Appendix

A Proofs of Main Theoretical Results

In this section, we provide proofs of our main results. We define below some crucial notations which we will use throughout. We use ODE$(\dots)$ to denote the backwards ODE under exact score $\nabla \log p_t(x)$. More specifically, given any $x \in \mathbb{R}^d$ and $s > r > 0$, let x_t denote the solution to the following ODE:

$$dx_t = -t\nabla \log p_t(x_t)dt. \tag{5}$$

ODE$(x, s \to r)$ is defined as "the value of x_r when initialized at $x_s = x$". It will also be useful to consider a "time-discretized ODE with drift $ts_\theta(x,t)$": let δ denote the discretization step size and let k denote any integer. Let δ denote a step size, let $\overline{x}_t$ denote the solution to

$$d\overline{x}_t = -ts_\theta(x_{k\delta}, k\delta)dt, \tag{6}$$

where for any t, k is the unique integer such that $t \in ((k-1)\delta, k\delta]$. We verify that the dynamics of Eq. (6) is equivalent to the following discrete-time dynamics for $t = k\delta, k \in \mathbb{Z}$:

$$\overline{x}_{(k-1)\delta} = \overline{x}_{k\delta} - \frac{1}{2}\left(((k-1)\delta)^2 - (k\delta)^2\right) s_\theta(x_{k\delta}, k\delta).$$

We similarly denote the value of $\overline{x}_r$ when initialized at $\overline{x}_s = x$ as ODE$_\theta(x, s \to r)$. Analogously, we let SDE$(x, s \to r)$ and SDE$_\theta(x, s \to r)$ denote solutions to

$$dy_t = -2t\nabla \log p_t(y_t)dt + \sqrt{2t}dB_t$$
$$d\overline{y}_t = -2ts_\theta(\overline{y}_t, t)dt + \sqrt{2t}dB_t$$

respectively. Finally, we will define the Restart$_\theta$ process as follows:

$$\begin{aligned} &\text{(Restart}_\theta\text{ forward process)} \qquad x^{i+1}_{t_{\text{max}}} = x^i_{t_{\text{min}}} + \varepsilon^i_{t_{\text{min}} \to t_{\text{max}}} \\ &\text{(Restart}_\theta\text{ backward process)} \qquad x^{i+1}_{t_{\text{min}}} = \text{ODE}_\theta(x^{i+1}_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}}), \end{aligned} \tag{7}$$

where $\varepsilon^i_{t_{\text{min}} \to t_{\text{max}}} \sim \mathcal{N}\left(\mathbf{0}, \left(t^2_{\text{max}} - t^2_{\text{min}}\right)\boldsymbol{I}\right)$. We use Restart$_\theta(x, K)$ to denote $x^K_{t_{\text{min}}}$ in the above processes, initialized at $x^0_{t_{\text{min}}} = x$. In various theorems, we will refer to a function $Q(r) : \mathbb{R}^+ \to [0, 1/2)$, defined as the Gaussian tail probability $Q(r) = Pr(a \geq r)$ for $a \sim \mathcal{N}(0,1)$.

A.1 Main Result

Theorem 3. *[Formal version of Theorem 1] Let t_{max} be the initial noise level. Let the initial random variables $\overline{x}_{t_{max}} = \overline{y}_{t_{max}}$, and*

$$\overline{x}_{t_{min}} = ODE_\theta(\overline{x}_{t_{max}}, t_{max} \to t_{min})$$
$$\overline{y}_{t_{min}} = SDE_\theta(\overline{y}_{t_{max}}, t_{max} \to t_{min}),$$

Let p_t denote the true population distribution at noise level t. Let $p^{ODE_\theta}_t, p^{SDE_\theta}_t$ denote the distributions for x_t, y_t respectively. Assume that for all x, y, s, t, $s_\theta(x,t)$ satisfies $\|ts_\theta(x,t) - ts_\theta(x,s)\| \leq L_0|s-t|$, $\|ts_\theta(x,t)\| \leq L_1$, $\|ts_\theta(x,t) - ts_\theta(y,t)\| \leq L_2\|x-y\|$, and the approximation error $\|ts_\theta(x,t) - t\nabla \log p_t(x)\| \leq \epsilon_{approx}$. Assume in addition that $\forall t \in [t_{min}, t_{max}]$, $\|x_t\| < B/2$ for any x_t in the support of p_t, $p^{ODE_\theta}_t$ or $p^{SDE_\theta}_t$, and $K \leq \frac{C}{L_2(t_{max} - t_{min})}$ for some universal constant C. Then

$$\begin{aligned} W_1(p^{ODE_\theta}_{t_{min}}, p_{t_{min}}) &\leq B \cdot TV\left(p^{ODE_\theta}_{t_{max}}, p_{t_{max}}\right) \\ &\quad + e^{L_2(t_{max} - t_{min})} \cdot (\delta(L_2L_1 + L_0) + \epsilon_{approx})(t_{max} - t_{min}) \end{aligned} \tag{8}$$

$$\begin{aligned} W_1(p^{SDE_\theta}_{t_{min}}, p_{t_{min}}) &\leq B \cdot \left(1 - \lambda e^{-BL_1/t_{min} - L_1^2 t^2_{max}/t^2_{min}}\right) TV(p^{SDE_\theta}_{t_{max}}, p_{t_{max}}) \\ &\quad + e^{2L_2(t_{max} - t_{min})}\left(\epsilon_{approx} + \delta L_0 + L_2\left(\delta L_1 + \sqrt{2\delta dt_{max}}\right)\right)(t_{max} - t_{min}) \end{aligned} \tag{9}$$

where $\lambda := 2Q\left(\frac{B}{2\sqrt{t^2_{max} - t^2_{min}}}\right)$.

Proof. Let us define $x_{t_{\text{max}}} \sim p_{t_{\text{max}}}$, and let $x_{t_{\text{min}}} = \text{ODE}(x_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}})$. We verify that $x_{t_{\text{min}}}$ has density $p_{t_{\text{min}}}$. Let us also define $\hat{x}_{t_{\text{min}}} = \text{ODE}_\theta(x_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}})$. We would like to bound the Wasserstein distance between $\bar{x}_{t_{\text{min}}}$ and $x_{t_{\text{min}}}$ (*i.e.*, $p_{t_{\text{min}}}^{\text{ODE}_\theta}$ and $p_{t_{\text{min}}}$), by the following triangular inequality:

$$W_1(\bar{x}_{t_{\text{min}}}, x_{t_{\text{min}}}) \le W_1(\bar{x}_{t_{\text{min}}}, \hat{x}_{t_{\text{min}}}) + W_1(\hat{x}_{t_{\text{min}}}, x_{t_{\text{min}}}) \tag{10}$$

By Lemma 2, we know that

$$\|\hat{x}_{t_{\text{min}}} - x_{t_{\text{min}}}\| \le e^{(t_{\text{max}} - t_{\text{min}})L_2} \left(\delta(L_2 L_1 + L_0) + \epsilon_{approx}\right)(t_{\text{max}} - t_{\text{min}}),$$

where we use the fact that $\|\hat{x}_{t_{\text{max}}} - x_{t_{\text{max}}}\| = 0$. Thus we immediately have

$$W_1(\hat{x}_{t_{\text{min}}}, x_{t_{\text{min}}}) \le e^{(t_{\text{max}} - t_{\text{min}})L_2} \left(\delta(L_2 L_1 + L_0) + \epsilon_{approx}\right)(t_{\text{max}} - t_{\text{min}}) \tag{11}$$

On the other hand,

$$\begin{aligned} W_1(\hat{x}_{t_{\text{min}}}, \overline{x}_{t_{\text{min}}}) \le & B \cdot TV\left(\hat{x}_{t_{\text{min}}}, \overline{x}_{t_{\text{min}}}\right) \\ \le & B \cdot TV\left(\hat{x}_{t_{\text{max}}}, \overline{x}_{t_{\text{max}}}\right) \end{aligned} \tag{12}$$

where the last equality is due to the data-processing inequality. Combining Eq. (11) , Eq. (12) and the triangular inequality Eq. (10), we arrive at the upper bound for ODE (Eq. (8)). The upper bound for SDE (Eq. (9)) shares a similar proof approach. First, let $y_{t_{\text{max}}} \sim p_{t_{\text{max}}}$. Let $\hat{y}_{t_{\text{min}}} = \text{SDE}_\theta(y_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}})$. By Lemma 5,

$$TV\left(\hat{y}_{t_{\text{min}}}, \overline{y}_{t_{\text{min}}}\right) \le \left(1 - 2Q\left(\frac{B}{2\sqrt{t_{\text{max}}^2 - t_{\text{min}}^2}}\right) \cdot e^{-BL_1/t_{\text{min}} - L_1^2 t_{\text{max}}^2 / t_{\text{min}}^2}\right) \cdot TV\left(\hat{y}_{t_{\text{max}}}, \overline{y}_{t_{\text{max}}}\right)$$

On the other hand, by Lemma 4,

$$\mathbb{E}\left[\|\hat{y}_{t_{\text{min}}} - y_{t_{\text{min}}}\|\right] \le e^{2L_2(t_{\text{max}} - t_{\text{min}})} \left(\epsilon_{approx} + \delta L_0 + L_2\left(\delta L_1 + \sqrt{2\delta d t_{\text{max}}}\right)\right)(t_{\text{max}} - t_{\text{min}}).$$

