpha_{t|s}^k} \mathbb{E}_{\mathcal{N}\left(x_s \middle| \alpha_{t|s}^{-1} x_t, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(x_s) \right] \right) \\
&= \frac{1}{\alpha_{t|s}} \nabla_{\alpha_{t|s}^{-1} x_t} \log \left(\mathbb{E}_{\mathcal{N}\left(\eta \middle| 0, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(\alpha_{t|s}^{-1} x_t + \eta) \right] \right) \\
&= \frac{\mathbb{E}_{\mathcal{N}\left(\eta \middle| 0, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[\nabla_{\alpha_{t|s}^{-1} x_t} q_s(\alpha_{t|s}^{-1} x_t + \eta) \right]}{\alpha_{t|s} \mathbb{E}_{\mathcal{N}\left(\eta \middle| 0, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(\alpha_{t|s}^{-1} x_t + \eta) \right]} \\
&= \frac{\mathbb{E}_{\mathcal{N}\left(\eta \middle| 0, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(\alpha_{t|s}^{-1} x_t + \eta) \nabla_{\alpha_{t|s}^{-1} x_t + \eta} \log q_s(\alpha_{t|s}^{-1} x_t + \eta) \right]}{\alpha_{t|s} \mathbb{E}_{\mathcal{N}\left(\eta \middle| 0, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(\alpha_{t|s}^{-1} x_t + \eta) \right]} \\
&= \frac{\mathbb{E}_{\mathcal{N}\left(x_s \middle| \alpha_{t|s}^{-1} x_t, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(x_s) \nabla_{x_s} \log q_s(x_s) \right]}{\alpha_{t|s} \mathbb{E}_{\mathcal{N}\left(x_s \middle| \alpha_{t|s}^{-1} x_t, \alpha_{t|s}^{-2} \sigma_{t|s}^2 \mathbf{I}\right)} \left[q_s(x_s) \right]} \\
&= \frac{\int \mathcal{N}\left(x_t \middle| \alpha_{t|s} x_s, \sigma_{t|s}^2 \mathbf{I}\right) q_s(x_s) \nabla_{x_s} \log q_s(x_s) dx_s}{\alpha_{t|s} \int \mathcal{N}\left(x_t \middle| \alpha_{t|s} x_s, \sigma_{t|s}^2 \mathbf{I}\right) q_s(x_s) dx_s} \\
&= \frac{1}{\alpha_{t|s}} \mathbb{E}_{q_{st}(x_s|x_t)} \left[\nabla_{x_s} \log q_s(x_s) \right].
\end{aligned}
\tag{1}
$$

Note that when the transition probability $q_{st}(x_t|x_s)$ corresponds to a well-defined forward process, there is $\alpha_t > 0$ for $\forall t \in [0, T]$, and thus we achieve $\alpha_t \nabla_{x_t} \log q_t(x_t) = \mathbb{E}_{q_{st}(x_s|x_t)} \left[\alpha_s \nabla_{x_s} \log q_s(x_s) \right]$. □

A.2 Proof of $\mathbb{E}_{q_0(x_0)} \left[\nabla_{x_0} \log q_0(x_0) \right] = 0$

Proof. The input variable $x \in \mathbb{R}^k$ and $q_0(x_0) \in \mathcal{C}^2$, where $\mathcal{C}^2$ denotes the family of functions with continuous second-order derivatives.[1] We use x^i denote the i-th element of x, then we can derive the expectation

$$
\begin{aligned}
\mathbb{E}_{q_0(x_0)} \left[\frac{\partial}{\partial x_0^i} \log q_0(x_0) \right] &= \int \cdots \int q_0(x_0) \frac{\partial}{\partial x_0^i} \log q_0(x_0) dx_0^1 dx_0^2 \cdots dx_0^k \\
&= \int \cdots \int \frac{\partial}{\partial x_0^i} q_0(x_0) dx_0^1 dx_0^2 \cdots dx_0^k \\
&= \int \frac{\partial}{\partial x_0^i} \left(\int q_0(x_0^i, x_0^{\setminus i}) dx_0^{\setminus i} \right) dx_0^i \\
&= \int \frac{d}{dx_0^i} q_0(x_0^i) dx_0^i = 0,
\end{aligned}
\tag{2}
$$

[1] This continuously differentiable assumption can be satisfied by adding a small Gaussian noise (e.g., with variance of 0.0001) on the original data distribution, as done in Song and Ermon [11].

where $x_0^{\setminus i}$ denotes all the $k-1$ elements in x_0 except for the i-th one. The last equation holds under the boundary condition that $\lim_{x_0^i \to \infty} q_0(x_0^i) = 0$ hold for any $i \in [K]$. Thus, we achieve the conclusion that $\mathbb{E}_{q_0(x_0)}\left[\nabla_{x_0} \log q_0(x_0)\right] = 0$. □

A.3 Concentration bounds

We describe concentration bounds [2, 1] of the martingale $\alpha_t \nabla_{x_t} \log q_t(x_t)$.

