CONTRACTIVE DIFFUSION PROBABILISTIC MODELS

WENPIN TANG AND HANYANG ZHAO

Abstract. Diffusion probabilistic models (DPMs) have emerged as a promising technique in generative modeling. The success of DPMs relies on two ingredients: time reversal of diffusion processes and score matching. Most existing works implicitly assume that score matching is close to perfect, while this assumption is questionable. In view of possibly unguaranteed score matching, we propose a new criterion – the contraction of backward sampling in the design of DPMs, leading to a novel class of contractive DPMs (CDPMs). The key insight is that the contraction in the backward process can narrow score matching errors and discretization errors. Thus, our proposed CDPMs are robust to both sources of error. For practical use, we show that CDPM can leverage pretrained DPMs by a simple transformation, and does not need retraining. We corroborated our approach by experiments on synthetic 1-dim examples, Swiss Roll, MNIST, CIFAR-10 32×32 and AFHQ 64×64 dataset. Notably, CDPM shows the best performance among all known SDE-based DPMs.

Key words: Contraction, diffusion probabilistic models, discretization, generative models, image synthesis, sampling, score matching, stochastic differential equations.

1. Introduction

Over the past decade, generative models have achieved remarkable success in creating instances across a wide variety of data modalities, including images [7, 24, 50], audio [6, 47], video [28], and text [8, 65, 69]. Diffusion probabilistic models (DPMs) have emerged as a promising generative approach that is observed to outperform generative adversarial nets on image and audio synthesis [17, 35], and underpins the major accomplishment in text-to-image creators such as DALL·E 2 [49] and Stable Diffusion [52], and the text-to-video generator Sora [45]. The concept of DPMs finds its roots in energy-based models [53], and is popularized by [26, 57, 60] in an attempt to produce from noise new samples (e.g. images, audio, text) that resemble the target data, while maintain diversity. See [11, 61, 68] for a review on DPMs.

DPMs relies on forward-backward Markov processes. The forward process starts with the target data distribution, and runs for some time until the signal is destroyed – this gives rise to *noise*. The backward process is then initiated with the noise, and reverses the forward process in time to generate samples whose distribution is close to the target distribution. In [2, 26, 57], DPMs are discrete time-indexed Markov chains; [13, 56, 60] model DPMs in continuous time as stochastic differential equations (SDEs). Nevertheless, there is no conceptual distinction between discrete and continuous DPMs as continuous DPMs can be viewed as the continuum limits of discrete DPMs, and discrete DPMs are time discretization of continuous DPMs. In this paper, we adopt a *continuous time* perspective, and algorithms are derived by discretization. Being concrete,

- The forward process $(X_t,\, 0 \le t \le T)$ is governed by the SDE:
$$dX_t = b(t, X_t)dt + \sigma(t)dB_t, \quad \text{with } X_0 \sim p_{\text{data}}(\cdot),$$
where B_t is Brownian motion, $b(\cdot,\cdot)$ and $\sigma(\cdot)$ are model parameters to be designed, and $p_{\text{data}}(\cdot)$ is the target data distribution.

Date: June 21, 2024.

Emails: {wt2319,hz2684}@columbia.edu, Department of Industrial Engineering and Operations Research, Columbia University.

- The backward process $(\overline{X}_t,\, 0 \le t \le T)$ is governed by the SDE:

$$d\overline{X}_t = \overline{b}(t, \overline{X}_t)dt + \overline{\sigma}(t)d\overline{B}_t, \quad \text{with } X_0 \sim p_{\text{noise}}(\cdot),$$

  where $\overline{B}_t$ is Brownian motion, $\overline{b}(\cdot,\cdot)$ and $\overline{\sigma}(\cdot)$ are *some* functions, and $p_{\text{noise}}(\cdot)$ is the noise distribution that does not depend on $p_{\text{data}}(\cdot)$.

The forward process transforms data to noise (with a suitable choice of $b(\cdot,\cdot)$ and $\sigma(\cdot)$). What is miraculous is how the backward process recovers the target data distribution $\overline{X}_T \approx p_{\text{data}}(\cdot)$ from noise $p_{\text{noise}}(\cdot)$. The secret consists of two key ingredients.

(1) *Time reversal of diffusion processes*: The backward process has an explicit form, where $\overline{b}(t,x)$ depends on $b(T-t,x)$, $\sigma(T-t)$ and Stein's score functions (the gradients of the log marginal density of the forward process), and $\overline{\sigma}(t) = \sigma(T-t)$ [9, 25].

(2) *Score matching*: The backward process is easily sampled given $b(\cdot,\cdot)$, $\sigma(\cdot)$ and score functions. The trick is to learn score functions via forward sampling, referred to as score-based DPMs. Leaning score functions, also known as score matching, features in a body of active research [27, 34, 59, 66].

In a nutshell, DPMs combine forward score learning with backward sampling.

As is clear, a DPM is specified by the pair $(b(\cdot,\,\cdot), \sigma(\cdot))$. Popular examples include Ornstein-Uhlenbeck (OU) processes [16], variance exploding (VE) SDEs, variance preserving (VP) SDEs, and sub-variance preserving (subVP) SDEs [60] (see Appendix B for definitions). Especially, VE and VP SDEs are the continuum limits of score matching with Langevin dynamics (SMLD) [57] and denoising diffusion probabilistic models (DDPMs) [26] respectively. A natural question is:

How do we design the pair $(b(\cdot,\,\cdot), \sigma(\cdot))$ for a DPM?

The first rule of thumb, which is satisfied by all the aforementioned examples, is:

Rule 1. *The conditional distributions of $X_t \,|\, X_0$ are easy to sample (e.g. Gaussian).*

This rule allows efficient sampling in the forward process for score matching via stochastic optimization. Notably, Rule 1 provides only instructions on the forward learning step, but no requirement on backward sampling. The purpose of this paper is to put forward a principle regarding backward sampling in the design of DPMs. Our proposal is:

Rule 2. *The backward process $\overline{X}$ is contractive.*

Roughly speaking, being contractive indicates that the process tends to be confined, or converge (see Section 3 for explanations). DPMs that comply with Rule 1 and 2 are called *contractive DPMs* (CDPMs). Here we abuse the term contractive DPMs by meaning that their backward processes, rather than the forward processes, exhibit contractive properties. At a high level, the contraction of the backward process will prevent score matching errors, which may be wild, from expanding over the time. The contributions of this work are summarized as follows.

Methodology: We propose a new criterion (**Rule 2**) for designing DPMs. This naturally leads to a novel class of DPMs, including contractive OU processes and contractive subVP SDEs. The idea of requiring the backward process to be contractive stems from sampling theory of SDEs, so our methodology is theory-oriented. To our best knowledge, this is the first paper to integrate contraction into the design of DPMs, with both theoretical guarantees and good empirical performance.

Theory: We prove Wasserstein bounds between contractive DPM samplers and the target data distribution. While most previous work (e.g. [13, 16, 37]) focused on Kullback–Leibler (KL) or total variation bounds for OU processes, we consider the Wasserstein metric because it has shown to align with human judgment on image similarity [5], and the standard evaluation metric – Fréchet inception distance (FID) is based on Wasserstein distance. Early work [16, 36] gave Wasserstein bounds for the existing DPMs (OU processes, VE and VP SDEs) with exponential dependence on T. This was improved in recent studies [13, 38, 23] under various assumptions of $p_{\text{data}}(\cdot)$, where the bounds are typically of form:

$$\underbrace{(\text{noise inaccuracy}) \cdot e^{-T}}_{\text{initialization error}} + \underbrace{(\text{score mismatch}) \cdot \texttt{Poly}(T)}_{\text{score error}} + \underbrace{\texttt{Poly}(\text{step size}) \cdot \texttt{Poly}(T)}_{\text{discretization error}},$$

with $\texttt{Poly}(\cdot)$ referring to polynomial in the variable. Our result gives a Wasserstein bound for CDPMs:

$$\underbrace{(\text{noise inaccuracy}) \cdot e^{-T}}_{\text{initialization error}} + \underbrace{(\text{score mismatch}) \cdot (1 - e^{-T})}_{\text{score error}} + \underbrace{\texttt{Poly}(\text{step size})}_{\text{discretization error}}.$$

Score matching is often trained using blackbox function approximations, and the errors incurred in this step may be large. So CDPMs are designed to be robust to score mismatch and discretization, at the cost of possible initialization bias.

Experiments: We apply the proposed CDPMs to both synthetic and real data. In dimension one, we compare contractive OU with OU by adding a fixed noise to the true score function, which yields the same score matching error. Our result shows that contractive OU consistently beats OU, and is robust to different error levels and time discretization. We further compare the performance of different models via Wasserstein-2 distance of the SWISS Roll dataset and FIDs of MNIST, which show that CDPMs outperform other SDE models. On the task of CIFAR-10 unconditional generation, we obtain an FID score of 2.47 and an inception score of 10.18 for CDPM, which requires no retraining by transforming the pretrained weights of VE-SDE in [60], surpassing all other SDE models.

Literature review. In the context of generative modeling, DPMs were initiated by [57] (SMLD) and [26] (DDPM) using forward-backward Markov chains. The work [60, 56] unified the previous models via a score-based SDE framework, which also led to deterministic ordinary differential equation (ODE) samplers. Since then the field has exploded, and lots of work has been built upon DPMs and their variants. Examples include DPMs in constrained domains [20, 41, 51, 18], DPMs on manifolds [46, 15, 10], DPMs in graphic models [42], variational DPMs [32, 63] and consistency models [55], just to name a few. Early theory [16, 15, 36] established the convergence of DPMs with exponential dependence on time horizon T and dimension d. Recently, polynomial convergence of various DPMs has been proved for stochastic samplers [13, 3, 40, 23, 37, 38] and deterministic samplers [4, 12, 14, 40], under suitable conditions on the target data distribution.

The remainder of the paper is organized as follows. In Section 2, we provide background on DPMs and score matching techniques. Theoretical results for CDPMs are presented in Section 3, and connections to VE are discussed in Section 4. Experiments are reported in Section 5. We conclude with Section 6.

2. Background

2.1. **Diffusion models.** Let's explain DPMs in the context of SDEs. We follow closely the presentation in [61]. Consider the forward process:

$$dX_t = b(t, X_t)dt + \sigma(t)dB_t, \quad \text{with } X_0 \sim p_{\text{data}}(\cdot), \tag{2.1}$$

where $b : \mathbb{R}_+ \times \mathbb{R}^d \to \mathbb{R}^d$ and $\sigma : \mathbb{R}_+ \to \mathbb{R}_+$ are model parameters. Some conditions on $b(\cdot, \cdot)$ and $\sigma(\cdot)$ are required so that the SDE (2.1) is at least well-defined (see [29, Chapter 5], [62, Section 3.1]). Let $p(t, \cdot)$ be the probability density of X_t.

Set $T > 0$ to be fixed, and run the SDE (2.1) until time T to get $X_T \sim p(T, \cdot)$. Now if we start with $p(T, \cdot)$ and run the process X backward, then we can generate a copy of $p(0, \cdot) = p_{\text{data}}(\cdot)$. Being precise, consider the time reversal $\overline{X}_t := X_{T-t}$ for $0 \le t \le T$. Assuming that $\overline{X}$ also satisfies an SDE, we can run the backward procedure by

$$d\overline{X}_t = \bar{b}(t, \overline{X}_t)dt + \bar{\sigma}(t)d\overline{B}_t, \quad \text{with } \overline{X}_0 \sim p(T, \cdot),$$

which samples the desired $\overline{X}_T \sim p_{\text{data}}(\cdot)$ at time T. Note that the distribution $p(T,\cdot)$ depends on the target distribution $p_{\text{data}}(\cdot)$. The idea of DPMs is, however, to generate samples from noise. So we need to replace $p(T,\cdot)$ by a proxy $p_{\text{noise}}(\cdot)$ that is independent of $p_{\text{data}}(\cdot)$. This yields the backward process:

$$d\overline{X}_t = \bar{b}(t, \overline{X}_t)dt + \bar{\sigma}(t)d\overline{B}_t, \quad \text{with } \overline{X}_0 \sim p_{\text{noise}}(\cdot). \tag{2.2}$$

Two questions arise:

(1) How can we choose $p_{\text{noise}}(\cdot)$?
(2) What are the parameters $\bar{b}(\cdot,\cdot)$ and $\bar{\sigma}(\cdot)$?