The SDE triangular upper bound on $W_1(\bar{y}_{t_{\text{min}}}, y_{t_{\text{min}}})$ follows by multiplying the first inequality by B (to bound $W_1(\bar{y}_{t_{\text{min}}}, \hat{y}_{t_{\text{min}}})$) and then adding the second inequality (to bound $W_1(y_{t_{\text{min}}}, \hat{y}_{t_{\text{min}}})$). Notice that by definition, $TV\left(\hat{y}_{t_{\text{max}}}, \overline{y}_{t_{\text{max}}}\right) = TV\left(y_{t_{\text{max}}}, \overline{y}_{t_{\text{max}}}\right)$. Finally, because of the assumption that $K \le \frac{C}{L_2(t_{\text{max}} - t_{\text{min}})}$ for some universal constant, we summarize the second term in the Eq. (8) and Eq. (9) into the big O in the informal version Theorem 1. □

Theorem 4. *[Formal version of Theorem 2] Consider the same setting as Theorem 3. Let $p_{t_{min}}^{Restart_\theta, i}$ denote the distributions after i^{th} Restart iteration,* i.e., *the distribution of $\overline{x}_{t_{min}}^i = Restart_\theta(\overline{x}_{t_{min}}^0, i)$. Given initial $\overline{x}_{t_{max}}^0 \sim p_{t_{max}}^{Restart,0}$, let $\overline{x}_{t_{min}}^0 = ODE_\theta(\overline{x}_{t_{max}}^0, t_{max} \to t_{min})$. Then*

$$\begin{aligned} W_1(p_{t_{min}}^{Restart_\theta, K}, p_{t_{min}}) \le & \underbrace{B \cdot (1-\lambda)^K \, TV(p_{t_{max}}^{Restart,0}, p_{t_{max}})}_{\text{upper bound on contracted error}} \\ & + \underbrace{e^{(K+1)L_2(t_{max} - t_{min})}(K+1)\left(\delta(L_2 L_1 + L_0) + \epsilon_{approx}\right)(t_{max} - t_{min})}_{\text{upper bound on additional sampling error}} \end{aligned} \tag{13}$$

where $\lambda = 2Q\left(\frac{B}{2\sqrt{t_{max}^2 - t_{min}^2}}\right)$.

Proof. Let $x_{t_{\text{max}}}^0 \sim p_{t_{\text{max}}}$. Let $x_{t_{\text{min}}}^K = \text{Restart}(x_{t_{\text{min}}}^0, K)$. We verify that $x_{t_{\text{min}}}^K$ has density $p_{t_{\text{min}}}$. Let us also define $\hat{x}_{t_{\text{min}}}^0 = \text{ODE}_\theta(x_{t_{\text{max}}}^0, t_{\text{max}} \to t_{\text{min}})$ and $\hat{x}_{t_{\text{min}}}^K = \text{Restart}_\theta(\hat{x}_{t_{\text{min}}}^0, K)$.

By Lemma 1,

$$\begin{aligned} TV\left(\overline{x}_{t_{\text{min}}}^K, \hat{x}_{t_{\text{min}}}^K\right) &\le \left(1 - 2Q\left(\frac{B}{2\sqrt{t_{\text{max}}^2 - t_{\text{min}}^2}}\right)\right)^K TV\left(\overline{x}_{t_{\text{min}}}^0, \hat{x}_{t_{\text{min}}}^0\right) \\ &\le \left(1 - 2Q\left(\frac{B}{2\sqrt{t_{\text{max}}^2 - t_{\text{min}}^2}}\right)\right)^K TV\left(\overline{x}_{t_{\text{max}}}^0, \hat{x}_{t_{\text{max}}}^0\right) \\ &= \left(1 - 2Q\left(\frac{B}{2\sqrt{t_{\text{max}}^2 - t_{\text{min}}^2}}\right)\right)^K TV\left(\overline{x}_{t_{\text{max}}}^0, x_{t_{\text{max}}}^0\right) \end{aligned}$$

The second inequality holds by data processing inequality. The above can be used to bound the 1-Wasserstein distance as follows:

$$W_1\left(\bar{x}^K_{t_{\min}}, \hat{x}^K_{t_{\min}}\right) \leq B \cdot TV\left(\bar{x}^K_{t_{\min}}, \hat{x}^K_{t_{\min}}\right) \leq \left(1 - 2Q\left(\frac{B}{2\sqrt{t^2_{\max} - t^2_{\min}}}\right)\right)^K TV\left(\bar{x}^0_{t_{\max}}, x^0_{t_{\max}}\right) \tag{14}$$

On the other hand, using Lemma 3,

$$\begin{aligned} W_1\left(x^K_{t_{\min}}, \hat{x}^K_{t_{\min}}\right) \leq & \left\| x^K_{t_{\min}} - \hat{x}^K_{t_{\min}} \right\| \\ \leq & e^{(K+1)L_2(t_{\max} - t_{\min})}(K+1)\left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{\max} - t_{\min}) \end{aligned} \tag{15}$$

We arrive at the result by combining the two bounds above (Eq. (14), Eq. (15)) with the following triangular inequality,

$$W_1(\bar{x}^K_{t_{\min}}, x^K_{t_{\min}}) \leq W_1(\bar{x}^K_{t_{\min}}, \hat{x}^K_{t_{\min}}) + W_1(\hat{x}^K_{t_{\min}}, x^K_{t_{\min}})$$

□

A.2 Mixing under Restart with exact ODE

Lemma 1. *Consider the same setup as Theorem 4. Consider the Restart$_\theta$ process defined in equation 7. Let*

$$\begin{aligned} x^i_{t_{min}} &= \textit{Restart}_\theta(x^0_{t_{min}}, i) \\ y^i_{t_{min}} &= \textit{Restart}_\theta(y^0_{t_{min}}, i). \end{aligned}$$

Let $p_t^{\textit{Restart}_\theta(i)}$ and $q_t^{\textit{Restart}_\theta(i)}$ denote the densities of x^i_t and y^i_t respectively. Then

$$TV\left(p^{\textit{Restart}_\theta(K)}_{t_{min}}, q^{\textit{Restart}_\theta(K)}_{t_{min}}\right) \leq (1-\lambda)^K TV\left(p^{\textit{Restart}_\theta(0)}_{t_{min}}, q^{\textit{Restart}_\theta(0)}_{t_{min}}\right),$$

where $\lambda = 2Q\left(\frac{B}{2\sqrt{t^2_{max} - t^2_{min}}}\right)$.

Proof. Conditioned on $x^i_{t_{\min}}, y^i_{t_{\min}}$, let $x^{i+1}_{t_{\max}} = x^i_{t_{\min}} + \sqrt{t^2_{\max} - t^2_{\min}}\xi^x_i$ and $y^{i+1}_{t_{\max}} = y^i_{t_{\min}} + \sqrt{t^2_{\max} - t^2_{\min}}\xi^y_i$. We now define a coupling between $x^{i+1}_{t_{\min}}$ and $y^{i+1}_{t_{\min}}$ by specifying the joint distribution over ξ^x_i and ξ^y_i.

If $x^i_{t_{\min}} = y^i_{t_{\min}}$, let $\xi^x_i = \xi^y_i$, so that $x^{i+1}_{t_{\min}} = y^{i+1}_{t_{\min}}$. On the other hand, if $x^i_{t_{\min}} \neq y^i_{t_{\min}}$, let $x^{i+1}_{t_{\max}}$ and $y^{i+1}_{t_{\max}}$ be coupled as described in the proof of Lemma 7, with $x' = x^{i+1}_{t_{\max}}, y' = y^{i+1}_{t_{\max}}, \sigma = \sqrt{t^2_{\max} - t^2_{\min}}$. Under this coupling, we verify that,

$$\begin{aligned} & \mathbb{E}\left[\mathbb{1}\left\{x^{i+1}_{t_{\min}} \neq y^{i+1}_{t_{\min}}\right\}\right] \\ \leq & \mathbb{E}\left[\mathbb{1}\left\{x^{i+1}_{t_{\max}} \neq y^{i+1}_{t_{\max}}\right\}\right] \\ \leq & \mathbb{E}\left[\left(1 - 2Q\left(\frac{\left\|x^i_{t_{\min}} - y^i_{t_{\min}}\right\|}{2\sqrt{t^2_{\max} - t^2_{\min}}}\right)\right)\mathbb{1}\left\{x^i_{t_{\min}} \neq y^i_{t_{\min}}\right\}\right] \\ \leq & \left(1 - 2Q\left(\frac{B}{2\sqrt{t^2_{\max} - t^2_{\min}}}\right)\right)\mathbb{E}\left[\mathbb{1}\left\{x^i_{t_{\min}} \neq y^i_{t_{\min}}\right\}\right]. \end{aligned}$$

Applying the above recursively,

$$\mathbb{E}\left[\mathbb{1}\left\{x^K_{t_{\min}} \neq y^K_{t_{\min}}\right\}\right] \leq \left(1 - 2Q\left(\frac{B}{2\sqrt{t^2_{\max} - t^2_{\min}}}\right)\right)^K \mathbb{E}\left[\mathbb{1}\left\{x^0_{t_{\min}} \neq y^0_{t_{\min}}\right\}\right].$$