Azuma's inequality. For discrete reverse timestep $t = T, T-1, \cdots, 0$, Assuming that there exist constants $0 < c_1, c_2, \cdots, < \infty$ such that for the i-th element of x,

$$A_t \leq \frac{\partial}{\partial x_{t-1}^i} \alpha_{t-1} \log q_{t-1}(x_{t-1}) - \frac{\partial}{\partial x_t^i} \alpha_t \log q_t(x_t) \leq B_t \text{ and } B_t - A_t \leq c_t \tag{3}$$

almost surely. Then $\forall \epsilon > 0$, the probability (note that $\alpha_0 = 1$)

$$P\left(\left|\frac{\partial}{\partial x_0^i} \log q_0(x_0) - \frac{\partial}{\partial x_T^i} \alpha_T \log q_T(x_T)\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^T c_t^2}\right). \tag{4}$$

Specially, considering that $q_T(x_T) \approx \mathcal{N}(x_T | 0, \widetilde{\sigma}^2 \mathbf{I})$, there is $\frac{\partial}{\partial x_T^i} \log q_T(x_T) \approx -\frac{x_T^i}{\widetilde{\sigma}^2}$. Thus, we can approximately obtain

$$P\left(\left|\frac{\partial}{\partial x_0^i} \log q_0(x_0) + \frac{\alpha_T x_T^i}{\widetilde{\sigma}^2}\right| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^T c_t^2}\right). \tag{5}$$

Doob's inequality. For continuous reverse timestep t from T to 0, if the sample paths of the martingale are almost surely right-continuous, then for the i-th element of x we have (note that $\alpha_0 = 1$)

$$P\left(\sup_{0 \leq t \leq T} \frac{\partial}{\partial x_t^i} \alpha_t \log q_t(x_t) \geq C\right) \leq \frac{\mathbb{E}_{q_0(x_0)}\left[\max\left(\frac{\partial}{\partial x_0^i} \log q_0(x_0), 0\right)\right]}{C}. \tag{6}$$

A.4 High-order SM objectives

Lu et al. [6] show that the KL divergence $\mathcal{D}_{\text{KL}}\left(q_0 \| p_0^{\text{ODE}}(\theta)\right)$ can be bounded as

$$\mathcal{D}_{\text{KL}}\left(q_0 \| p_0^{\text{ODE}}(\theta)\right) \leq \mathcal{D}_{\text{KL}}\left(q_T \| p_T\right) + \sqrt{\mathcal{J}_{\text{SM}}(\theta; g(t)^2)} \cdot \sqrt{\mathcal{J}_{\text{Fisher}}(\theta)}, \tag{7}$$

where $\mathcal{J}_{\text{Fisher}}(\theta)$ is a weighted sum of Fisher divergence between $q_t(x_t)$ and $p_t^{\text{ODE}}(\theta)$ as

$$\mathcal{J}_{\text{Fisher}}(\theta) = \frac{1}{2} \int_0^T g(t)^2 D_F\left(q_t \| p_t^{\text{ODE}}(\theta)\right) dt. \tag{8}$$

Moreover, Lu et al. [6] prove that if $\forall t \in [0, T]$ and $\forall x_t \in \mathbb{R}^k$, there exist a constant C_F such that the spectral norm of Hessian matrix $\|\nabla_{x_t}^2 \log p_t^{\text{ODE}}(x_t; \theta)\|_2 \leq C_F$, and there exist $\delta_1, \delta_2, \delta_3 > 0$ such that

$$\begin{aligned}
&\|\boldsymbol{s}_\theta^t(x_t) - \nabla_{x_t} \log q_t(x_t)\|_2 \leq \delta_1, \\
&\|\nabla_{x_t} \boldsymbol{s}_\theta^t(x_t) - \nabla_{x_t}^2 \log q_t(x_t)\|_F \leq \delta_2, \\
&\|\nabla_{x_t} \mathbf{tr}\left(\nabla_{x_t} \boldsymbol{s}_\theta^t(x_t)\right) - \nabla_{x_t} \mathbf{tr}\left(\nabla_{x_t}^2 \log q_t(x_t)\right)\|_2 \leq \delta_3,
\end{aligned} \tag{9}$$

where $\|\cdot\|_F$ is the Frobenius norm of matrix. Then there exist a function $U(t; \delta_1, \delta_2, \delta_3, q)$ that independent of θ and strictly increasing (if $g(t) \neq 0$) w.r.t. δ_1, δ_2, and δ_3, respectively, such that the Fisher divergence can be bounded as $D_F\left(q_t \| p_t^{\text{ODE}}(\theta)\right) \leq U(t; \delta_1, \delta_2, \delta_3, q)$.