For (1), the noise $p_{\text{noise}}(\cdot)$ is often derived by decoupling the conditional distribution of $X_t \mid X_0$ from X_0. It is expected that the closer the distributions $p(T,\cdot)$ and $p_{\text{noise}}(\cdot)$ are, the closer the distribution of $\overline{X}_T$ sampled from (2.2) is to $p_{\text{data}}(\cdot)$. For (2), it relies on the time reversal of SDEs [1, 25].

Theorem 1. *Under suitable conditions on $b(\cdot,\cdot)$, $\sigma(\cdot)$ and $\{p(t,\cdot)\}_{0\le t\le T}$, we have*

$$\bar{\sigma}(t) = \sigma(T-t), \quad \bar{b}(t,x) = -b(T-t,x) + \sigma^2(T-t)\nabla \log p(T-t,x), \tag{2.3}$$

where the term $\nabla \log p(\cdot,\cdot)$ is called Stein's score function.

We give a derivation of Theorem 1 with further references in Appendix A. As a consequence, the backward process is:

$$d\overline{X}_t = \left(-b(T-t,\overline{X}_t) + \sigma^2(T-t)\nabla \log p(T-t,\overline{X}_t)\right)dt + \sigma(T-t)d\overline{B}_t. \tag{2.4}$$

Various examples of DPMs are provided in Appendix B. Since $b(\cdot,\cdot)$ and $\sigma(\cdot)$ are chosen in advance, all but the term $\nabla \log p(T-t,\overline{X}_t)$ in (2.4) are available.

2.2. **Score matching.** As previously mentioned, we need to compute the score function $\nabla \log p(t,x)$ for backward sampling. The idea from recently developed score-based generative modeling [26, 57, 60] is to estimate $\nabla \log p(t,x)$ by function approximations. More precisely, denote by $\{s_\theta(t,x)\}_\theta$ a family of functions on $\mathbb{R}_+ \times \mathbb{R}^d$ parametrized by θ. Fixing t, the goal is to solve the problem:

$$\min_\theta \mathcal{J}_{\text{ESM}}(\theta) := \mathbb{E}_{p(t,\cdot)}|s_\theta(t,X) - \nabla \log p(t,X)|^2, \tag{2.5}$$

which is known as the *explicit score matching* (ESM) objective. The stochastic optimization (2.5) is far-fetched since the scores $\nabla \log p(t,x)$ are not available. Interestingly, this problem has been studied in the context of estimating statistical models with unknown normalizing constant. The following result [27] shows that the score matching problem (2.5) can be recast into a feasible stochastic optimization with no $\nabla \log p(t,X)$-term, known as *implicit score matching* (ISM) objective.

Theorem 2. *Let $\mathcal{J}_{ISM}(\theta) := \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2 + 2\,\nabla\cdot s_\theta(t,X)\right]$. Under suitable conditions on s_θ, we have $\mathcal{J}_{ISM}(\theta) = \mathcal{J}_{ESM}(\theta) + C$ for some C independent of θ. Consequently, the minimum point of $\mathcal{J}_{ISM}$ and that of $\mathcal{J}_{ESM}$ coincide.*

We give a proof of this theorem in Appendix C.1 for completeness. In practice, the score matching problem with a continuous weighted combination is considered:

$$\min_\theta \tilde{\mathcal{J}}_{\text{ESM}}(\theta) = \mathbb{E}_{t\in\mathcal{U}(0,T)}\mathbb{E}_{p(t,\cdot)}\left[\lambda(t)|s_\theta(t,X) - \nabla \log p(t,X)|^2\right]. \tag{2.6}$$

where $\mathcal{U}(0,T)$ denotes a uniform distribution on $[0,T]$, and $\lambda : \mathbb{R} \to \mathbb{R}_+$ is a positive weighting function. We can alternate the inside part by ISM to solve the problem:

$$\min_\theta \tilde{\mathcal{J}}_{\text{ISM}}(\theta) = \mathbb{E}_{t\in\mathcal{U}(0,T)}\mathbb{E}_{p(t,\cdot)}\left[\lambda(t)\left(|s_\theta(t,X)|^2 + 2\,\nabla\cdot s_\theta(t,X)\right)\right]. \tag{2.7}$$

However, the problem (2.7) can still be computationally costly when the dimension d is large. Using a neural network for $s_\theta(t,x)$, we need to conduct d times of backward propagation of all parameters to compute $\nabla \cdot s_\theta(t,x)$. This means that the computation of the gradient scales linearly with the dimension, thus making the gradient descent methods not efficient for solving the problem (2.7) with respect to examples such as image data in high dimension. *Denoising score matching* (DSM) [66] serves as a scalable alternative:

$$\tilde{\mathcal{J}}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,T)} \left\{\lambda(t)\mathbb{E}_{X_0\sim p_{data}(\cdot)}\mathbb{E}_{X_t|X_0}\left[|s_\theta(t,X(t)) - \nabla_{X_t}\log p(t,X_t \mid X_0)|^2\right]\right\}.$$

Its equivalent form and other methods such as sliced score matching [59] are discussed in Appendix C.2. Now we replace $\nabla p(t,x)$ with the matched score $s_\theta(t,x)$ in (2.2) to get the backward process:

$$d\overline{X}_t = \left(-b(T-t,\overline{X}_t) + \sigma^2(T-t)\, s_\theta(T-t,\overline{X}_t)\right)dt + \sigma(T-t)d\overline{B}_t,\ \overline{X}_0 \sim p_{\mathrm{noise}}(\cdot). \tag{2.8}$$

3. Theory for contractive DPMs

In this section, we introduce the idea of CDPMs, and present supportive theoretical results. For a DPM specified by (2.1)-(2.8), we consider the Euler-Maruyama discretization of its backward process $\overline{X}$. Fix $\delta > 0$ as the step size, and set $t_k := k\delta$ for $k = 0,\dots,N := \frac{T}{\delta}$. Let $\widehat{X}_0 = \overline{X}_0$, and

$$\begin{aligned}\widehat{X}_k := \widehat{X}_{k-1} + (-b(T-t_k,\widehat{X}_{k-1})&+a(T-t_{k-1})s_\theta(T-t_{k-1},\widehat{X}_{k-1}))\delta \\ &+ \sigma(T-t_{k-1})(B_{t_k} - B_{t_{k-1}}), \quad \text{for } k=1,\dots,N.\end{aligned} \tag{3.1}$$

Our goal is to bound the Wasserstein-2 distance $W_2(p_{\mathrm{data}}(\cdot),\widehat{X}_N)$. Clearly,

$$W_2(p_{\mathrm{data}}(\cdot),\widehat{X}_N) \le W_2(p_{\mathrm{data}}(\cdot),\overline{X}_T) + \left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}}, \tag{3.2}$$

where the first term on the right side of (3.2) is the sampling error at the continuous level, and the second term is the discretization error. We will study these two terms in the next two subsections.

3.1. Sampling error in continuous time.

We are concerned with the term $W_2(p_{\mathrm{data}}(\cdot),\overline{X}_T)$. Existing work [13, 38, 16] established W_2 bounds for OU processes under a bounded support assumption. Closer to our result (and proof) is the concurrent work [23], where a W_2 bound is derived for a class of DPMs with $b(t,x)$ separable in t and x, under a strongly log-concavity assumption.

Assumption 3. *The following conditions hold:*

1. *There exists $r_b : [0,T] \to \mathbb{R}$ such that $(x-x')\cdot(b(t,x)-b(t,x')) \ge r_b(t)|x-x'|^2$ for all t and x,x'.*
2. *There exists $L > 0$ such that $|\nabla\log p(t,x) - \nabla\log p(t,x')| \le L|x-x'|$ for all t and x,x'.*
3. *There exists $\varepsilon > 0$ such that $\mathbb{E}|s_\theta(t,\overline{X}_{T-t}) - \nabla\log p(t,\overline{X}_{T-t})|^2 \le \varepsilon^2$ for all t.*

The condition (1) assumes the monotonicity of $b(t,\cdot)$ and (2) assumes the Lipschitz property of the score functions. In the previous examples, $b(t,x)$ is linear in x so the density $p(t,\cdot)$ is Gaussian-like, and its score is almost affine. Thus, it is reasonable to assume (2). Conditions (1) and (2) are used to quantify how a perturbation of the model parameters in an SDE affects its distribution. The condition (3) specifies how accurate Stein's score is estimated by a blackbox estimation. There has been work (e.g. [10, 34, 44]) on the efficiency of score approximations. So it is possible to replace the condition (3) with those score approximation bounds.

Theorem 4. *Let Assumption 3 hold, and $h > 0$. Define $\eta := W_2(p(T,\cdot),p_{noise}(\cdot))$, and*

$$u(t) := \int_{T-t}^{T}\left(-2r_b(s) + (2L+2h)\sigma^2(s)\right)ds. \tag{3.3}$$

Then we have

$$W_2(p_{data}(\cdot),\overline{X}_T) \le \sqrt{\eta^2 e^{u(T)} + \frac{\varepsilon^2}{2h}\int_0^T \sigma^2(t)e^{u(T)-u(T-t)}dt}. \tag{3.4}$$

The proof of Theorem 4 is deferred to Appendix D. A similar result was given in [36] under an unconventional assumption that the score matching functions $s_\theta(t,x)$, rather than the score functions $\nabla p(t,x)$, are Lipschitz. In fact, the (impractical) assumption that the score matching functions are Lipschitz is not needed at the continuous level, and can be replaced with the Lipschitz condition on the score functions. On the other hand, the Lipschitz property of the score matching functions are required, for technical purposes, to bound the discretization error in Section 3.2.

It is possible to establish sharper bounds under extra (structural) conditions on $(b(\cdot,\cdot),\sigma(\cdot))$, and also specify the dependence in dimension d (e.g. [13, 23]). For instance, if we assume $b(t,x)$ is separable in t and x and linear in x, and $p_{\text{data}}(\cdot)$ is strongly log-concave, then the term $L\sigma^2(s)$ in (3.3) will become $-L'\sigma^2(s)$ for some $L' > 0$. Since the purpose of this paper is to introduce the methodology of CDPMs, we leave the full investigation of its theory to the future work.