The conclusion follows by noticing that $TV\left(p^{\text{Restart}_\theta(K)}_{t_{\min}}, q^{\text{Restart}_\theta(K)}_t\right) \leq Pr\left(x^K_{t_{\min}} \neq y^K_{t_{\min}}\right) = \mathbb{E}\left[\mathbb{1}\left\{x^K_{t_{\min}} \neq y^K_{t_{\min}}\right\}\right]$, and by selecting the initial coupling so that $Pr\left(x^0_{t_{\min}} \neq y^0_{t_{\min}}\right) = TV\left(p^{\text{Restart}_\theta(0)}_{t_{\min}}, q^{\text{Restart}_\theta(0)}_{t_{\min}}\right)$. □

A.3 W_1 discretization bound

Lemma 2 (Discretization bound for ODE)**.** *Let $x_{t_{min}} = ODE\left(x_{t_{max}}, t_{max} \to t_{min}\right)$ and let $\overline{x}_{t_{min}} = ODE_\theta\left(\overline{x}_{t_{max}}, t_{max} \to t_{min}\right)$. Assume that for all x, y, s, t, $s_\theta(x,t)$ satisfies $\|ts_\theta(x,t) - ts_\theta(x,s)\| \leq L_0|s-t|$, $\|ts_\theta(x,t)\| \leq L_1$ and $\|ts_\theta(x,t) - ts_\theta(y,t)\| \leq L_2\|x-y\|$. Then*

$$\|x_{t_{min}} - \overline{x}_{t_{min}}\| \leq e^{(t_{max}-t_{min})L_2}\left(\|x_{t_{max}} - \overline{x}_{t_{max}}\| + \left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{max} - t_{min})\right)$$

Proof. Consider some fixed arbitrary k, and recall that δ is the step size. Recall that by definition of ODE and ODE_θ, for $t \in ((k-1)\delta, k\delta]$,

$$\begin{aligned} dx_t &= -t\nabla \log p_t(x_t)dt \\ d\overline{x}_t &= -ts_\theta(\overline{x}_{k\delta}, k\delta)dt. \end{aligned}$$

For $t \in [t_{\min}, t_{\max}]$, let us define a time-reversed process $x_t^{\leftarrow} := x_{-t}$. Let $v(x,t) := \nabla \log p_{-t}(x)$. Then for $t \in [-t_{\max}, -t_{\min}]$

$$dx_t^{\leftarrow} = tv(x_t^{\leftarrow}, t)ds.$$

Similarly, define $\overline{x}_t^{\leftarrow} := \overline{x}_{-t}$ and $\overline{v}(x,t) := s_\theta(x, -t)$. It follows that

$$d\overline{x}_t^{\leftarrow} = t\overline{v}(\overline{x}_{k\delta}^{\leftarrow}, k\delta)ds,$$

where k is the unique (negative) integer satisfying $t \in [k\delta, (k+1)\delta)$. Following these definitions,

$$\begin{aligned} &\frac{d}{dt}\|x_t^{\leftarrow} - \overline{x}_t^{\leftarrow}\| \\ \leq &\|tv(x_t^{\leftarrow}, t) - t\overline{v}(x_t^{\leftarrow}, t)\| \\ &\quad + \|t\overline{v}(x_t^{\leftarrow}, t) - t\overline{v}(\overline{x}_t^{\leftarrow}, t)\| \\ &\quad + \|t\overline{v}(\overline{x}_t^{\leftarrow}, t) - t\overline{v}(\overline{x}_t^{\leftarrow}, k\delta)\| \\ &\quad + \|t\overline{v}(\overline{x}_t^{\leftarrow}, k\delta) - t\overline{v}(\overline{x}_{k\delta}^{\leftarrow}, k\delta)\| \\ \leq &\epsilon_{approx} + L_2\|x_t^{\leftarrow} - \overline{x}_t^{\leftarrow}\| + \delta L_0 + L_2\|\overline{x}_t^{\leftarrow} - \overline{x}_{k\delta}^{\leftarrow}\| \\ \leq &\epsilon_{approx} + L_2\|x_t^{\leftarrow} - \overline{x}_t^{\leftarrow}\| + \delta L_0 + \delta L_2 L_1. \end{aligned}$$

Applying Gronwall's Lemma over the interval $t \in [-t_{\max}, -t_{\min}]$,

$$\begin{aligned} &\|x_{t_{\min}} - \overline{x}_{t_{\min}}\| \\ = &\left\|x_{-t_{\min}}^{\leftarrow} - \overline{x}_{-t_{\min}}^{\leftarrow}\right\| \\ \leq &e^{L_2(t_{\max}-t_{\min})}\left(\left\|x_{-t_{\max}}^{\leftarrow} - \overline{x}_{-t_{\max}}^{\leftarrow}\right\| + \left(\epsilon_{approx} + \delta L_0 + \delta L_2 L_1\right)(t_{\max} - t_{\min})\right) \\ = &e^{L_2(t_{\max}-t_{\min})}\left(\|x_{t_{\max}} - \overline{x}_{t_{\max}}\| + \left(\epsilon_{approx} + \delta L_0 + \delta L_2 L_1\right)(t_{\max} - t_{\min})\right). \end{aligned}$$

□

Lemma 3. *Given initial $x^0_{t_{max}}$, let $x^0_{t_{min}} = ODE\left(x^0_{t_{max}}, t_{max} \to t_{min}\right)$, and let $\hat{x}^0_{t_{min}} = ODE_\theta\left(x^0_{t_{max}}, t_{max} \to t_{min}\right)$. We further denote the variables after K Restart iterations as $x^K_{t_{min}} = Restart(x^0_{t_{min}}, K)$ and $\hat{x}^K_{t_{min}} = Restart_\theta(\hat{x}^0_{t_{min}}, K)$, with true field and learned field respectively. Then there exists a coupling between $x^K_{t_{min}}$ and $\hat{x}^K_{t_{min}}$ such that*

$$\left\|x^K_{t_{min}} - \hat{x}^K_{t_{min}}\right\| \leq e^{(K+1)L_2(t_{max}-t_{min})}(K+1)\left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{max} - t_{min}).$$

Proof. We will couple $x^i_{t_{\min}}$ and $\hat{x}^i_{t_{\min}}$ by using the same noise $\varepsilon^i_{t_{\min}\to t_{\max}}$ in the Restart forward process for $i = 0 \dots K-1$ (see Eq. (7)). For any i, let us also define $y^{i,j}_{t_{\min}} := \text{Restart}_\theta\left(x^i_{t_{\min}}, j - i\right)$, and this process uses the same noise $\varepsilon^i_{t_{\min}\to t_{\max}}$ as previous ones. From this definition, $y^{K,K}_{t_{\min}} = x^K_{t_{\min}}$. We can thus bound

$$\left\|x^K_{t_{\min}}, \hat{x}^K_{t_{\min}}\right\| \leq \left\|y^{0,K}_{t_{\min}} - \hat{x}^K_{t_{\min}}\right\| + \sum_{i=0}^{K-1}\left\|y^{i,K}_{t_{\min}} - y^{i+1,K}_{t_{\min}}\right\| \tag{16}$$

Using the assumption that $ts_\theta(\cdot, t)$ is L_2 Lipschitz,

$$\begin{aligned}
&\left\|y^{0,i+1}_{t_{\text{min}}} - \hat{x}^{i+1}_{t_{\text{min}}}\right\| \\
=&\left\|\text{ODE}_\theta(y^{0,i}_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}}) - \text{ODE}_\theta(\hat{x}^{i}_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}})\right\| \\
\leq& e^{L_2(t_{\text{max}} - t_{\text{min}})}\left\|y^{0,i}_{t_{\text{max}}} - \hat{x}^{i}_{t_{\text{max}}}\right\| \\
=& e^{L_2(t_{\text{max}} - t_{\text{min}})}\left\|y^{0,i}_{t_{\text{min}}} - \hat{x}^{i}_{t_{\text{min}}}\right\|,
\end{aligned}$$

where the last equality is because we add the same additive Gaussian noise $\varepsilon^i_{t_{\text{min}} \to t_{\text{max}}}$ to $y^{0,i}_{t_{\text{min}}}$ and $\hat{x}^i_{t_{\text{min}}}$ in the Restart forward process. Applying the above recursively, we get

$$\begin{aligned}
\left\|y^{0,K}_{t_{\text{min}}} - \hat{x}^{K}_{t_{\text{min}}}\right\| \leq& e^{KL_2(t_{\text{max}} - t_{\text{min}})}\left\|y^{0,0}_{t_{\text{min}}} - \hat{x}^{0}_{t_{\text{min}}}\right\| \\
\leq& e^{KL_2(t_{\text{max}} - t_{\text{min}})}\left\|x^{0}_{t_{\text{min}}} - \hat{x}^{0}_{t_{\text{min}}}\right\| \\
\leq& e^{(K+1)L_2(t_{\text{max}} - t_{\text{min}})}\left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{\text{max}} - t_{\text{min}}), \qquad (17)
\end{aligned}$$

where the last line follows by Lemma 2 when setting $x_{t_{\text{max}}} = \bar{x}_{t_{\text{max}}}$. We will now bound $\left\|y^{i,K}_{t_{\text{min}}} - y^{i+1,K}_{t_{\text{min}}}\right\|$ for some $i \leq K$. It follows from definition that

$$\begin{aligned}
y^{i,i+1}_{t_{\text{min}}} &= \text{ODE}_\theta\left(x^i_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}}\right) \\
y^{i+1,i+1}_{t_{\text{min}}} &= x^{i+1}_{t_{\text{min}}} = \text{ODE}\left(x^i_{t_{\text{max}}}, t_{\text{max}} \to t_{\text{min}}\right).
\end{aligned}$$