The case after calibration. When we impose the calibration term $\eta_t^* = \mathbb{E}_{q_t(x_t)}\left[\boldsymbol{s}_\theta^t(x_t)\right]$ to get the score model $\boldsymbol{s}_\theta^t(x_t) - \eta_t^*$, there is $\nabla_{x_t} \eta_t^* = 0$ and thus $\nabla_{x_t}\left(\boldsymbol{s}_\theta^t(x_t) - \eta_t^*\right) = \nabla_{x_t} \boldsymbol{s}_\theta^t(x_t)$. Then we have

$$\begin{aligned}
&\|\boldsymbol{s}_\theta^t(x_t) - \eta_t^* - \nabla_{x_t} \log q_t(x_t)\|_2 \leq \delta_1' \leq \delta_1, \\
&\|\nabla_{x_t}\left(\boldsymbol{s}_\theta^t(x_t) - \eta_t^*\right) - \nabla_{x_t}^2 \log q_t(x_t)\|_F \leq \delta_2, \\
&\|\nabla_{x_t} \mathbf{tr}\left(\nabla_{x_t}\left(\boldsymbol{s}_\theta^t(x_t) - \eta_t^*\right)\right) - \nabla_{x_t} \mathbf{tr}\left(\nabla_{x_t}^2 \log q_t(x_t)\right)\|_2 \leq \delta_3.
\end{aligned} \tag{10}$$

From these, we know that the Fisher divergence $D_F\left(q_t \| p_t^{\text{ODE}}(\theta, \eta_t^*)\right) \leq U(t; \delta_1', \delta_2, \delta_3, q) \leq U(t; \delta_1, \delta_2, \delta_3, q)$, namely, $D_F\left(q_t \| p_t^{\text{ODE}}(\theta, \eta_t^*)\right)$ has a lower upper bound compared to $D_F\left(q_t \| p_t^{\text{ODE}}(\theta)\right)$. Consequently, we can get lower upper bounds for both $\mathcal{J}_{\text{Fisher}}(\theta, \eta_t^*)$ and $\mathcal{D}_{\text{KL}}\left(q_0 \| p_0^{\text{ODE}}(\theta, \eta_t^*)\right)$, compared to $\mathcal{J}_{\text{Fisher}}(\theta)$ and $\mathcal{D}_{\text{KL}}\left(q_0 \| p_0^{\text{ODE}}(\theta)\right)$, respectively.

B Model parametrization

This section introduces different parametrizations used in diffusion models and provides their calibrated instantiations.

B.1 Preliminary

Along the research routine of diffusion models, different model parametrizations have been used, including score prediction $\boldsymbol{s}_\theta^t(x_t)$ [11, 13], noise prediction $\boldsymbol{\epsilon}_\theta^t(x_t)$ [3, 8], data prediction $\boldsymbol{x}_\theta^t(x_t)$ [5, 7], and velocity prediction $\boldsymbol{v}_\theta^t(x_t)$ [9, 4]. Taking the DSM objective as the training loss, its instantiation at timestep $t \in [0, T]$ is written as

$$\mathcal{J}_{\text{DSM}}^t(\theta) = \begin{cases} \frac{1}{2}\mathbb{E}_{q_0(x_0), q(\epsilon)}\left[\|\boldsymbol{s}_\theta^t(x_t) + \frac{\epsilon}{\sigma_t}\|_2^2\right], & \text{score prediction;} \\ \frac{\alpha_t^2}{2\sigma_t^4}\mathbb{E}_{q_0(x_0), q(\epsilon)}\left[\|\boldsymbol{x}_\theta^t(x_t) - x_0\|_2^2\right], & \text{data prediction;} \\ \frac{1}{2\sigma_t^2}\mathbb{E}_{q_0(x_0), q(\epsilon)}\left[\|\boldsymbol{\epsilon}_\theta^t(x_t) - \epsilon\|_2^2\right], & \text{noise prediction;} \\ \frac{\alpha_t^2}{2\sigma_t^2}\mathbb{E}_{q_0(x_0), q(\epsilon)}\left[\|\boldsymbol{v}_\theta^t(x_t) - (\alpha_t\epsilon - \sigma_t x_0)\|_2^2\right], & \text{velocity prediction.} \end{cases} \tag{11}$$