Now let's explain contractive DPMs. Looking at the bound (3.4), the sampling error $W_2(p_{\text{data}}(\cdot),\overline{X}_T)$ is linear in the score matching error ε and the initialization error η, and these errors may be amplified in time T – in most aforementioned DPMs, $r_b(t) \le 0$ so $u(t)$ is positive and at least linear. As mentioned earlier, it is problematic if we don't know how good a blackbox score matching $s_\theta(t,x)$ is. Our idea is simply to make $u(t)$ be negative, that is to set $r_b(t) > 0$ sufficiently large, in order to prevent the score matching error from propagating in backward sampling. This yields the class of CDPMs, which is inherently different from existing DPMs in the sense that these DPMs often have contractive forward processes, while our proposal requires contractive backward processes. Quantitatively, we can set for some $\alpha > 0$,

$$\inf_{0 \le t \le T} \left(r_b(t) - (L+h)\sigma^2(t) \right) \ge \alpha. \tag{3.5}$$

In practice, it suffices to design $(b(\cdot,\cdot),\sigma(\cdot))$ with a positive $r_b(t)$. We present three examples, contractive OU processes and contractive subVP SDEs:

(a) Contractive Ornstein-Ulenback (COU) process: $b(t,x) = \theta(x-\mu)$ with $\theta > 0$, $\mu \in \mathbb{R}^d$ and $\sigma(t) = \sigma$. The backward process is:

$$d\overline{X}_t = \left(-\theta(\overline{X}_t - \mu) + \sigma^2 \nabla \log p(T-t,\overline{X}_t) \right) dt + \sigma dB_t, \quad \overline{X}_0 \sim \mathcal{N}\left(0, \frac{\sigma^2}{2\theta}(e^{2\theta T}-1)I\right). \tag{3.6}$$

(b) Contractive variance preserving (CVP) SDE: $b(t,x) = \frac{1}{2}\beta(t)x$ and $\sigma(t) = \sqrt{\beta(t)}$, where $\beta(t) = \beta_{\min} + \frac{t}{T}(\beta_{\max} - \beta_{\min})$. The backward process is:

$$\begin{aligned} d\overline{X}_t = &\left(-\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)\nabla \log p(T-t,\overline{X}_t) \right) dt \\ &+ \sqrt{\beta(T-t)}dB_t, \quad \overline{X}_0 \sim \mathcal{N}\left(0, (e^{\frac{T}{2}(\beta_{\max}+\beta_{\min})} - 1)I\right). \end{aligned} \tag{3.7}$$

(c) Contractive sub-variance preserving (CsubVP) SDEs: $b(t,x) = \frac{1}{2}\beta(t)x$ and $\sigma(t) = \sqrt{\beta(t)(e^{2\int_0^t \beta(s)ds} - 1)}$. By setting $\gamma(t) = e^{2\int_0^t \beta(s)ds}$, the backward process is:

$$\begin{aligned} d\overline{X}_t = &\left(-\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)(\gamma(T-t)-1)\nabla \log p(T-t,\overline{X}_t) \right) dt \\ &+ \sqrt{\beta(T-t)(\gamma(T-t)-1)}dB_t, \quad \overline{X}_0 \sim \mathcal{N}\left(0, (e^{\frac{T}{2}(\beta_{\max}+\beta_{\min})} - 1)^2 I\right). \end{aligned} \tag{3.8}$$

To illustrate, we give a bound for CVP. Recall that a function $\ell : \mathbb{R}^d \to \mathbb{R}$ is κ-strongly concave if $(\nabla \ell(x) - \nabla \ell(y)) \cdot (x-y) \le -\kappa |x-y|^2$.

Theorem 5. *Let $(\overline{X}_t,\, 0 \le t \le T)$ be specified by (3.7) (the backward process of CVP). Assume that $\log p_{data}(\cdot)$ is κ-strongly log-concave, and $\mathbb{E}_{p_{data}(\cdot)}|x|^2 < \infty$. We have*

$$W_2^2(p_{data}(\cdot),\overline{X}_T) \le e^{-2\left(\frac{\kappa}{1+\kappa} - \beta_{\max} hT + \mathcal{O}(e^{-\beta_{\min}T})\right)} \mathbb{E}_{p_{data}(\cdot)}|x|^2 + \frac{\varepsilon^2}{2h(1-2h)}. \tag{3.9}$$

The proof of Theorem 5 is given in Appendix E. It is easy to see from the theorem that CVP (and other CDPMs) allow to control the score matching error ε, at the possible cost of initialization error coming from η. Note that if $\beta_{\max}T$ is asymptotically small, this error is bounded. This requires scaling the hyperparameters with respect to T in the model. We observe in the experiment that tuning a moderate level β is important to let CDPM benefit from contraction while not suffer from the initialization error. Also note that it is not necessary to send $T \to \infty$, and $T = 1$ is taken in [60].

3.2. **Discretization error.** We study the discretization error $\left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}}$ for CDPMs. Classical SDE theory [33] indicates that $\left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}} \le C(T)\delta$, with the constant $C(T)$ exponential in T. Here we show that by a proper choice of the pair $(b(\cdot,\cdot), \sigma(\cdot))$ leading to CDPMs, the constant $C(T)$ can be made independent of T. In other words, the discretization error will not expand over the time. We need some technical assumptions.

Assumption 6. *The following conditions hold:*

(1) There exists $L_\sigma > 0$ such that $|\sigma(t) - \sigma(t')| \le L_\sigma|t - t'|$ for all t, t'.
(2) There exists $R_\sigma > 0$ such that $\sigma(t) \le R_\sigma$ for all t.
(3) There exists $L_b > 0$ such that $|b(t,x) - b(t',x')| \le L_b(|t - t'| + |x - x'|)$ for all t, t' and x, x'.
(4) There exists $L_s > 0$ such that $|s_\theta(t,x) - s_\theta(t',x')| \le L_s(|t - t'| + |x - x'|)$ for all t, t' and x, x'.
(5) There exists $R_s > 0$ such that $|s_\theta(T,x)| \le R_s(1 + |x|)$ for all x.

Next we introduce a contractive assumption that is consistent with (3.5)

Assumption 7. *There exists $\beta > 0$ such that*

$$\int_{T-t}^{T} (r_b(s) - L_s\sigma^2(s))\, ds \ge \beta t, \quad \text{for all } t, \tag{3.10}$$

or simply

$$\beta := \inf_{0 \le t \le T} \left(r_b(t) - L_s\sigma^2(t)\right) > 0. \tag{3.11}$$

Theorem 8. *Let Assumptions 3, 6 and 7 hold. Then there exists $C > 0$ (independent of δ, T) such that for $\delta > 0$ sufficiently small,*

$$\left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}} \le C\sqrt{\delta}. \tag{3.12}$$

The proof of Theorem 8 is given in Appendix F.

4. Connections between CDPM and VE

In this section, we draw connections between CDPM and VE. We first show that VE exhibits some hidden contractive property. Then we show how to exploit the pretrained models such as VE to achieve CDPM, which does not require retraining.

4.1. **VE is implicit CDPM at earlier denoising steps.** We illustrate with an example that the backward process of VE also yields the contractive property at earlier stages of the denoising process. In Figure 1, the angles correspond to the scores of a normal distribution with mean 0 (the case of the VE prior). We see that the two points becomes closer after a denosing step , which provides an explanation of the hidden contraction. However, VE may lose this contractive property when the distribution is close to the target data distribution, which motivates the design of CDPM.

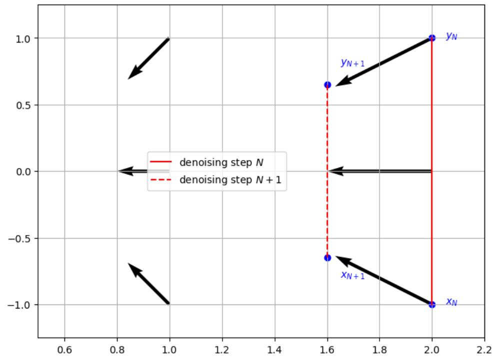


FIGURE 1. Contraction of VE

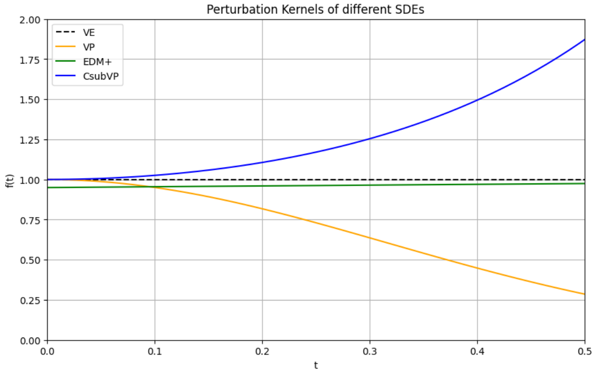


FIGURE 2. Perturbation kernels

4.2. **CDPM is a change of variables of VE.** We show that CDPM, though different from existing DPMs, can be derived from VE via a time/space change, so does not require pretraining the score matching objectives. Note that for COU, CVP and CsubVP, the parameter $b(t, x)$ is separate in t and x. We follow [30], and define the perturbation kernel of CDPM as:

$$X_t \mid X_0 \sim \mathcal{N}\left(f(t)X_0,\, f(t)^2 g(t)^2 I\right),$$

where

$$f(t) = e^{\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})+\frac{t}{2}\beta_{\min}} \quad \text{and} \quad g(t) = f(t) - f^{-1}(t). \tag{4.1}$$

Figure 2 plots different perturbation kernels of SDE models: existing models lead to either a constant or a decreasing kernel, while we propose the kernel be increasing. This yields CsubVP and EDM+, which we will show in Section 5.3.

Denote by $p_{\mathrm{VE}}(t,\cdot)$ and $p_{\mathrm{CsubVP}}(t,\cdot)$ the probability distribution of X_t following VE and CsubVP respectively. Assume that we have access to a pretrained VE score matching $s_{pre}(t,x) \approx \nabla \log p_{VE}(t,x)$. We can then compute the CDPM score by the following transformation, which is read from [30, Equation (12) and (19)].

Theorem 9. *Assume that $\sigma^2_{\max} - \sigma^2_{\min} > g^2(T)$. We have for $t \in [0,T]$,*

$$p_{CsubVP}(t,x) = f(t)^{-d} p_{VE}(\tau(t), x/f(t)), \tag{4.2}$$

where

$$\tau(t) = \frac{T}{2}\frac{\log(1+\frac{g^2(t)}{\sigma^2_{\min}})}{\log(\sigma_{\max}/\sigma_{\min})}. \tag{4.3}$$

So it suffices to take $\nabla \log p_{\mathrm{CsubVP}}(t,x) \approx s_{pre}(\tau(t), x/f(t))$, meaning that we can exploit existing score matching neural nets, or pretrained weights for CDPM sampling.

5. EXPERIMENTS

In this section, we report empirical results on the proposed contractive approach and CDPMs. We conduct experiments on a 1-dimensional synthetic example, Swiss Roll, MNIST, CIFAR10 32×32 and AFHQv2 64×64 datasets.

5.1. **CDPM shows better performance with the same scoring matching error.** The goal is to learn/generate a single point mass at $x_0 = -1$ in dimension one. Since we can compute the score explicitly, we test the performance of different SDE models by adding the SAME noise level of error at each time/point. Implementation details are given in Appendix G.1. Table 1 compares the W_2 errors for OU and COU, with different noise levels and time discretization. As is expected from the theory, COU is more robust to score matching error and time discretization.

Noise Level \ $W_2 \downarrow$	*OU*	*COU*
$\epsilon = 0.02$	0.245	0.22
$\epsilon = 0.05$	0.265	0.227
$\epsilon = 0.1$	0.30	0.23
$\epsilon = 0.2$	0.39	0.25
$\epsilon = 0.5$	0.7	0.42
$\epsilon = 1$	1.3	0.8

(A) time discretization $\Delta t = 0.02$

Noise Level \ $W_2 \downarrow$	*OU*	*COU*
$\epsilon = 0.02$	0.41	0.35
$\epsilon = 0.05$	0.44	0.36
$\epsilon = 0.1$	0.48	0.36
$\epsilon = 0.2$	0.58	0.36
$\epsilon = 0.5$	0.92	0.43
$\epsilon = 1$	1.5	0.7

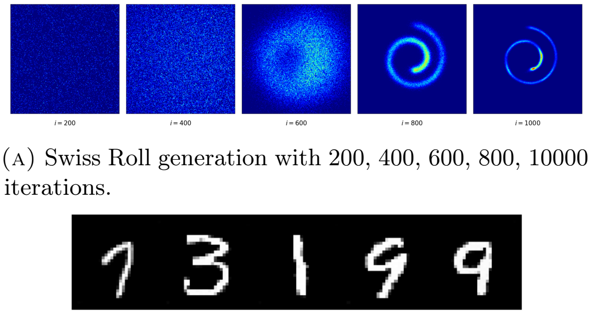
(B) time discretization $\Delta t = 0.05$

Table 1. W_2 distance under the same score matching error.