By Lemma 2,

$$\left\|y^{i,i+1}_{t_{\text{min}}} - y^{i+1,i+1}_{t_{\text{min}}}\right\| \leq e^{L_2(t_{\text{max}} - t_{\text{min}})}\left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{\text{max}} - t_{\text{min}})$$

For the remaining steps from $i+2 \ldots K$, both $y^{i,\cdot}$ and $y^{i+1,\cdot}$ evolve with ODE_θ in each step. Again using the assumption that $ts_\theta(\cdot, t)$ is L_2 Lipschitz,

$$\left\|y^{i,K}_{t_{\text{min}}} - y^{i+1,K}_{t_{\text{min}}}\right\| \leq e^{(K-i)L_2(t_{\text{max}} - t_{\text{min}})}\left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{\text{max}} - t_{\text{min}})$$

Summing the above for $i = 0 \ldots K-1$, and combining with Eq. (16) and Eq. (17) gives

$$\left\|x^K_{t_{\text{min}}} - \hat{x}^K_{t_{\text{min}}}\right\| \leq e^{(K+1)L_2(t_{\text{max}} - t_{\text{min}})}(K+1)\left(\delta(L_2L_1 + L_0) + \epsilon_{approx}\right)(t_{\text{max}} - t_{\text{min}}).$$

□

Lemma 4. *Consider the same setup as Theorem 3. Let $x_{t_{min}} = SDE\left(x_{t_{max}}, t_{max} \to t_{min}\right)$ and let $\overline{x}_{t_{min}} = SDE\left(\overline{x}_{t_{max}}, t_{max} \to t_{min}\right)$. Then there exists a coupling between x_t and $\overline{x}_t$ such that*

$$\begin{aligned}
\mathbb{E}\left[\|x_{t_{min}} - \overline{x}_{t_{min}}\|\right] \leq& e^{2L_2(t_{max} - t_{min})}\mathbb{E}\left[\|x_{t_{max}} - \overline{x}_{t_{max}}\|\right] \\
&+ e^{2L_2(t_{max} - t_{min})}\left(\epsilon_{approx} + \delta L_0 + L_2\left(\delta L_1 + \sqrt{2\delta dt_{max}}\right)\right)(t_{max} - t_{min})
\end{aligned}$$

Proof. Consider some fixed arbitrary k, and recall that δ is the stepsize. By definition of SDE and SDE_θ, for $t \in ((k-1)\delta, k\delta]$,

$$\begin{aligned}
dx_t &= -2t\nabla \log p_t(x_t)dt + \sqrt{2t}dB_t \\
d\overline{x}_t &= -2ts_\theta(\overline{x}_{k\delta}, k\delta)dt + \sqrt{2t}dB_t.
\end{aligned}$$

Let us define a coupling between x_t and $\overline{x}_t$ by identifying their respective Brownian motions. It will be convenient to define the time-reversed processes $x^{\leftarrow}_t := x_{-t}$, and $\overline{x}^{\leftarrow}_t := \overline{x}_{-t}$, along with $v(x,t) := \nabla \log p_{-t}(x)$ and $\overline{v}(x,t) := s_\theta(x, -t)$. Then there exists a Brownian motion $B^{\leftarrow}_t$, such that for $t \in [-t_{\text{max}}, -t_{\text{min}}]$,

$$\begin{aligned}
dx^{\leftarrow}_t &= -2tv(x^{\leftarrow}_t, t)dt + \sqrt{-2t}dB^{\leftarrow}_t \\
d\overline{x}^{\leftarrow}_t &= -2t\overline{v}(\overline{x}^{\leftarrow}_{k\delta}, k\delta)dt + \sqrt{-2t}dB^{\leftarrow}_t \\
\Rightarrow \quad d(x^{\leftarrow}_t - \overline{x}^{\leftarrow}_t) &= -2t\left(v(x^{\leftarrow}_t, t) - \overline{v}(\overline{x}^{\leftarrow}_{k\delta}, k\delta)\right)dt,
\end{aligned}$$

where k is the unique negative integer such that $t \in [k\delta, (k+1)\delta)$. Thus

$$\begin{aligned}
&\frac{d}{dt}\mathbb{E}\left[\|x_t^{\leftarrow} - \overline{x}_t^{\leftarrow}\|\right] \\
\leq& 2\left(\mathbb{E}\left[\|tv(x_t^{\leftarrow}, t) - t\overline{v}(x_t^{\leftarrow}, t)\|\right] + \mathbb{E}\left[\|t\overline{v}(x_t^{\leftarrow}, t) - t\overline{v}(\overline{x}_t^{\leftarrow}, t)\|\right]\right) \\
&\quad + 2\left(\mathbb{E}\left[\|t\overline{v}(\overline{x}_t^{\leftarrow}, t) - t\overline{v}(\overline{x}_t^{\leftarrow}, k\delta)\|\right] + \mathbb{E}\left[\|t\overline{v}(\overline{x}_t^{\leftarrow}, k\delta) - t\overline{v}(\overline{x}_{k\delta}^{\leftarrow}, k\delta)\|\right]\right) \\
\leq& 2\left(\epsilon_{approx} + L_2\mathbb{E}\left[\|x_t^{\leftarrow} - \overline{x}_t^{\leftarrow}\|\right] + \delta L_0 + L_2\mathbb{E}\left[\|\overline{x}_t^{\leftarrow} - \overline{x}_{k\delta}^{\leftarrow}\|\right]\right) \\
\leq& 2\left(\epsilon_{approx} + L_2\mathbb{E}\left[\|x_t^{\leftarrow} - \overline{x}_t^{\leftarrow}\|\right] + \delta L_0 + L_2\left(\delta L_1 + \sqrt{2\delta dt_{\text{max}}}\right)\right).
\end{aligned}$$

By Gronwall's Lemma,

$$\begin{aligned}
&\mathbb{E}\left[\|x_{t_{\text{min}}} - \overline{x}_{t_{\text{min}}}\|\right] \\
=&\mathbb{E}\left[\left\|x_{-t_{\text{min}}}^{\leftarrow} - \overline{x}_{-t_{\text{min}}}^{\leftarrow}\right\|\right] \\
\leq& e^{2L_2(t_{\text{max}} - t_{\text{min}})}\left(\mathbb{E}\left[\left\|x_{-t_{\text{max}}}^{\leftarrow} - \overline{x}_{-t_{\text{max}}}^{\leftarrow}\right\|\right] + \left(\epsilon_{approx} + \delta L_0 + L_2\left(\delta L_1 + \sqrt{2\delta dt_{\text{max}}}\right)\right)(t_{\text{max}} - t_{\text{min}})\right) \\
=& e^{2L_2(t_{\text{max}} - t_{\text{min}})}\left(\mathbb{E}\left[\|x_{t_{\text{max}}} - \overline{x}_{t_{\text{max}}}\|\right] + \left(\epsilon_{approx} + \delta L_0 + L_2\left(\delta L_1 + \sqrt{2\delta dt_{\text{max}}}\right)\right)(t_{\text{max}} - t_{\text{min}})\right)
\end{aligned}$$

□

A.4 Mixing Bounds

Lemma 5. *Consider the same setup as Theorem 3. Assume that $\delta \leq t_{min}$. Let*

$$\begin{aligned}
x_{t_{min}} &= SDE_\theta\left(x_{t_{max}}, t_{max} \to t_{min}\right) \\
y_{t_{min}} &= SDE_\theta\left(y_{t_{max}}, t_{max} \to t_{min}\right).
\end{aligned}$$

Then there exists a coupling between x_s and y_s such that

$$TV\left(x_{t_{min}}, y_{t_{min}}\right) \leq \left(1 - 2Q\left(\frac{B}{2\sqrt{t_{max}^2 - t_{min}^2}}\right) \cdot e^{-BL_1/t_{min} - L_1^2 t_{max}^2/t_{min}^2}\right) TV\left(x_{t_{max}}, y_{t_{max}}\right)$$