B.2 Calibrated instantiation

Under different model parametrizations, we can derive the optimal calibration terms η_t^* that minimizing $\mathcal{J}_{\text{DSM}}^t(\theta, \eta_t)$ as

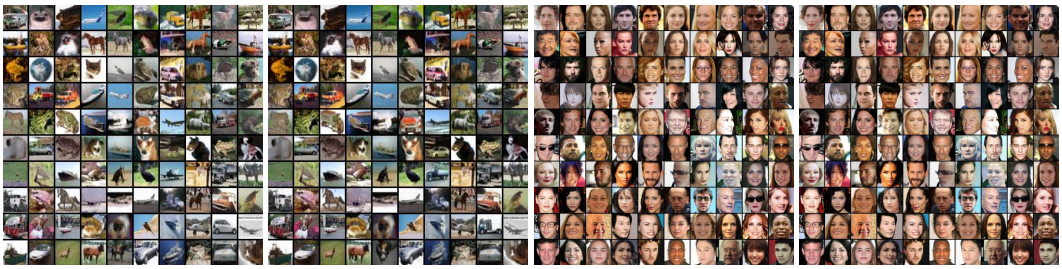
$$\eta_t^* = \begin{cases} \mathbb{E}_{q_t(x_t)}\left[\boldsymbol{s}_\theta^t(x_t)\right], & \text{score prediction;} \\ \mathbb{E}_{q_t(x_t)}\left[\boldsymbol{x}_\theta^t(x_t)\right] - \mathbb{E}_{q_0(x_0)}\left[x_0\right], & \text{data prediction;} \\ \mathbb{E}_{q_t(x_t)}\left[\boldsymbol{\epsilon}_\theta^t(x_t)\right], & \text{noise prediction;} \\ \mathbb{E}_{q_t(x_t)}\left[\boldsymbol{v}_\theta^t(x_t)\right] + \sigma_t\mathbb{E}_{q_0(x_0)}\left[x_0\right], & \text{velocity prediction.} \end{cases} \tag{12}$$

Taking η_t^* into $\mathcal{J}_{\text{DSM}}^t(\theta, \eta_t)$ we can obtain the gap

$$\mathcal{J}_{\text{DSM}}^t(\theta) - \mathcal{J}_{\text{DSM}}^t(\theta, \eta_t^*) = \begin{cases} \frac{1}{2}\|\mathbb{E}_{q_t(x_t)}\left[\boldsymbol{s}_\theta^t(x_t)\right]\|_2^2, & \text{score prediction;} \\ \frac{\alpha_t^2}{2\sigma_t^4}\|\mathbb{E}_{q_t(x_t)}\left[\boldsymbol{x}_\theta^t(x_t)\right] - \mathbb{E}_{q_0(x_0)}\left[x_0\right]\|_2^2, & \text{data prediction;} \\ \frac{1}{2\sigma_t^2}\|\mathbb{E}_{q_t(x_t)}\left[\boldsymbol{\epsilon}_\theta^t(x_t)\right]\|_2^2, & \text{noise prediction;} \\ \frac{\alpha_t^2}{2\sigma_t^2}\|\mathbb{E}_{q_t(x_t)}\left[\boldsymbol{v}_\theta^t(x_t)\right] + \sigma_t\mathbb{E}_{q_0(x_0)}\left[x_0\right]\|_2^2, & \text{velocity prediction.} \end{cases} \tag{13}$$

C Visualization of the generations

We further show generated images in Figure 1 to double confirm the efficacy of our calibration method. Our calibration could help to reduce ambiguous generations on both CIFAR-10 and CelebA.



(a) CIFAR-10, w/ calibration (b) CIFAR-10, w/o calibration (c) CelebA, w/ calibration (d) CelebA, w/o calibration

Figure 1: Unconditional generation results on CIFAR-10 and CelebA using models from [3] and [10] respectively. The number of sampling steps is 20 based on the results in Tables 1 and 2.

References

[1] Kazuoki Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

[2] Joseph L Doob and Joseph L Doob. *Stochastic processes*, volume 7. Wiley New York, 1953.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[4] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[5] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[6] Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International Conference on Machine Learning (ICML)*, 2022.

[7] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

[11] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 11895–11907, 2019.

[12] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[13] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.