5.2. **Swiss Roll and MNIST datasets.** We apply CsubVP to Swiss Roll and MNIST datasets. Implementation details are reported in Appendix G.2. Figure 2a shows the evolution process of CsubVP on the Swiss Roll data. Figure 2b provides image synthesis by CsubVP on MNIST. Table 2 shows a clear advantage of CDPMs over other SDE models in terms of W_2 error and FID score.

Model (SDE)	$W_2 \downarrow$	FIDs $\downarrow$
OU	0.29	-
VP [60]	0.33	0.79
subVP [60]	0.34	0.52
VE [60]	0.18	0.20
CDPMs		
COU	**0.10**	-
CsubVP	**0.14**	**0.03**

Table 2. W_2 metric on Swiss Roll and FIDs on MNIST synthesis.



(A) Swiss Roll generation with 200, 400, 600, 800, 10000 iterations.

(B) MNIST synthesis by CsubVP.

5.3. **CIFAR-10 dataset.** We first test the performance of our proposed CsubVP on the task of unconditional synthesis of the CIFAR-10 dataset. We compute and compare FID and inception scores of CsubVP and other SDE models. Implementation details are given in Appendix G.3.

Figure 3 provides image synthesis on CIFAR-10. From Table 3, CsubVP shows the best performance among all known classes of SDE-based diffusion models. In particular, it outperforms VE SDEs (non-deep version in [60]) by achieving both smaller FIDs and higher Inception Scores. ($*$ the model evaluation is conducted on our own machine (4 4090RTX GPUs) given the checkpoints provided by [60]).

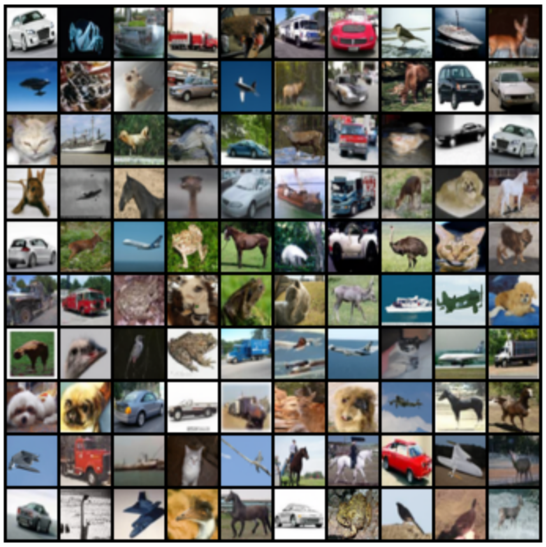


FIGURE 3. CIFAR-10 Synthesis (CsubVP).

Model	Inception ↑	FID ↓
PixelCNN [64]	4.60	65.9
IGEBM [19]	6.02	40.6
ViTGAN [39]	9.30	6.66
StyleGAN2-ADA [31]	9.83	2.92
NCSN [57]	8.87	25.32
NCSNv2 [58]	8.40	10.87
DDPM [26]	9.46	3.17
DDIM, $T = 50$ [54]	-	4.67
DDIM, $T = 100$ [54]	-	4.16
NCSN++		
VP SDE [60]	9.58	2.55
subVP SDE [60]	9.56	2.61
VE SDE [60]	9.68*	2.50*
CsubVP	**10.18**	**2.47**

TABLE 3. Inception & FID.

We also show how to improve the baseline pretrained models using our idea of contraction. Given the checkpoints of EDM [30] on CIFAR10 and AFHQv2 datasets, we modify the perturbation kernel to let $s(0) = 1 - \epsilon < 1$, leading to an increasing function as in Figure 2. This simple technique, motivated by our contraction approach, yields improvement to the EDM baselines as in Table 4. Moreover, we observe the improvement of the sample quality by comparing the images generated by EDM and EDM with contraction, see Figure 4.

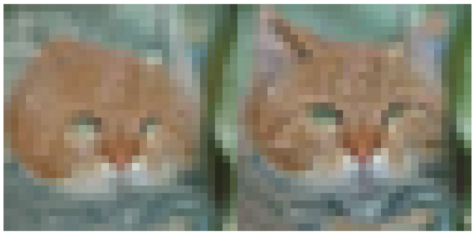


FIGURE 4. (CIFAR10 sample) **LEFT**: EDM, **RIGHT**: EDM with contraction.

Model/FID ↓	EDM*[30]	'+' Contraction	NFE
CIFAR10 32×32			
VP SDE (cond)	1.85	1.83	35
VE SDE (cond)	1.83	1.82	35
VP SDE (uncond)	1.96	1.94	35
VE SDE (uncond)	1.97	1.97	35
AFHQv2 64×64			
VP SDE (uncond)	2.10	2.08	79
VE SDE (uncond)	2.24	2.20	79

TABLE 4. FID scores (*our reruns).

6. CONCLUSION

In this paper, we propose a new criterion – the contraction of backward sampling in the design of SDE-based DPMs. This naturally leads to a novel class of contractive DPMs. The main takeaway is that the contraction of the backward process limits score matching errors from propagating, and controls discretization error as well. Our proposal is supported by theoretical considerations, and is corroborated by experiments. Notably, our proposed contractive subVP outperforms other SDE-based DPMs on CIFAR 10 dataset. Though the intention of this paper is not to beat SOTA diffusion models, CDPMs show promising results that we hope to trigger further research.

There are a few directions to extend this work. First, we assume that score matching errors are bounded in L^2 as in [13, 23, 38]. It is interesting to see whether this assumption can be relaxed to more realistic conditions given the application domain. Second, it is desirable to establish sharp theory for CDPMs, with dimension dependence. Finally, our formulation is based on SDEs, and hence stochastic samplers. We don't look into ODE samplers as in [60, 67, 40, 12]. This leaves open the problems such as whether the proposed CDPMs perform well by ODE samplers, and why the ODE samplers derived from SDEs outperform those directly learnt, as observed in previous studies.

Acknowledgement: We thank Jason Altschuler, Yuxin Chen, Sinho Chewi, Xuefeng Gao, Dan Lacker, Yuting Wei and David Yao for helpful discussions. We gratefully acknowledges financial support through NSF grants DMS-2113779 and DMS-2206038, and by a start-up grant at Columbia University.

References

[1] B. D. O. Anderson. Reverse-time diffusion equation models. *Stochastic Process. Appl.*, 12(3):313–326, 1982.

[2] J. Austin, D. D. Johnson, J. Ho, D. Tarlow, and R. Van Den Berg. Structured denoising diffusion models in discrete state-spaces. In *Neurips*, volume 34, pages 17981–17993, 2021.

[3] J. Benton, V. De Bortoli, A. Doucet, and G. Deligiannidis. Linear convergence bounds for diffusion models via stochastic localization. *arXiv:2308.03686*, 2023.

[4] J. Benton, G. Deligiannidis, and A. Doucet. Error bounds for flow matching methods. *arXiv:2305.16860*, 2023.

[5] A. Borji. Pros and cons of GAN evaluation measures. *Comput. Vis. Image Underst.*, 179:41–65, 2019.

[6] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, and M. Tagliasacchi. AudioLM: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[7] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2018.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell. Language models are few-shot learners. In *Neurips*, volume 33, pages 1877–1901, 2020.

[9] P. Cattiaux, G. Conforti, I. Gentil, and C. Léonard. Time reversal of diffusion processes under a finite entropy condition. *Ann. Inst. Henri Poincaré Probab. Stat.*, 59(4):1844–1881, 2023.

[10] M. Chen, K. Huang, T. Zhao, and M. Wang. Score approximation, estimation and distribution recovery of diffusion models on low-dimensional data. In *ICML*, volume 40, pages 4672–4712, 2023.

[11] M. Chen, S. Mei, J. Fan, and M. Wang. An overview of diffusion models: applications, guided generation, statistical rates and optimization. 2024. arXiv:2404.07771.

[12] S. Chen, S. Chewi, H. Lee, Y. Li, J. Lu, and A. Salim. The probability flow ode is provably fast. 2023. arXiv:2305.11798.

[13] S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *ICLR*, 2023.

[14] S. Chen, G. Daras, and A. Dimakis. Restoration-degradation beyond linear diffusions: A non-asymptotic analysis for ddim-type samplers. In *ICML*, volume 40, pages 4462–4484, 2023.

[15] V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. *arXiv:2208.05314*, 2022.

[16] V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion Schrödinger bridge with applications to score-based generative modeling. In *Neurips*, volume 34, pages 17695–17709, 2021.

[17] P. Dhariwal and A. Nichol. Diffusion models beat GANs on image synthesis. In *Neurips*, volume 34, pages 8780–8794, 2021.

[18] S. Dieleman, L. Sartran, A. Roshannai, N. Savinov, Y. Ganin, P. H. Richemond, A. Doucet, R. Strudel, C. Dyer, and C. Durkan. Continuous diffusion for categorical data. *arXiv:2211.15089*, 2022.

[19] Y. Du, S. Li, J. Tenenbaum, and I. Mordatch. Improved contrastive divergence training of energy based models. *arXiv:2012.01316*, 2020.

[20] N. Fishman, L. Klarner, V. D. Bortoli, E. Mathieu, and M. J. Hutchinson. Diffusion models for constrained domains. *Transactions on Machine Learning Research*, 2023.

[21] H. Föllmer. An entropy approach to the time reversal of diffusion processes. In *Stochastic differential systems (Marseille-Luminy, 1984)*, volume 69 of *Lect. Notes Control Inf. Sci.*, pages 156–163. Springer, Berlin, 1985.

[22] H. Föllmer. Time reversal on Wiener space. In *Stochastic processes—mathematics and physics (Bielefeld, 1984)*, volume 1158 of *Lecture Notes in Math.*, pages 119–129. Springer, Berlin, 1986.

[23] X. Gao, H. M. Nguyen, and L. Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *arXiv:2311.11003*, 2023.

[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, volume 27, pages 2672–2680, 2014.

[25] U. G. Haussmann and E. Pardoux. Time reversal of diffusions. *Ann. Probab.*, 14(4):1188–1205, 1986.

[26] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Neurips*, volume 33, pages 6840–6851, 2020.

[27] A. Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005.

[28] N. Kalchbrenner, A. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu. Video pixel networks. In *ICML*, volume 34, pages 1771–1779, 2017.

[29] I. Karatzas and S. E. Shreve. *Brownian motion and stochastic calculus*, volume 113 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, second edition, 1991.

[30] T. Karras, M. Aittala, T. Aila, and S. Laine. Elucidating the design space of diffusion-based generative models. In *Neurips*, volume 35, pages 26565–26577, 2022.

[31] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila. Training generative adversarial networks with limited data. In *Neurips*, volume 33, pages 12104–12114, 2020.

[32] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. In *Neurips*, volume 34, pages 21696–21707, 2021.

[33] P. E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations*, volume 23 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1992.

[34] F. Koehler, A. Heckett, and A. Risteski. Statistical efficiency of score matching: The view from isoperimetry. In *ICLR*, 2023.

[35] Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *ICLR*, 2021.

[36] D. Kwon, Y. Fan, and K. Lee. Score-based generative modeling secretly minimizes the Wasserstein distance. In *Neurips*, volume 35, pages 20205–20217, 2022.

[37] H. Lee, J. Lu, and Y. Tan. Convergence for score-based generative modeling with polynomial complexity. In *Neurips*, volume 35, pages 22870–22882, 2022.