Proof. We will construct a coupling between x_t and y_t. First, let $(x_{t_{\text{max}}}, y_{t_{\text{max}}})$ be sampled from the optimal TV coupling, *i.e.*, $Pr(x_{t_{\text{max}}} \neq y_{t_{\text{max}}}) = \frac{1}{2}TV(x_{t_{\text{max}}}, y_{t_{\text{max}}})$. Recall that by definition of SDE$_\theta$, for $t \in ((k-1)\delta, k\delta]$,

$$dx_t = -2ts_\theta(x_{k\delta}, k\delta)dt + \sqrt{2t}dB_t.$$

Let us define a time-rescaled version of x_t: $\overline{x}_t := x_{t^2}$. We verify that

$$d\overline{x}_t = -s_\theta(\overline{x}_{(k\delta)^2}, k\delta)dt + dB_t,$$

where k is the unique integer satisfying $t \in [((k-1)\delta)^2, k^2\delta^2)$. Next, we define the time-reversed process $\overline{x}_t^{\leftarrow} := \overline{x}_{-t}$, and let $v(x,t) := s_\theta(x, -t)$. We verify that there exists a Brownian motion B_t^x such that, for $t \in [-t_{\text{max}}^2, -t_{\text{min}}^2]$,

$$d\overline{x}_t^{\leftarrow} = v_t^x dt + dB_t^x,$$

where $v_t^x = s_\theta(\overline{x}_{-(k\delta)^2}^{\leftarrow}, -k\delta)$, where k is the unique positive integer satisfying $-t \in (((k-1)\delta)^2, (k\delta)^2]$. Let $d\overline{y}_t^{\leftarrow} = v_t^y dt + dB_t^y$, be defined analogously. For any positive integer k and for any $t \in [-(k\delta)^2, -((k-1)\delta)^2)$, let us define

$$z_t = \overline{x}_{-k^2\delta^2}^{\leftarrow} - \overline{y}_{-k^2\delta^2}^{\leftarrow} + (2k-1)\delta^2\left(v_{-(k\delta)^2}^x - v_{-(k\delta)^2}^y\right) + \left(B_t^x - B_{-(k\delta)^2}^x\right) - \left(B_t^y - B_{-(k\delta)^2}^y\right).$$

Let $\gamma_t := \frac{z_t}{\|z_t\|}$. We will now define a coupling between dB_t^x and dB_t^y as

$$dB_t^y = \left(I - 2\mathbb{1}\left\{t \leq \tau\right\}\gamma_t\gamma_t^T\right)dB_t^x,$$

where $\mathbb{1}\{\}$ denotes the indicator function, i.e. $\mathbb{1}\{t \le \tau\} = 1$ if $t \le \tau$, and τ is a stopping time given by the first hitting time of $z_t = 0$. Let $r_t := \|z_t\|$. Consider some $t \in \left(-i^2\delta^2, -(i-1)^2\delta^2\right)$, and Let $j := \frac{t_{\max}}{\delta}$ (assume w.l.o.g that this is an integer), then

$$\begin{aligned} r_t - r_{-t_{\max}^2} &\le \sum_{k=i}^{j} (2k-1)\delta^2 \left\| (v^x_{-(k\delta)^2} - v^y_{-(k\delta)^2}) \right\| + \int_{-t_{\max}^2}^{t} \mathbb{1}\{t \le \tau\} 2dB_s^1 \\ &\le \sum_{k=i}^{j} \left(k^2 - (k-1)^2\right) \delta^2 2L_1 / \left(t_{\min}\right) + \int_{-t_{\max}^2}^{t} \mathbb{1}\{t \le \tau\} 2dB_t^1 \\ &= \int_{-t_{\max}^2}^{-(i-1)\delta^2} \frac{2L_1}{t_{\min}} ds + \int_{-t_{\max}^2}^{t} \mathbb{1}\{t \le \tau\} 2dB_s^1, \end{aligned}$$

where $dB_s^1 = \langle \gamma_t, dB_s^x - dB_s^y \rangle$ is a 1-dimensional Brownian motion. We also verify that

$$\begin{aligned} r_{-t_{\max}^2} &= \left\| z_{-t_{\max}^2} \right\| \\ &= \left\| \overleftarrow{x}_{-t_{\max}^2} - \overleftarrow{y}_{-t_{\max}^2} + (2j-1)\delta^2 \left(v^x_{-t_{\max}^2} - v^y_{-t_{\max}^2} \right) + \left(B_t^x - B^x_{-t_{\max}^2} \right) - \left(B_t^y - B^y_{-t_{\max}^2} \right) \right\| \\ &\le \left\| \overleftarrow{x}_{-t_{\max}^2} + (2j-1)\delta^2 v^x_{-t_{\max}^2} + \left(B^x_{-(j-1)^2\delta^2} - B^x_{-t_{\max}^2} \right) \right\| \\ &\quad + \left\| \overleftarrow{y}_{-t_{\max}^2} + (2j-1)\delta^2 v^y_{-t_{\max}^2} + \left(B^x_{-(j-1)^2\delta^2} - B_t^x + B_t^y - B^y_{-t_{\max}^2} \right) \right\| \le B \end{aligned}$$

where the third relation is by adding and subtracting $B^x_{-(j-1)^2\delta^2} - B_t^x$ and using triangle inequality. The fourth relation is by noticing that $\overleftarrow{x}_{-t_{\max}^2} + (2j-1)\delta^2 v^x_{-t_{\max}^2} + \left(B^x_{-(j-1)^2\delta^2} - B^x_{-t_{\max}^2} \right) = \overleftarrow{x}_{-(j-1)^2\delta^2}$ and that $\overleftarrow{y}_{-t_{\max}^2} + (2j-1)\delta^2 v^y_{-t_{\max}^2} + \left(B^x_{-(j-1)^2\delta^2} - B_t^x + B_t^y - B^y_{-t_{\max}^2} \right) \overset{d}{=} \overleftarrow{y}_{-(j-1)^2\delta^2}$, and then using our assumption in the theorem statement that all processes are supported on a ball of radius $B/2$.

We now define a process s_t defined by $ds_t = 2L_1/t_{\min} dt + 2dB_t^1$, initialized at $s_{-t_{\max}^2} = B \ge r_{-t_{\max}^2}$. We can verify that, up to time τ, $r_t \le s_t$ with probability 1. Let τ' denote the first-hitting time of s_t to 0, then $\tau \le \tau'$ with probability 1. Thus

$$Pr(\tau \le -t_{\min}^2) \ge Pr(\tau' \le -t_{\min}^2) \ge 2Q\left(\frac{B}{2\sqrt{t_{\max}^2 - t_{\min}^2}} \right) \cdot e^{-BL_1/t_{\min} - L_1^2 t_{\max}^2 / t_{\min}^2}$$

where we apply Lemma 6. The proof follows by noticing that, if $\tau \le -t_{\min}^2$, then $x_{t_{\min}} = y_{t_{\min}}$. This is because if $\tau \in [-k^2\delta^2, -(k-1)^2\delta^2]$, then $\overleftarrow{x}_{-(k-1)^2\delta^2} = \overleftarrow{y}_{-(k-1)^2\delta^2}$, and thus $\overleftarrow{x}_t = \overleftarrow{y}_t$ for all $t \ge -(k-1)^2\delta^2$, in particular, at $t = -t_{\min}^2$.

□

Lemma 6. *Consider the stochastic process*

$$dr_t = dB_t^1 + cdt.$$

Assume that $r_0 \le B/2$. Let τ denote the hitting time for $r_t = 0$. Then for any $T \in \mathbb{R}^+$,

$$Pr(\tau \le T) \ge 2Q\left(\frac{B}{2\sqrt{T}} \right) \cdot e^{-ac - \frac{c^2 T}{2}},$$

where Q is the tail probability of a standard Gaussian defined in Definition 1.

Proof. We will use he following facts in our proof:

1. For $x \sim \mathcal{N}(0, \sigma^2)$, $Pr(x > r) = \frac{1}{2}\left(1 - erf\left(\frac{r}{\sqrt{2}\sigma}\right)\right) = \frac{1}{2} erfc\left(\frac{r}{\sqrt{2}\sigma}\right)$.
2. $\int_0^T \frac{a \exp\left(-\frac{a^2}{2t}\right)}{\sqrt{2\pi t^3}} dt = erfc\left(\frac{a}{\sqrt{2T}}\right) = 2Pr\left(\mathcal{N}(0,T) > a\right) = 2Q\left(\frac{a}{\sqrt{T}}\right)$ by definition of Q.

Let $dr_t = dB_t^1 + cdt$, with $r_0 = a$. The density of the hitting time τ is given by

$$p(\tau = t) = f(a, c, t) = \frac{a \exp\left(-\frac{(a+ct)^2}{2t}\right)}{\sqrt{2\pi t^3}}. \tag{18}$$

(see e.g. [3]). From item 2 above,

$$\int_0^T f(a, 0, t)dt = 2Q\left(\frac{a}{\sqrt{T}}\right).$$

In the case of a general $c \neq 0$, we can bound $\frac{(a+ct)^2}{2t} = \frac{a^2}{2t} + ac + \frac{c^2 t}{2}$. Consequently,

$$f(a, c, t) \geq f(a, 0, t) \cdot e^{-ac - \frac{c^2 t}{2}}.$$

Therefore,

$$Pr(\tau \leq T) = \int_0^T f(a, c, t)dt \geq \int_0^T f(a, 0, t)dte^{-c} = 2Q\left(\frac{B}{2\sqrt{T}}\right) \cdot e^{-ac - \frac{c^2 T}{2}}.$$

□

A.5 TV Overlap

Definition 1. *Let x be sampled from standard normal distribution $\mathcal{N}(0, 1)$. We define the Gaussian tail probability $Q(a) := Pr(x \geq a)$.*

Lemma 7. *We verify that for any two random vectors $\xi_x \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$ and $\xi_y \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})$, each belonging to $\mathbb{R}^d$, the total variation distance between $x' = x + \xi_x$ and $y' = y + \xi_y$ is given by*

$$TV(x', y') = 1 - 2Q\left(r\right) \leq 1 - \frac{2r}{r^2 + 1}\frac{1}{\sqrt{2\pi}}e^{-r^2/2},$$

where $r = \frac{\|x-y\|}{2\sigma}$, and $Q(r) = Pr(\xi \geq r)$, when $\xi \sim \mathcal{N}(0, 1)$.