[38] H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In *ALT*, pages 946–985. PMLR, 2023.

[39] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu. ViTGAN: Training GANs with vision transformers. In *ICLR*, 2022.

[40] G. Li, Y. Wei, Y. Chen, and Y. Chi. Towards faster non-asymptotic convergence for diffusion-based generative models. *arXiv:2306.09251*, 2023.

[41] A. Lou and S. Ermon. Reflected diffusion models. *arXiv:2304.04740*, 2023.

[42] S. Mei and Y. Wu. Deep networks as denoising algorithms: Sample-efficient learning of diffusion models in high-dimensional graphical models. *arXiv:2309.11420*, 2023.

[43] G. N. Milstein and M. V. Tretyakov. *Stochastic numerics for mathematical physics*. Scientific Computation. Springer-Verlag, Berlin, 2004.

[44] K. Oko, S. Akiyama, and T. Suzuki. Diffusion models are minimax optimal distribution estimators. In *ICML*, volume 40, page 26517–26582, 2023.

[45] OpenAI. Sora: Creating video from text. 2024. Available at `https://openai.com/sora`.

[46] J. Pidstrigach. Score-based generative models detect manifolds. In *Neurips*, volume 35, pages 35852–35865, 2022.

[47] R. Prenger, R. Valle, and B. Catanzaro. Waveglow: A flow-based generative network for speech synthesis. In *ICASSP*, pages 3617–3621, 2019.

[48] J. Quastel. Time reversal of degenerate diffusions. In *In and out of equilibrium (Mambucaba, 2000)*, volume 51 of *Progr. Probab.*, pages 249–257. Birkhäuser Boston, Boston, MA, 2002.

[49] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv:2204.06125*, 2022.

[50] A. Razavi, A. Van den Oord, and O. Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Neurips*, volume 32, pages 14866–14876, 2019.

[51] P. H. Richemond, S. Dieleman, and A. Doucet. Categorical SDEs with simplex diffusion. *arXiv:2210.14784*, 2022.

[52] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[53] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, volume 32, pages 2256–2265, 2015.

[54] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. In *ICLR*, 2021.

[55] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. In *ICML*, volume 40, page 32211–32252, 2023.

[56] Y. Song, C. Durkan, I. Murray, and S. Ermon. Maximum likelihood training of score-based diffusion models. In *Neurips*, volume 34, pages 1415–1428, 2021.

[57] Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Neurips*, volume 32, page 11918–11930, 2019.

[58] Y. Song and S. Ermon. Improved techniques for training score-based generative models. In *Neurips*, volume 33, pages 12438–12448, 2020.

[59] Y. Song, S. Garg, J. Shi, and S. Ermon. Sliced score matching: A scalable approach to density and score estimation. In *UAI*, volume 35, pages 574–584, 2020.

[60] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

[61] W. Tang and H. Zhao. Score-based diffusion models via stochastic differential equations–a technical tutorial. 2024. arXiv:2402.07487.

[62] W. Tang and X. Y. Zhou. Tail probability estimates of continuous-time simulated annealing processes. *Numer. Algebra Control Optim.*, 13(3-4):473–485, 2023.

[63] A. Vahdat, K. Kreis, and J. Kautz. Score-based generative modeling in latent space. In *Neurips*, volume 34, pages 11287–11302, 2021.

[64] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, and A. Graves. Conditional image generation with pixelcnn decoders. In *NIPS*, volume 29, pages 4797–4805, 2016.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, volume 30, pages 6000–6010, 2017.

[66] P. Vincent. A connection between score matching and denoising autoencoders. *Neural Comput.*, 23(7):1661–1674, 2011.

[67] Y. Xu, Z. Liu, M. Tegmark, and T. Jaakkola. Poisson flow generative models. In *Neurips*, volume 35, pages 16782–16795, 2022.

[68] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Comput. Surv.*, 56(4):1–39, 2023.

[69] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Neurips*, volume 32, pages 5753–5763, 2019.

[70] H. Zhao, W. Tang, and D. D. Yao. Policy optimization for continuous reinforcement learning. In *Neurips*, volume 36, 2023.

Appendix

A. Proof of Theorem 1.

Proof. Here we give a heuristic derivation of the time reversal formula (2.3). First, the infinitesimal generator of X is $\mathcal{L} := \frac{1}{2}\sigma^2(t)\Delta + b\cdot\nabla$. It is known that the density $p(t,x)$ satisfies the the Fokker–Planck equation:

$$\frac{\partial}{\partial t}p(t,x) = \mathcal{L}^* p(t,x), \tag{.1}$$

where $\mathcal{L}^* := \frac{1}{2}\sigma^2(t)\Delta - \nabla\cdot b$ is the adjoint of $\mathcal{L}$. Let $\overline{p}(t,x) := p(T-t,x)$ be the probability density of the time reversal $\overline{X}$. By (.1), we get

$$\frac{\partial}{\partial t}\overline{p}(t,x) = -\frac{1}{2}\sigma^2(t)\Delta\overline{p}(t,x) + \nabla\cdot\left(b(T-t,x)\,\overline{p}(t,x)\right). \tag{.2}$$

On the other hand, we expect the generator of $\overline{X}$ to be $\overline{\mathcal{L}} := \frac{1}{2}\overline{\sigma}^2(t)\Delta + \overline{b}\cdot\nabla$. The Fokker-Planck equation for $\overline{p}(t,x)$ is

$$\frac{\partial}{\partial t}\overline{p}(t,x) = \frac{1}{2}\overline{\sigma}^2(t)\Delta\overline{p}(t,x) - \nabla\cdot\left(\overline{b}(t,x)\,\overline{p}(t,x)\right). \tag{.3}$$

Comparing (.2) and (.3), we set $\overline{\sigma}(t) = \sigma(T-t)$ and then get

$$\left(b(T-t,x) + \overline{b}(t,x)\right)\overline{p}(t,x) = \sigma^2(T-t)\,\nabla\overline{p}(t,x).$$

This yields the desired result. □

Let's comment on Theorem 1. [25] proved the result by assuming that $b(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ are globally Lipschitz, and the density $p(t,x)$ satisfies an a priori H^1 bound. The implicit condition on $p(t,x)$ is guaranteed if $\partial_t + \mathcal{L}$ is hypoelliptic. These conditions were relaxed in [48]. In another direction, [21, 22] used an entropy argument to prove the time reversal formula in the case $\sigma(t) = \sigma$. This approach was further developed in [9] which made connections to optimal transport.

B. Examples of DPMs. We present a few examples of DPMs.

(a) OU processes: $b(t,x) = \theta(\mu - x)$ with $\theta > 0$, $\mu\in\mathbb{R}^d$; $\sigma(t) = \sigma > 0$. The distribution of $(X_t\,|\,X_0 = x)$ is $\mathcal{N}(\mu + (x-\mu)e^{-\theta t}, \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t})I)$, which is approximately $\mathcal{N}(\mu, \frac{\sigma^2}{2\theta}I)$ as t is large. The backward process specializes to

$$d\overline{X}_t = (\theta(\overline{X}_t - \mu) + \sigma^2\nabla\log p(T-t,\overline{X}_t))dt + \sigma dB_t,\ \overline{X}_0 \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{2\theta}I\right). \tag{.4}$$

(b) VE-SDE: $b(t,x) = 0$ and $\sigma(t) = \sigma_{\min}\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{t}{T}}\sqrt{\frac{2}{T}\log\frac{\sigma_{\max}}{\sigma_{\min}}}$ with $\sigma_{\min} \ll \sigma_{\max}$. The distribution of $(X_t\,|\,X_0 = x)$ is $\mathcal{N}\left(x, \sigma^2_{\min}\left(\left(\frac{\sigma_{\max}}{\sigma_{\min}}\right)^{\frac{2t}{T}} - 1\right)I\right)$, which can be approximated by $\mathcal{N}(0,\sigma^2_{\max}I)$ at $t = T$. The backward process is:

$$d\overline{X}_t = \sigma^2(T-t))\nabla\log p(T-t,\overline{X}_t) + \sigma(T-t)dB_t,\ \overline{X}_0 \sim \mathcal{N}(0,\sigma^2_{\max}I). \tag{.5}$$

(c) VP-SDE: $b(t,x) = -\frac{1}{2}\beta(t)x$ and $\sigma(t) = \sqrt{\beta(t)}$, where $\beta(t) := \beta_{\min} + \frac{t}{T}(\beta_{\max} - \beta_{\min})$ with $\beta_{\min} \ll \beta_{\max}$. The distribution of $(X_t\,|\,X_0 = x)$ is

$$\mathcal{N}(e^{-\frac{t^2}{4T}(\beta_{\max}-\beta_{\min}) - \frac{t}{2}\beta_{\min}}x, (1 - e^{-\frac{t^2}{2T}(\beta_{\max}-\beta_{\min}) - t\beta_{\min}})I),$$

which can be approximated by $\mathcal{N}(0,I)$ at $t = T$. The backward process is:

$$\begin{aligned} d\overline{X}_t = &\left(\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)\nabla\log p(T-t,\overline{X}_t)\right)dt \\ &+ \sqrt{\beta(T-t))}dB_t,\ \overline{X}_0 \sim \mathcal{N}(0,I). \end{aligned} \tag{.6}$$

(d) subVP-SDE: $b(t,x) = -\frac{1}{2}\beta(t)x$ and $\sigma(t) = \sqrt{\beta(t)(1-e^{-2\int_0^t \beta(s)ds})}$. The distribution of $(X_t \,|\, X_0 = x)$ is $\mathcal{N}(e^{-\frac{t^2}{4T}(\beta_{\max}-\beta_{\min})-\frac{t}{2}\beta_{\min}}x, (1-e^{-\frac{t^2}{2T}(\beta_{\max}-\beta_{\min})-t\beta_{\min}})^2 I)$, which can be approximated by $\mathcal{N}(0,I)$ at $t=T$. The backward process is:

$$\begin{aligned} d\overline{X}_t = & \left(\frac{1}{2}\beta(T-t)\overline{X}_t + \beta(T-t)(1-\gamma(T-t))\nabla \log p(T-t, \overline{X}_t)\right)dt \\ & + \sqrt{\beta(T-t)(1-\gamma(T-t))}dB_t, \ \overline{X}_0 \sim \mathcal{N}(0,I), \end{aligned} \tag{.7}$$

where $\gamma(t) := e^{-2\int_0^t \beta(s)ds} = e^{-\frac{t^2}{T}(\beta_{\max}-\beta_{\min})-2t\beta_{\min}}$.

C. Score matching.

C.1. *Proof of Theorem 2.*

Proof. We have

$$\begin{aligned} \nabla_\theta \mathcal{J}_{\mathrm{ISM}}(\theta) &= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] - 2\mathbb{E}_{p(t,\cdot)}\left[\nabla_\theta s_\theta(t,X)\cdot\nabla \log p(t,X)\right] \\ &= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] - 2\int \nabla_\theta s_\theta(t,x)\cdot \nabla p(t,x)dx \\ &= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] - 2\,\nabla_\theta \int s_\theta(t,x)\cdot \nabla p(t,x)dx \\ &= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2\right] + 2\nabla_\theta \int \nabla\cdot s_\theta(t,x)\, p(t,x)dx \\ &= \nabla_\theta \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2 + 2\,\nabla\cdot s_\theta(t,X)\right] = \nabla_\theta \widetilde{\mathcal{J}}(\theta), \end{aligned}$$

where we use the divergence theorem in the fourth equation. □

C.2. *Scalable score matching methods.*

(a) Sliced score matching. One way is that we further address the term $\nabla_x \cdot s_\theta(t,x)$ by random projections. The method proposed in [59] is called sliced score matching. Considering the Jacobian matrix $\nabla s_\theta(t,x) \in \mathbb{R}^{d\times d}$, we have

$$\nabla\cdot s_\theta(t,x) = \mathrm{Tr}(\nabla s_\theta(t,x)) = \mathbb{E}_{v\sim\mathcal{N}(0,I)}\left[v^\top \nabla s_\theta(t,x) v\right].$$

We can then rewrite the training objective as:

$$\min_\theta \widetilde{\mathcal{J}}_{\mathrm{SSM}}(\theta) = \mathbb{E}_{t\in\mathcal{U}(0,T)}\mathbb{E}_{v_t\sim\mathcal{N}(0,I)}\mathbb{E}_{p(t,\cdot)}\left[\lambda(t)\left(\|s_\theta(t,X)\|^2 + 2\,v^\top\nabla(v^\top s_\theta(t,x))\right)\right]. \tag{.8}$$

which can be computed easily. It requires only two times of back propagation, as $v^\top s_\theta(t,x)$ can be seen as adding a layer of the inner product between v and s_θ.