Proof. Let $\gamma := \frac{x-y}{\|x-y\|}$. We decompose x', y' into the subspace/orthogonal space defined by γ:

$$x' = x^\perp + \xi_x^\perp + x^\parallel + \xi_x^\parallel$$
$$y' = y^\perp + \xi_y^\perp + y^\parallel + \xi_y^\parallel$$

where we define

$$\begin{aligned} x^\parallel &:= \gamma\gamma^T x & x^\perp &:= x - x^\parallel \\ y^\parallel &:= \gamma\gamma^T y & y^\perp &:= y - y^\parallel \\ \xi_x^\parallel &:= \gamma\gamma^T \xi_x & \xi_x^\perp &:= \xi_x - \xi_x^\parallel \\ \xi_y^\parallel &:= \gamma\gamma^T \xi_y & \xi_y^\perp &:= \xi_y - \xi_y^\parallel \end{aligned}$$

We verify the independence $\xi_x^\perp \perp\!\!\!\perp \xi_x^\parallel$ and $\xi_y^\perp \perp\!\!\!\perp \xi_y^\parallel$ as they are orthogonal decompositions of the standard Gaussian. We will define a coupling between x' and y' by setting $\xi_x^\perp = \xi_y^\perp$. Under this coupling, we verify that

$$\left(x^\perp + \xi_x^\perp\right) - \left(y^\perp + \xi_y^\perp\right) = x - y - \gamma\gamma^T(x - y) = 0$$

Therefore, $x' = y'$ if and only if $x^\parallel + \xi_x^\parallel = y^\parallel + \xi_y^\parallel$. Next, we draw (a, b) from the optimal coupling between $\mathcal{N}(0, 1)$ and $\mathcal{N}(\frac{\|x-y\|}{\sigma}, 1)$. We verify that $x^\parallel + \xi_x^\parallel$ and $y^\parallel + \xi_y^\parallel$ both lie in the span of γ. Thus it suffices to compare $\left\langle \gamma, x^\parallel + \xi_x^\parallel \right\rangle$ and $\left\langle \gamma, y^\parallel + \xi_y^\parallel \right\rangle$. We verify that $\left\langle \gamma, x^\parallel + \xi_x^\parallel \right\rangle =$

$\langle \gamma, y^{\parallel} \rangle + \langle \gamma, x^{\parallel} - y^{\parallel} \rangle + \left\langle \gamma, \xi_x^{\parallel} \right\rangle \sim \mathcal{N}(\langle \gamma, y^{\parallel} \rangle + \|x - y\|, \sigma^2) \overset{d}{=} \langle \gamma, y^{\parallel} \rangle + \sigma b$. We similarly verify that $\left\langle \gamma, y^{\parallel} + \xi_y^{\parallel} \right\rangle = \langle \gamma, y^{\parallel} \rangle + \left\langle \gamma, \xi_y^{\parallel} \right\rangle \sim \mathcal{N}(\langle \gamma, y^{\parallel} \rangle, \sigma^2) \overset{d}{=} \langle \gamma, y^{\parallel} \rangle + \sigma a$.

Thus $TV(x', y') = TV(\sigma a, \sigma b) = 1 - 2Q\left(\frac{\|x-y\|}{2\sigma}\right)$. The last inequality follows from

$$Pr(\mathcal{N}(0,1) \geq r) \geq \frac{r}{r^2+1} \frac{1}{\sqrt{2\pi}} e^{-r^2/2}$$

□

B More on Restart Algorithm

B.1 EDM Discretization Scheme

[13] proposes a discretization scheme for ODE given the starting t_{max} and end time t_{min}. Denote the number of steps as N, then the EDM discretization scheme is:

$$t_{i<N} = \left(t_{\text{max}}^{\frac{1}{\rho}} + \frac{i}{N-1} (t_{\text{min}}^{\frac{1}{\rho}} - t_{\text{max}}^{\frac{1}{\rho}}) \right)^{\rho}$$

with $t_0 = t_{\text{max}}$ and $t_{N-1} = t_{\text{min}}$. ρ is a hyperparameter that determines the extent to which steps near t_{min} are shortened. We adopt the value $\rho = 7$ suggested by [13] in all of our experiments. We apply the EDM scheme to creates a time discretization in each Restart interval $[t_{\text{max}}, t_{\text{min}}]$ in the Restart backward process, as well as the main backward process between $[0, T]$ (by additionally setting $t_{\text{min}} = 0.002$ and $t_N = 0$ as in [13]). It is important to note that t_{min} should be included within the list of time steps in the main backward process to seamlessly incorporate the Restart interval into the main backward process. We summarize the scheme as a function in Algorithm 1.

Algorithm 1 EDM_Scheme($t_{\text{min}}, t_{\text{max}}, N, \rho = 7$)

1: **return** $\left\{ (t_{\text{max}}^{\frac{1}{\rho}} + \frac{i}{N-1}(t_{\text{min}}^{\frac{1}{\rho}} - t_{\text{max}}^{\frac{1}{\rho}}))^{\rho} \right\}_{i=0}^{N-1}$

B.2 Restart Algorithm

We present the pseudocode for the Restart algorithm in Algorithm 2. In this pseudocode, we describe a more general case that applies l-level Restarting strategy. For each Restart segment, the include the number of steps in the Restart backward process N_{Restart}, the Restart interval $[t_{\text{min}}, t_{\text{max}}]$ and the number of Restart iteration K. We further denote the number of steps in the main backward process as N_{main}. We use the EDM discretization scheme (Algorithm 1) to construct time steps for the main backward process ($t_0 = T, t_{N_{\text{main}}} = 0$) as well as the Restart backward process, when given the starting/end time and the number of steps.

Although Heun's 2nd order method [2] (Algorithm 3) is the default ODE solver in the pseudocode, it can be substituted with other ODE solvers, such as Euler's method or the DPM solver [16].

The provided pseudocode in Algorithm 2 is tailored specifically for diffusion models [13]. To adapt Restart for other generative models like PFGM++ [28], we only need to modify the Gaussian perturbation kernel in the Restart forward process (line 10 in Algorithm 2) to the one used in PFGM++.

C Experimental Details

In this section, we discuss the configurations for different samplers in details. All the experiments are conducted on eight NVIDIA A100 GPUs.

C.1 Configurations for Baselines

We select **Vanilla SDE** [23], **Improved SDE** [13], **Gonna Go Fast** [12] as SDE baselines and the **Heun**'s 2nd order method [2] (Alg 3) as ODE baseline on standard benchmarks CIFAR-10 and

Algorithm 2 Restart sampling

Input: Score network s_θ, time steps in main backward process $t_{i\in\{0,N_{\text{main}}\}}$, Restart parameters $\{(N_{\text{Restart},j}, K_j, t_{\text{min},j}, t_{\text{max},j})\}_{j=1}^{l}$
Round $t_{\text{min},j\in\{1,l\}}$ to its nearest neighbor in $t_{i\in\{0,N_{\text{main}}\}}$
Sample $x_0 \sim \mathcal{N}(\mathbf{0}, T^2\boldsymbol{I})$
for $i = 0 \dots N_{\text{main}} - 1$ **do** ▷ Main backward process
  $x_{t_{i+1}} = \text{OneStep\_Heun}(s_\theta, t_i, t_{i+1})$ ▷ Running single step ODE
  if $\exists j \in \{1, \dots, l\}, t_{i+1} = t_{\text{min},j}$ **then**
    $t_{\text{min}} = t_{\text{min},j}, t_{\text{max}} = t_{\text{max},j}, K = K_j, N_{\text{Restart}} = N_{\text{Restart},j}$
    $x^0_{t_{\text{min}}} = x_{t_{i+1}}$
    for $k = 0 \dots K - 1$ **do** ▷ **Restart for K iterations**
      $\varepsilon_{t_{\text{min}}\to t_{\text{max}}} \sim \mathcal{N}(\mathbf{0}, (t^2_{\text{max}} - t^2_{\text{min}})\boldsymbol{I})$
      $x^{k+1}_{t_{\text{max}}} = x^k_{t_{\text{min}}} + \varepsilon_{t_{\text{min}}\to t_{\text{max}}}$ ▷ **Restart forward process**
      $\{\bar{t}_m\}_{m=0}^{N_{\text{Restart}}-1} = \text{EDM\_Scheme}(t_{\text{min}}, t_{\text{max}}, N_{\text{Restart}})$
      for $m = 0 \dots N_{\text{Restart}} - 1$ **do** ▷ **Restart backward process**
        $x^{k+1}_{\bar{t}_{m+1}} = \text{OneStep\_Heun}(s_\theta, \bar{t}_m, \bar{t}_{m+1})$
      end for
    end for
  end if
end for
return $x_{t_{N_{\text{main}}}}$

Algorithm 3 OneStep_Heun$(s_\theta, x_{t_i}, t_i, t_{i+1})$

$d_i = t_i s_\theta(x_{t_i}, t_i)$
$x_{t_{i+1}} = x_{t_i} - (t_{i+1} - t_i)d_i$
if $t_{i+1} \neq 0$ **then**
  $d'_i = t_{i+1} s_\theta(x_{t_{i+1}}, t_{i+1})$
  $x_{t_{i+1}} = x_{t_i} - (t_{i+1} - t_i)(\frac{1}{2}d_i + \frac{1}{2}d'_i)$
end if
return $x_{t_{i+1}}$

ImageNet 64×64. We choose **DDIM** [22], **Heun**'s 2nd order method, and **DDPM** [9] for comparison on Stable Diffusion model.