(b) Denoising score matching. The second way is that we go back to the objective (2.6), and use a nonparametric estimation. The idea stems from [27, 66], in which it was shown that $\mathcal{J}_{ESM}$ is equivalent to the following denoising score matching (DSM) objective:

$$\tilde{\mathcal{J}}_{\mathrm{DSM}}(\theta) = \mathbb{E}_{t\sim\mathcal{U}(0,T)}\left\{\lambda(t)\mathbb{E}_{X_0\sim p_{data}(\cdot)}\mathbb{E}_{X_t|X_0}\left[|s_\theta(t,X(t)) - \nabla_{X_t}\log p(t,X_t \mid X_0)|^2\right]\right\}$$

Theorem 10. *Let* $\mathcal{J}_{DSM}(\theta) := \mathbb{E}_{X_0\sim p_{data}(\cdot)}\mathbb{E}_{X_t|X_0}\left[|s_\theta(t,X(t)) - \nabla_{X_t}\log p(t,X_t \mid X_0)|^2\right]$. *Under suitable conditions on* s_θ*, we have* $\mathcal{J}_{DSM}(\theta) = \mathcal{J}_{ESM}(\theta) + C$ *for some* C *independent of* θ*. Consequently, the minimum point of* $\mathcal{J}_{DSM}$ *and that of* $\mathcal{J}_{ESM}$ *coincide.*

Proof. We have

$$\begin{aligned}\mathcal{J}_{\text{ESM}}(\theta) &= \mathbb{E}_{p(t,\cdot)}|s_\theta(t,X) - \nabla \log p(t,X)|^2 \\ &= \mathbb{E}_{p(t,\cdot)}\left[|s_\theta(t,X)|^2 - 2s_\theta(t,X)^\top \nabla \log p(t,X) + |\nabla \log p(t,X)|^2\right].\end{aligned}$$

Consider the inner product term, rewriting it as:

$$\begin{aligned}\mathbb{E}_{p(t,\cdot)}\left[s_\theta(t,X)^\top \nabla \log p(t,X)\right] &= \int_x p(t,x) s_\theta(t,x)^\top \nabla \log p(t,x) \mathrm{d}x \\ &= \int_x s_\theta(t,x)^\top \nabla p(t,x) \mathrm{d}x \\ &= \int_x s_\theta(t,x)^\top \nabla \int_{x_0} p(0,x_0) p(t,x|x_0) \mathrm{d}x_0 \mathrm{d}x \\ &= \int_{x_0}\int_x s_\theta(t,x)^\top p(0,x_0) \nabla p(t,x|x_0) \mathrm{d}x \mathrm{d}x_0 \\ &= \int_{x_0} p(0,x_0) \int_x s_\theta(t,x)^\top p(t,x|x_0) \nabla \log p(t,x|x_0) \mathrm{d}x \mathrm{d}x_0 \\ &= \mathbb{E}_{X_0 \sim p(0,\cdot)} \mathbb{E}_{X_t|X_0}\left[s_\theta(t,X(t))^\top \nabla_{X_t} \log p(t,X_t \mid X_0)\right],\end{aligned}$$

combining $\mathbb{E}_X|s_\theta(t,X)|^2 = \mathbb{E}_{X_0}\mathbb{E}_{X|X_0}|s_\theta(t,X)|^2$ concludes our proof. □

The intuition of DSM is that following the gradient s_θ of the log density at some corrupted point $\tilde{x}$ should ideally move us towards the clean sample x. The reason that the objective $\mathcal{J}_{DSM}$ is comparatively easy to solve is that conditional distribution usually satisfies a good distribution, like Gaussian kernel, i.e. $p(X_t \mid X_0) \sim N\left(X_t; \mu_t(X_0), \sigma_t^2 I\right)$, which is satisfied in many cases of DPM, we can explicitly compute that:

$$\nabla_{X_t} \log p(t, X_t \mid X_0) = \frac{1}{\sigma_t^2}(\mu_t(X_0) - X_t).$$

Direction $\frac{1}{\sigma_t^2}(X_0 - \mu_t(X_0))$ clearly corresponds to moving from $\tilde{x}$ back towards clean sample x, and we want s_θ to match that as best it can. Moreover, empirically validated by e.g. [60], a good candidate of $\lambda(t)$ is chosen as:

$$\lambda(t) \propto 1/\mathbb{E}\left[|\nabla_{X_t} \log p(t, X_t \mid X_0)|^2\right] = \sigma_t^2$$

Thus, our final optimization objective is:

$$\begin{aligned}\tilde{\mathcal{J}}_{\text{DSM}}(\theta) &= \mathbb{E}_{t\sim\mathcal{U}(0,T)}\left\{\sigma_t^2 \mathbb{E}_{X_0\sim p_{data}(\cdot)} \mathbb{E}_{X_t|X_0}\left[|s_\theta(t,X(t)) - \nabla_{X_t}\log p(t,X_t \mid X_0)|^2\right]\right\} \\ &= \mathbb{E}_{t\sim\mathcal{U}(0,T)}\left\{\mathbb{E}_{X_0\sim p_{data}(\cdot)} \mathbb{E}_{X_t|X_0}\left[\left|\sigma_t s_\theta(t,X(t)) + \frac{X_t - \mu_t(X_0)}{\sigma_t}\right|^2\right]\right\} \\ &= \mathbb{E}_{t\sim\mathcal{U}(0,T)}\left\{\mathbb{E}_{X_0\sim p_{data}(\cdot)} \mathbb{E}_{\epsilon_t\sim\mathcal{N}(0,I)}\left[|\sigma_t s_\theta(t, \mu_t(X_0) + \sigma_t \epsilon_t) + \epsilon_t|^2\right]\right\}\end{aligned}$$

where the second equality holds when $X_t \mid X_0$ follows a conditionally normal and the third equality follows from a reparameterization/change of variables.

D. Proof of Theorem 4. The idea relies on coupling, which is similar to [70, Lemma 4]. Consider the coupled SDEs:

$$\begin{cases} dY_t = \left(-b(T-t,Y_t) + \sigma^2(T-t)\nabla \log p(T-t,Y_t)\right)dt + \sigma(T-t)dB_t, \\ dZ_t = \left(-b(T-t,Z_t) + \sigma^2(T-t)s_\theta(T-t,Z_t)\right)dt + \sigma(T-t)dB_t, \end{cases}$$

where (Y_0, Z_0) are coupled to achieve $W_2(p(T,\cdot), p_{\text{noise}}(\cdot))$, i.e. $\mathbb{E}|Y_0 - Z_0|^2 = W_2(p(T,\cdot), p_{\text{noise}}(\cdot))$. It is easy to see that

$$W_2^2(p_{\text{data}}(\cdot), \overline{X}_T) \le \mathbb{E}|Y_T - Z_T|^2. \tag{.9}$$

So the goal is to bound $\mathbb{E}|Y_T - Z_T|^2$. By Itô's formula, we get

$$\begin{aligned} d|Y_t - Z_t|^2 = 2(Y_t - Z_t) \cdot (&-b(T-t, Y_t) + \sigma^2(T-t)\nabla \log p(T-t, Y_t) \\ &+ b(T-t, Z_t) - \sigma^2(T-t) s_\theta(T-t, Z_t))dt \end{aligned}$$

which implies that

$$\begin{aligned} \frac{d\,\mathbb{E}|Y_t - Z_t|^2}{dt} &= -2 \underbrace{\mathbb{E}((Y_t - Z_t) \cdot (b(T-t, Y_t) - b(T-t, Z_t)))}_{(a)} \\ &\quad + 2 \underbrace{\mathbb{E}((Y_t - Z_t) \cdot \sigma^2(T-t)(\nabla \log p(T-t, Y_t) - s_\theta(T-t, Z_t)))}_{(b)}. \end{aligned} \tag{.10}$$

By Assumption 3 (1), we get

$$(a) \ge r_b(T-t)\, \mathbb{E}|Y_t - Z_t|^2. \tag{.11}$$

Moreover,

$$\begin{aligned} (b) &= \sigma^2(T-t)\Bigg(\mathbb{E}((Y_t - Z_t) \cdot (\nabla \log p(T-t, Y_t) - \nabla \log p(T-t, Z_t))) \\ &\qquad + \mathbb{E}((Y_t - Z_t) \cdot (\nabla \log p(T-t, Z_t) - s_\theta(T-t, Z_t)))\Bigg) \\ &\le \sigma^2(T-t) \left(L\, \mathbb{E}|Y_t - Z_t|^2 + h\mathbb{E}|Y_t - Z_t|^2 + \frac{1}{4h}\varepsilon^2 \right), \end{aligned} \tag{.12}$$

where we use Assumption 3 (2)(3) in the last inequality. Combining (.10), (.11) and (.12), we have

$$\frac{d\,\mathbb{E}|Y_t - Z_t|^2}{dt} \le \big(-2r_b(T-t) + (2h + 2L)\sigma^2(T-t)\big)\, \mathbb{E}|Y_t - Z_t|^2 + \frac{\varepsilon^2}{2h}\sigma^2(T-t). \tag{.13}$$

Applying Grönwall's inequality, we have:

$$\mathbb{E}|Y_T - Z_T|^2 \le e^{u(T)}\mathbb{E}|Y_0 - Z_0|^2 + \frac{\varepsilon^2}{2h}\int_0^T \sigma^2(T-t)e^{u(T)-u(t)}dt,$$

which combined with (.9) yields (3.4).

E. Proof of Theorem 5. Recall that $r_b(t) = \frac{1}{2}\beta(t)$, $\sigma(t) = \sqrt{\beta(t)}$ and $p_{\text{noise}}(\cdot) \sim \mathcal{N}(0, (e^{\frac{T}{2}(\beta_{\max}+\beta_{\min})} - 1)I)$. By [23, Proposition 10], if $\log p_{\text{data}}(\cdot)$ is κ-strongly log-concave, then $\nabla \log p(T-t,\cdot)$ is $\kappa\left(e^{\int_0^{T-t}\beta(s)ds} + \kappa\int_0^{T-t} e^{\int_s^{T-t}\beta(v)dv}\beta(s)ds\right)^{-1}$-strongly concave. Thus, the term $\mathbb{E}((Y_t - Z_t)\cdot(\nabla \log p(T-t, Y_t) - \nabla \log p(T-t, Z_t)))$ in (.12) is bounded from above by

$$-\frac{\kappa}{e^{\int_0^{T-t}\beta(s)ds} + \kappa\int_0^{T-t} e^{\int_s^{T-t}\beta(v)dv}\beta(s)ds}\mathbb{E}|Y_t - Z_t|^2,$$

instead of $L\mathbb{E}|Y_t - Z_t|^2$. Consequently, the bound (3.4) holds by replacing $u(t)$ with

$$u_{\text{CVP}}(t) := \int_{T-t}^T \beta(s)\left(-1 + 2h - \frac{2\kappa}{e^{\int_0^s \beta(v)dv} + \kappa\int_0^s e^{\int_v^s \beta(u)du}\beta(v)dv}\right) ds. \tag{.14}$$

Note that $u_{\text{CVP}}(T) \le -\int_0^T \beta(s)ds + 2\beta_{\max}hT - \frac{2\kappa}{\kappa+1}\left(1 - e^{-\beta_{\min}T}\right)$ and $u_{\text{CVP}}(T) - u_{\text{CVP}}(T-t) \le \beta_{\max}(2h-1)t$. Moreover, $W_2^2(p(T,\cdot), p_{\text{noise}}(\cdot)) \le e^{\int_0^T \beta(s)ds}\mathbb{E}_{p_{\text{data}}(\cdot)}|x|^2$. Combining (3.4) with the above estimates yields (3.9).

F. Proof of Theorem 8. The analysis of the error $\left(\mathbb{E}|\overline{X}_T - \widehat{X}_N|^2\right)^{\frac{1}{2}}$ relies on the following lemmas. The lemma below proves the contraction of the backward SDE $\overline{X}$.

Lemma 11. *Let $(\overline{X}^x_t,\, 0 \le t \le T)$ be defined by (2.8) with $\overline{X}^x_0 = x$. Let Assumptions 3, 6 and 7 hold. Then*

$$\left(\mathbb{E}|\overline{X}^x_t - \overline{X}^y_t|^2\right)^{\frac{1}{2}} \le \left(\mathbb{E}|x-y|^2\right)^{\frac{1}{2}} \exp(-2\beta t), \quad \textit{for all } t, \tag{.15}$$

where $\overline{X}^x$ and $\overline{X}^y$ be coupled, i.e. they are driven by the same Brownian motion with (different) initial values x and y respectively (x and y represent two random variables).

Proof. Note that

$$\begin{aligned} d|\overline{X}^x_s - \overline{X}^y_s|^2 = 2\left(\overline{X}^x_s - \overline{X}^y_s\right)\cdot\Big(& - b(T-s,\overline{X}^x_s) + \sigma^2(T-s)s_\theta(T-s,\overline{X}^x_s) \\ & + b(T-s,\overline{X}^y_s) - \sigma^2(T-s)s_\theta(\overline{X}^y_s)\Big)ds. \end{aligned}$$

Thus,

$$\begin{aligned} \frac{d}{ds}\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2 = & -2\underbrace{\mathbb{E}\left[(\overline{X}^x_s - \overline{X}^y_s)\cdot(b(T-s,\overline{X}^x_s) - b(T-s,\overline{X}^y_s))\right]}_{(a)} \\ & + 2\underbrace{\mathbb{E}\left[(\overline{X}^x_s - \overline{X}^y_s)\sigma^2(T-s)(s_\theta(T-s,\overline{X}^x_s) - s_\theta(T-s,\overline{X}^y_s))\right]}_{(b)}. \end{aligned} \tag{.16}$$

By Assumption 3 (1), we get

$$(a) \ge r_b(T-s)\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2. \tag{.17}$$

By Assumption 6 (4), we obtain

$$(b) \le L_s\sigma^2(T-s)\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2. \tag{.18}$$

Combining (.16), (.17) and (.18) yields

$$\frac{d}{ds}\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2 \le -2(r_b(T-s) - L_s\sigma^2(T-s))\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2.$$

By Grönwall's inequality, we have

$$\mathbb{E}|\overline{X}^x_s - \overline{X}^y_s|^2 \le \mathbb{E}|x-y|^2 \exp\left(-2\int_0^t (r_b(T-s) - L_s\sigma^2(T-s))ds\right),$$

which, by the condition (3.10), yields (.15) □

Next we deal with the *local (one-step) discretization error* of the process $\overline{X}$. Fixing $t_\star \le T-\delta$, the (one-step) discretization of $\overline{X}$ starting at $\overline{X}_{t_\star} = x$ is:

$$\widehat{X}^{t_\star,x}_1 = x + (-b(T-t_\star,x) + \sigma^2(T-t_\star)s_\theta(T-t_\star,x))\delta + \sigma(T-t_\star)(B_{t_\star+\delta} - B_{t_\star}). \tag{.19}$$

The following lemma provides an estimate of the local discretization error.

Lemma 12. *Let $(\overline{X}^{t_\star,x}_t,\, t_\star \le t \le T)$ be defined by (2.8) with $\overline{X}^{t_\star,x}_{t_\star} = x$, and $\widehat{X}^{t_\star,x}_1$ be given by (.19). Let Assumption 6 hold. Then for δ sufficiently small (i.e. $\delta \le \overline{\delta}$ for some $\overline{\delta} < 1$), there exists $C_1, C_2 > 0$ independent of δ and x such that*

$$\left(\mathbb{E}|\overline{X}^{t_\star,x}_{t_\star+\delta} - \widehat{X}^{t_\star,x}_1|^2\right)^{\frac{1}{2}} \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^{\frac{3}{2}}, \tag{.20}$$

$$|\mathbb{E}(\overline{X}^{t_\star,x}_{t_\star+\delta} - \widehat{X}^{t_\star,x}_1)| \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^{\frac{3}{2}}. \tag{.21}$$

Proof. For ease of presentation, we write $\overline{X}_t$ (resp. $\widehat{X}_1$) for $\overline{X}_t^{t_\star,x}$ (resp. $\widehat{X}_1^{t_\star,x}$). Without loss of generality, set $t_\star = 0$. We have

$$\begin{aligned}\overline{X}_\delta &= x + \int_0^\delta -b(T-t,\overline{X}_t) + \sigma^2(T-t)s_\theta(T-t,\overline{X}_t)dt + \int_0^\delta \sigma(T-t)dB_t,\\ \widehat{X}_1 &= x + \int_0^\delta -b(T,x) + \sigma^2(T)s_\theta(T,x)dt + \int_0^\delta \sigma(T)dB_t.\end{aligned}$$

So

$$\begin{aligned}&\mathbb{E}|\overline{X}_\delta - \widehat{X}_1|^2\\ &= \mathbb{E}\bigg|\int_0^\delta b(T,x) - b(T-t,\overline{X}_t)dt + \int_0^\delta \sigma^2(T-t)s_\theta(T-t,\overline{X}_t) - \sigma^2(T)s_\theta(T,x)dt\\ &\qquad\qquad + \int_0^\delta \sigma(T-t) - \sigma(T)dB_t\bigg|^2\\ &\le 3\mathbb{E}\bigg(\bigg|\int_0^\delta b(T,x) - b(T-t,\overline{X}_t)dt\bigg|^2 + \bigg|\int_0^\delta \sigma^2(T-t)s_\theta(T-t,\overline{X}_t) - \sigma^2(T)s_\theta(T,x)dt\bigg|^2\\ &\qquad\qquad + \bigg|\int_0^\delta \sigma(T-t) - \sigma(T)dB_t\bigg|^2\bigg) \qquad (.22)\\ &\le 3\bigg(\underbrace{\delta\int_0^\delta \mathbb{E}|b(T,x) - b(T-t,\overline{X}_t)|^2dt}_{(a)} + \underbrace{\delta\int_0^\delta \mathbb{E}|\sigma^2(T-t)s_\theta(T-t,\overline{X}_t) - \sigma^2(T)s_\theta(T,x)|^2dt}_{(b)}\\ &\qquad\qquad + \underbrace{\int_0^\delta |\sigma(T-t) - \sigma(T)|^2dt}_{(c)}\bigg),\end{aligned}$$

where we use the Cauchy–Schwarz inequality and Itô's isometry in the last inequality. By Assumption 6 (1), we get

$$(c) \le \int_0^\delta L_\sigma^2 t^2 dt = \frac{L_\sigma^2}{3}\delta^3. \tag{.23}$$

By Assumption 6 (3), we have

$$(a) \le \int_0^\delta 2L_b^2(t^2 + \mathbb{E}|\overline{X}_t - x|^2)dt = 2L_b^2\left(\frac{\delta^3}{3} + \int_0^\delta \mathbb{E}|\overline{X}_t - x|^2dt\right).$$

According to [33, Theorem 4.5.4], we have $\mathbb{E}|\overline{X}_t - x|^2 \le C(1+\mathbb{E}|x|^2)te^{Ct}$ for some $C > 0$ (independent of x). Consequently, for $t \le \delta$ sufficiently small (bounded by $\overline{\delta} < 1$),

$$\mathbb{E}|\overline{X}_t - x|^2 \le C'(1+\mathbb{E}|x|^2)t, \quad \text{for some } C' > 0 \text{ (independent of } \delta, x).$$

We then get

$$(a) \le 2L_b^2\left(\frac{\delta^3}{3} + \frac{C'(1+\mathbb{E}|x|^2)}{2}\delta^2\right) \le 2L_b^2\left(\frac{1}{3} + \frac{C'}{2} + \frac{C'}{2}\mathbb{E}|x|^2\right)\delta^2. \tag{.24}$$

Similarly, we obtain by Assumption 6 (1)(2)(4)(5):

$$(b) \le C''(1+\mathbb{E}|x|^2)\delta^2, \quad \text{for some } C'' > 0 \text{ (independent of } \delta, x). \tag{.25}$$

Combining (.22), (.23), (.24) and (.25) yields the estimate (.20).

Next we have

$$
\begin{aligned}
&|\mathbb{E}(\overline{X}_\delta - \widehat{X}_1)| \\
&= \left|\mathbb{E}\int_0^\delta b(T,x) - b(T-t,\overline{X}_t)dt + \mathbb{E}\int_0^\delta \sigma^2(T-t)s_\theta(T-t,\overline{X}_t) - \sigma^2(T)s_\theta(T,x)dt\right| \\
&\le \int_0^\delta \mathbb{E}|b(T,x) - b(T-t,\overline{X}_t)|dt + \int_0^\delta \mathbb{E}|\sigma^2(T-t)s_\theta(T-t,\overline{X}_t) - \sigma^2(T)s_\theta(T,x)|dt \\
&\le C''' \int_0^\delta \big(t(1+\mathbb{E}|x|) + \mathbb{E}|\overline{X}_t - x|\big)\,dt \\
&\le C''''(1+\sqrt{\mathbb{E}|x|^2})\delta^{\frac{3}{2}}, \quad \text{for some } C'''' > 0 \text{ (independent of } \delta, x).
\end{aligned}
$$

where the third inequality follows from Assumption 6, and the last inequality is due to the fact that $\mathbb{E}|\overline{X}_t - x| \le \big(\mathbb{E}|\overline{X}_t - x|^2\big)^{\frac{1}{2}} \le \sqrt{C'(1+\mathbb{E}|x|^2)t}$. This yields the estimate (.21). □

Proof of Theorem 8. The proof is split into four steps.