Vanilla SDE denotes the reverse-time SDE sampler in [23]. For Improved SDE, we use the recommended dataset-specific hyperparameters (*e.g.*, $S_{\text{max}}, S_{\text{min}}, S_{\text{churn}}$) in Table 5 of the EDM paper [13]. They obtained these hyperparameters by grid search. Gonna Go Fast [12] applied an adaptive step size technique based on Vanilla SDE and we directly report the FID scores listed in [12] for Gonna Go Fast on CIFAR-10 (VP). For fair comparison, we use the EDM discretization scheme [13] for Vanilla SDE, Improved SDE, Heun as well as Restart.

We borrow the hyperparameters such as discretization scheme or initial noise scale on Stable Diffusion models in the diffuser [2] code repository. We directly use the DDIM and DDPM samplers implemented in the repo. We apply the same set of hyperparameters to Heun and Restart.

C.2 Configurations for Restart

We report the configurations for Restart for different models and NFE on standard benchmarks CIFAR-10 and ImageNet 64×64. The hyperparameters of Restart include the number of steps in the main backward process N_{main}, the number of steps in the Restart backward process N_{Restart}, the Restart interval $[t_{\text{min}}, t_{\text{max}}]$ and the number of Restart iteration K. In Table 3 (CIFAR-10, VP) we provide the quintuplet $(N_{\text{main}}, N_{\text{Restart}}, t_{\text{min}}, t_{\text{max}}, K)$ for each experiment. Since we apply the multi-level Restart strategy for ImageNet 64×64, we provide N_{main} as well as a list of quadruple $\{(N_{\text{Restart},i}, K_i, t_{\text{min},i}, t_{\text{max},i})\}_{i=1}^{l}$ (l is the number of Restart interval depending on experiments) in Table 5. In order to integrate the Restart time interval to the main backward process, we round $t_{\text{min},i}$

[2] https://github.com/huggingface/diffusers

to its nearest neighbor in the time steps of main backward process, as shown in line 2 of Algorithm 2. We apply Heun method for both main/backward process. The formula for NFE calculation is NFE $= \underbrace{2 \cdot N_{\text{main}} - 1}_{\text{main backward process}} + \sum_{i=1}^{l} \underbrace{K_i}_{\text{number of repetitions}} \cdot \underbrace{(2 \cdot (N_{\text{Restart},i} - 1))}_{\text{per iteration in } i^{\text{th}} \text{ Restart interval}}$ in this case. Inspired by [13], we inflate the additive noise in the Restart forward process by multiplying $S_{\text{noise}} = 1.003$ on ImageNet 64×64, to counteract the over-denoising tendency of neural networks. We also observe that setting $\gamma = 0.05$ in Algorithm 2 of EDM [13] would sligtly boost the Restart performance on ImageNet 64×64 when $t \in [0.01, 1]$.

We further include the configurations for Restart on Stable Diffusion models in Table 10, with a varying guidance weight w. Similar to ImageNet 64×64, we use multi-level Restart with a fixed number of steps $N_{\text{main}} = 30$ in the main backward process. We utilize the Euler method for the main backward process and the Heun method for the Restart backward process, as our empirical observations indicate that the Heun method doesn't yield significant improvements over the Euler method, yet necessitates double the steps. The number of steps equals to $N_{\text{main}} + \sum_{i=1}^{l} K_i \cdot (2 \cdot (N_{\text{Restart},i} - 1))$ in this case. We set the total number of steps to 66, including main backward process and Restart backward process.

Given the prohibitively large search space for each Restart quadruple, a comprehensive enumeration of all possibilities is impractical due to computational limitations. Instead, we adjust the configuration manually, guided by the heuristic that weaker/smaller models or more challenging tasks necessitate a stronger Restart strength (e.g., larger K, wider Restart interval, etc). On average, we select the best configuration from 5 sets for each experiment; these few trials have empirically outperformed previous SDE/ODE samplers. We believe that developing a systematic approach for determining Restart configurations could be of significant value in the future.

C.3 Pre-trained Models

For CIFAR-10 dataset, we use the pre-trained VP and EDM models from the EDM repository [3], and PFGM++ ($D = 2048$) model from the PFGM++ repository [4]. For ImageNet 64×64, we borrow the pre-trained EDM model from EDM repository as well.

C.4 Classifier-free Guidance

We follow the convention in [20], where each step in classifier-free guidance is as follows:

$$\tilde{s}_\theta(x, c, t) = w s_\theta(x, c, t) + (1 - w) s_\theta(x, t)$$

where c is the conditions, and$s_\theta(x, c, t)/s_\theta(x, t)$ is the conditional/unconditional models, sharing parameters. Increasing w would strength the effect of guidance, usually leading to a better text-image alignment [20].

C.5 More on the Synthetic Experiment

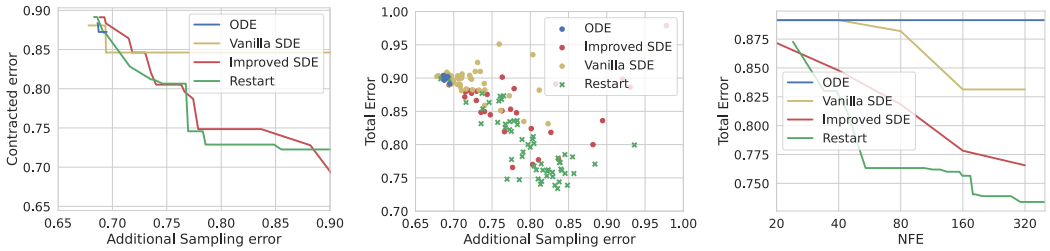
C.5.1 Discrete Dataset

We generate the underlying discrete dataset S with $|S| = 2000$ as follows. Firstly, we sample 2000 points, denoted as S_1, from a mixture of two Gaussians in $\mathbb{R}^4$. Next, we project these points onto $\mathbb{R}^{20}$. To ensure a variance of 1 on each dimension, we scale the coordinates accordingly. This setup aims to simulate data points that primarily reside on a lower-dimensional manifold with multiple modes.

The specific details are as follows: $S_1 \sim 0.3N(a, s^2 I) + 0.7(-a, s^2 I)$, where $a = (3, 3, 3, 3) \subset \mathbb{R}^4$ and $s = 1$. Then, we randomly select a projection matrix $P \in \mathbb{R}^{20 \times 4}$, where each entry is drawn from $N(0, 1)$, and compute $S_2 = PS_1$. Finally, we scale each coordinate by a constant factor to ensure a variance of 1.

[3] https://github.com/NVlabs/edm

[4] https://github.com/Newbeeer/pfgmpp



(a) (b) (c)

Figure 7: Comparison of additional sampling error versus **(a)** contracted error (plotting the Pareto frontier) and **(b)** total error (using a scatter plot). **(c)** Pareto frontier of NFE versus total error.

C.5.2 Model Architecture

We employ a common MLP architecture with a latent size of 64 to learn the score function. The training method is adapted from [13], which includes the preconditioning technique and denoising score-matching objective [25].

C.5.3 Varying Hyperparameters

To achieve the best trade-off between contracted error and additional sampling error, and optimize the NFE versus FID (Fréchet Inception Distance) performance, we explore various hyperparameters. [13] shows that the Vanilla SDE can be endowed with additional flexibility by varying the coefficient $\beta(t)$ (Eq.(6) in [13]). Hence, regarding SDE, we consider NFE values from $\{20, 40, 80, 160, 320\}$, and multiply the original $\beta(t) = \dot{\sigma}(t)/\sigma(t)$ [13] with values from $\{0, 0.25, 0.5, 1, 1.5, 2, 4, 8\}$. It is important to note that larger NFE values do not lead to further performance improvements. For restarts, we tried the following two settings: first we set the number of steps in Restart backward process to 40 and vary the number of Restart iterations K in the range $\{0, 5, 10, 15, 20, 25, 30, 35\}$. We also conduct a grid search with the number of Restart iterations K ranging from 5 to 25 and the number of steps in Restart backward process varying from 2 to 7. For ODE, we experiment with the number of steps set to $\{20, 40, 80, 160, 320, 640\}$.

Additionally, we conduct an experiment for Improved SDE in EDM. We try different values of S_{churn} in the range of $\{0, 1, 2, 4, 8, 16, 32, 48, 64\}$. We also perform a grid search where the number of steps ranged from 20 to 320 and S_{churn} takes values of $[0.2 \times \text{steps}, 0.5 \times \text{steps}, 20, 60]$. The plot combines the results from SDE and is displayed in Figure 7.