Step 1. Recall that $t_k = k\delta$ for $k = 0, \dots, N$. Denote $\overline{X}_k := \overline{X}_{t_k}$, and let

$$e_k := \left(\mathbb{E}|\overline{X}_k - \widehat{X}_k|^2\right)^{\frac{1}{2}}.$$

The idea is to build a recursion for the sequence $(e_k)_{k=0,\dots,N}$. Also write $(\overline{X}_t^{t_\star,x},\ t_\star \le t \le T)$ to emphasize that the reversed SDE (2.8) starts at $\overline{X}_{t_\star}^{t_\star,x} = x$, so $\overline{X}_{k+1} = \overline{X}_{t_{k+1}}^{t_k,\overline{X}_k}$. We have

$$
\begin{aligned}
e_{k+1}^2 &= \mathbb{E}\left|\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} + \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1}\right|^2 \\
&= \underbrace{\mathbb{E}|\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k}|^2}_{(a)} + \underbrace{\mathbb{E}|\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1}|^2}_{(b)} + 2\underbrace{\mathbb{E}\left[(\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k})(\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1})\right]}_{(c)}.
\end{aligned}
\tag{.26}
$$

Step 2. We analyze the term (a) and (b). By Lemma 11 (the contraction property), we get

$$(a) = \mathbb{E}|\overline{X}_{t_{k+1}}^{t_k,\overline{X}_k} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k}|^2 \le e_k^2 \exp(-2\beta\delta). \tag{.27}$$

By (.20) (in Lemma 12), we have

$$(b) \le \left(C_1 + C_2\mathbb{E}|\widehat{X}_k|^2\right)\delta^3. \tag{.28}$$

Step 3. We analyze the cross-product (c). By splitting

$$\overline{X}_{k+1} - \overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} = (\overline{X}_k - \widehat{X}_k) + \underbrace{\left[(\overline{X}_{k+1} - \overline{X}_k) - (\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_k)\right]}_{:=d_\delta(\overline{X}_k,\widehat{X}_k)},$$

we obtain

$$(c) = \underbrace{\mathbb{E}\left[(\overline{X}_k - \widehat{X}_k)(\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1})\right]}_{(d)} + \underbrace{\mathbb{E}\left[d_\delta(\overline{X}_k,\widehat{X}_k)(\overline{X}_{t_{k+1}}^{t_k,\widehat{X}_k} - \widehat{X}_{k+1})\right]}_{(e)}. \tag{.29}$$

For the term (d), we have

$$\begin{aligned}(d) &= \mathbb{E}\left[(\overline{X}_k - \widehat{X}_k)\,\mathbb{E}(\overline{X}^{t_k,\widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1}|\mathcal{F}_k)\right] \\ &\le e_k\left(\mathbb{E}|\mathbb{E}(\overline{X}^{t_k,\widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1}|\mathcal{F}_k)|^2\right)^{\frac{1}{2}} \\ &\le e_k\left(C_1 + C_2\sqrt{\mathbb{E}|\widehat{X}_k|^2}\right)\delta^{\frac{3}{2}},\end{aligned} \tag{.30}$$

where we use the tower property (of the conditional expectation) in the first equation, the Cauchy-Schwarz inequality in the second inequality, and (.21) in the final inequality. According to [43, Lemma 1.3], there exists $C_0 > 0$ (independent of $\delta, \widehat{X}_k$) such that

$$\left(\mathbb{E}d^2_\delta(\overline{X}_k, \widehat{X}_k)\right)^{\frac{1}{2}} \le C_0 e_k\sqrt{\delta}. \tag{.31}$$

Thus,

$$\begin{aligned}(e) &\le \left(\mathbb{E}d^2_\delta(\overline{X}_k, \widehat{X}_k)\right)^{\frac{1}{2}}\left(\mathbb{E}|\overline{X}^{t_k,\widehat{X}_k}_{t_{k+1}} - \widehat{X}_{k+1}|^2\right)^{\frac{1}{2}} \\ &\le C_0 e_k\left(C_1 + C_2\sqrt{\mathbb{E}|\widehat{X}_k|^2}\right)\delta^2.\end{aligned} \tag{.32}$$

where we use (.20) and (.31) in the last inequality. Combining (.29), (.30) and (.32) yields for δ sufficiently small,

$$(c) \le e_k\left(C'_1 + C'_2\sqrt{\mathbb{E}|\widehat{X}_k|^2}\right)\delta^{\frac{3}{2}}, \quad \text{for some } C'_1, C'_2 > 0 \text{ (independent of } \delta, \widehat{X}_k). \tag{.33}$$

Step 4. Combining (.26) with (.27), (.28) and (.33) yields

$$e^2_{k+1} \le e^2_k\exp(-2\beta\delta) + \left(C_1 + C_2\mathbb{E}|\widehat{X}_k|^2\right)\delta^3 + e_k\left(C'_1 + C'_2\sqrt{\mathbb{E}|\widehat{X}_k|^2}\right)\delta^{\frac{3}{2}}.$$

A standard argument shows that Lemma 11 (the contraction property) implies $\mathbb{E}|\overline{X}_t|^2 \le C$ for some $C > 0$. Thus, $\mathbb{E}|\widehat{X}_k|^2 \le 2(C + e^2_k)$. As a result, for δ sufficiently small,

$$e^2_{k+1} \le e^2_k\left(1 - \frac{3}{4}\beta\delta\right) + D_1\delta^3 + D_2 e^2_k\left(\delta^3 + \delta^{\frac{3}{2}}\right) + D_3 e_k\delta^{\frac{3}{2}}, \tag{.34}$$

for some $D_1, D_2, D_3 > 0$ (independent of δ). Note that

$$D_2 e^2_k\left(\delta^3 + \delta^{\frac{3}{2}}\right) \le \frac{1}{4}e^2_k\beta\delta, \quad \text{for } \delta \text{ sufficiently small,}$$

and

$$D_3 e_k\delta^{\frac{3}{2}} \le \frac{1}{4}e^2_k\beta\delta + \frac{2D_3^2}{\beta}\delta^2.$$

Thus, the estimate (.34) leads to

$$e^2_{k+1} \le e^2_k\left(1 - \frac{1}{4}\beta\delta\right) + D\delta^2, \quad \text{for some } D > 0 \text{ (independent of } \delta). \tag{.35}$$

Unfolding the inequality (.35) yields the estimate (3.12). □

As a remark, if we can improve the estimate in (.21) to

$$|\mathbb{E}(\overline{X}^{t_\star,x}_{t_\star+\delta} - \widehat{X}^{t_\star,x}_1)| \le (C_1 + C_2\sqrt{\mathbb{E}|x|^2})^{\frac{1}{2}}\delta^2, \tag{.36}$$

(i.e. δ^2 local error instead of $\delta^{\frac{3}{2}}$), then the discretization error is $C\delta$.

G. Experimental Details. We provide implementation details of the experiments in Section 5. All codes will soon be available at github.

G.1. 1-dimensional data experiments. For this 1-dimensional experiments, we start from a single point mass, and the parameters of the DPMs (OU and COU) are chosen as $\theta \equiv 0.2$, $\sigma(t) \equiv 0.5$.

G.2. Swiss Roll and MNIST experiments. In both experiments, we use contractive subVP with predictor-corrector sampler. We recommend to use $\beta_{min} = 0.01$ and $\beta_{max} = 8$ for a good result. The signal-noise-ratio is set to be 0.2 for Swiss Roll, and 0.1 for MNIST. The Jupyter Notebooks can be found in the zip file of supplementary materials.

G.3. CIFAR10 experiments. We use contractive subVP with predictor-corrector sampler. with all the other settings the same as VE SDE in [60]. We recommend to use $\beta_{min} = 0.01$ and $\beta_{max} = 8$ for a good result. The signal-noise-ratio is set to be 0.11 for the best result. The experiments based on NCSN++ [60] are as below in Table 5.

For EDM [30] with contraction, we adopt $f(0) = 0.98$ to yield the best result by adding contraction, i.e. $\epsilon = 0.02$. We find that $f(0) = 0.97$ or 0.99 also leads to better results than original EDM. Our denoising sampler is based on the deterministic sampler (2^{nd} order scheme) of EDM paper and yields the same settings of sampling steps or other sampler constants. For computational resource, we used 4 L40S GPUs, and these results could readily be realized within 0.5 hours for each separate task.

Parameter	Value
β_{min}	0.01
β_{max}	8
snr	0.11
grad_{clip}	10
α_{lr}	$5e^{-4}$

Table 5. Parameters of the CDPM for CIFAR-10.

G.4 More examples of CIFAR10 synthesis by CsubVP. Here we provide more examples of CIFAR10 32×32 synthesis by CsubVP SDEs in Figure 5.

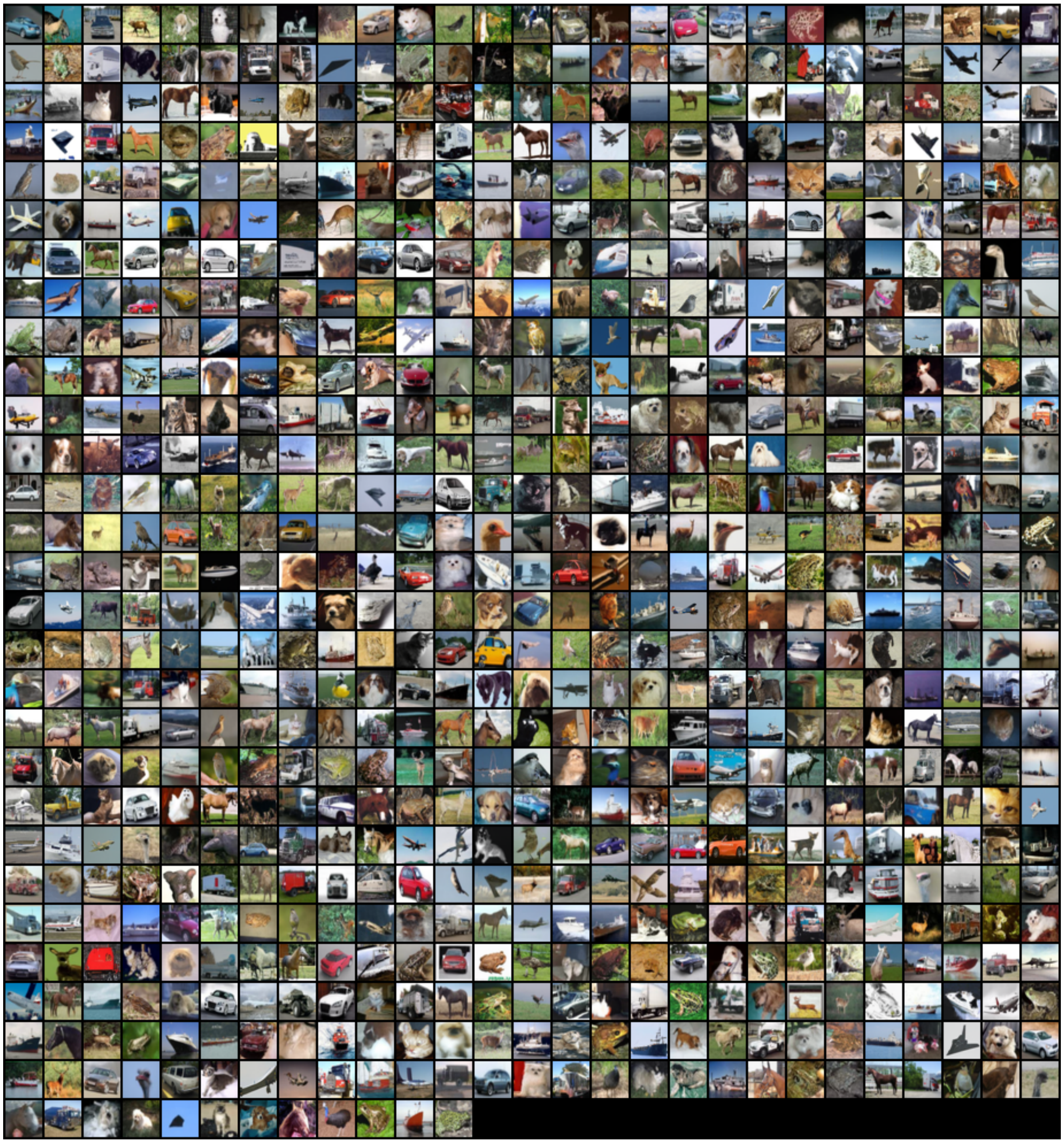


Figure 5. CsubVP CIFAR10 samples