To mitigate the impact of randomness, we collect the data by averaging the results from five runs with the same hyperparameters. To compute the Wasserstein distance between two discrete distributions, we use minimum weight matching.

C.5.4 Plotting the Pareto frontier

We generate the Pareto frontier plots as follows. For the additional sampling error versus contracted error plot, we first sort all the data points based on their additional sampling error and then connect the data points that represent prefix minimums of the contracted error. Similarly, for the NFE versus FID plot, we sort the data points based on their NFE values and connect the points where the FID is a prefix minimum.

D Extra Experimental Results

D.1 Numerical Results

In this section, we provide the corresponding numerical reuslts of Fig. 3(a) and Fig. 3(b), in Table 2, 3 (CIFAR-10 VP) and Table 4, 5 (ImageNet 64×64 EDM), respectively. We also include the performance of Vanilla SDE in those tables. For the evaluation, we compute the Fréchet distance between 50000 generated samples and the pre-computed statistics of CIFAR-10 and ImageNet 64×64. We follow the evaluation protocol in EDM [13] that calculates each FID scores three times with different seeds and report the minimum.

We also provide the numerical results on the Stable Diffusion model [19], with a classifier guidance weight $w = 2, 3, 5, 8$ in Table 6, 7, 8, 9. As in [17], we report the zero-shot FID score on 5K random prompts sampled from the COCO validation set. We evaluate CLIP score [6] with the open-sourced ViT-g/14 [11], Aesthetic score by the more recent LAION-Aesthetics Predictor V2 [5]. We average the CLIP and Aesthetic scores over 5K generated samples. The number of function evaluations is two times the sampling steps in Stable Diffusion model, since each sampling step involves the evaluation of the conditional and unconditional model.

Table 2: CIFAR-10 sample quality (FID score) and number of function evaluations (NFE) on VP [23] for baselines

	NFE	FID
ODE (Heun) [13]	1023	2.90
	511	2.90
	255	2.90
	127	2.90
	63	2.89
	35	2.97
Vanilla SDE [23]	1024	2.79
	512	4.01
	256	4.79
	128	12.57
Gonna Go Fast [12]	1000	2.55
	329	2.70
	274	2.74
	179	2.59
	147	2.95
	49	72.29
Improved SDE [13]	1023	2.35
	511	2.37
	255	2.40
	127	2.58
	63	2.88
	35	3.45

Table 3: CIFAR-10 sample quality (FID score), number of function evaluations (NFE) and configurations on VP [23] for Restart

NFE	FID	Configuration $(N_{\text{main}}, N_{\text{Restart},i}, K_i, t_{\text{min},i}, t_{\text{max},i})$
519	2.11	$(20, 9, 30, 0.06, 0.20)$
115	2.21	$(18, 3, 20, 0.06, 0.30)$
75	2.27	$(18, 3, 10, 0.06, 0.30)$
55	2.45	$(18, 3, 5, 0.06, 0.30)$
43	2.70	$(18, 3, 2, 0.06, 0.30)$

[5] https://github.com/christophschuhmann/improved-aesthetic-predictor

Table 4: ImageNet 64×64 sample quality (FID score) and number of function evaluations (NFE) on EDM [13] for baselines

	NFE	FID (50k)
ODE (Heun) [13]	1023	2.24
	511	2.24
	255	2.24
	127	2.25
	63	2.30
	35	2.46
Vanilla SDE [23]	1024	1.89
	512	3.38
	256	11.91
	128	59.71
Improved SDE [13]	1023	1.40
	511	1.45
	255	1.50
	127	1.75
	63	2.24
	35	2.97

Table 5: ImageNet 64×64 sample quality (FID score), number of function evaluations (NFE) and configurations on EDM [13] for Restart

NFE	FID (50k)	Configuration $N_{\text{main}}, \{(N_{\text{Restart},i}, K_i, t_{\text{min},i}, t_{\text{max},i})\}_{i=1}^{l}$
623	1.36	36, {(10, 3, 19.35, 40.79),(10, 3, 1.09, 1.92), (7, 6, 0.59, 1.09), (7, 6, 0.30, 0.59), (7, 25, 0.06, 0.30)}
535	1.39	36, {(6, 1, 19.35, 40.79),(6, 1, 1.09, 1.92), (7, 6, 0.59, 1.09), (7, 6, 0.30, 0.59), (7, 25, 0.06, 0.30)}
385	1.41	36, {(3, 1, 19.35, 40.79),(6, 1, 1.09, 1.92), (6, 5, 0.59, 1.09), (6, 5, 0.30, 0.59), (6, 20, 0.06, 0.30)}
203	1.46	36, {(4, 1, 19.35, 40.79),(4, 1, 1.09, 1.92), (4, 5, 0.59, 1.09), (4, 5, 0.30, 0.59), (6, 6, 0.06, 0.30)}
165	1.51	18, {(3, 1, 19.35, 40.79),(4, 1, 1.09, 1.92), (4, 5, 0.59, 1.09), (4, 5, 0.30, 0.59), (4, 10, 0.06, 0.30)}
99	1.71	18, {(3, 1, 19.35, 40.79),(4, 1, 1.09, 1.92), (4, 4, 0.59, 1.09), (4, 1, 0.30, 0.59), (4, 4, 0.06, 0.30)}
67	1.95	18, {(5, 1, 19.35, 40.79),(5, 1, 1.09, 1.92), (5, 1, 0.59, 1.09), (5, 1, 0.06, 0.30)}
39	2.38	14, {(3, 1, 19.35, 40.79), (3, 1, 1.09, 1.92), (3, 1, 0.06, 0.30)}

Table 6: Numerical results on Stable Diffusion v1.5 with a classifier-free guidance weight $w = 2$

	Steps	FID (5k) ↓	CLIP score ↑	Aesthetic score ↑
DDIM [22]	50	16.08	0.2905	5.13
	100	15.35	0.2920	5.15
Heun	51	18.80	0.2865	5.14
	101	18.21	0.2871	5.15
DDPM [9]	100	13.53	0.3012	5.20
	200	13.22	0.2999	5.19
Restart	66	13.16	0.2987	5.19

Table 7: Numerical results on Stable Diffusion v1.5 with a classifier-free guidance weight $w = 3$

	Steps	FID (5k) ↓	CLIP score ↑	Aesthetic score ↑
DDIM [22]	50	14.28	0.3056	5.22
	100	14.30	0.3056	5.22
Heun	51	15.63	0.3022	5.20
	101	15.40	0.3026	5.21
DDPM [9]	100	15.72	0.3129	5.28
	200	15.13	0.3131	5.28
Restart	66	14.48	0.3079	5.25

Table 8: Numerical results on Stable Diffusion v1.5 with a classifier-free guidance weight $w = 5$

	Steps	FID (5k) ↓	CLIP score ↑	Aesthetic score ↑
DDIM [22]	50	16.60	0.3154	5.31
	100	16.80	0.3157	5.31
Heun	51	16.26	0.3135	5.28
	101	16.38	0.3136	5.29
DDPM [9]	100	19.62	0.3197	5.36
	200	18.88	0.3200	5.35
Restart	66	16.21	0.3179	5.33

Table 9: Numerical results on Stable Diffusion v1.5 with a classifier-free guidance weight $w = 8$

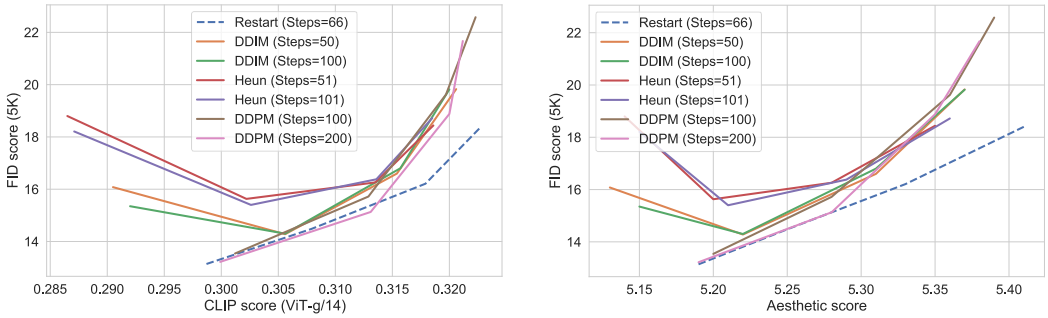
	Steps	FID (5k) ↓	CLIP score ↑	Aesthetic score ↑
DDIM [22]	50	19.83	0.3206	5.37
	100	19.82	0.3200	5.37
Heun	51	18.44	0.3186	5.35
	101	18.72	0.3185	5.36
DDPM [9]	100	22.58	0.3223	5.39
	200	21.67	0.3212	5.38
Restart	47	18.40	0.3228	5.41

D.2 Study on Adjusting t_{min}

We also investigate the impact of varying t_{min} when $t_{\text{max}} = t_{\text{min}} + 0.3$. Fig. **??** reveals that FID scores achieve a minimum at a t_{min} close to 0 on VP, indicating higher accumulated errors at the end of

Table 10: Restart (Steps=66) configurations on Stable Diffusion v1.5

w	Configuration N_{main}, $\{(N_{\text{Restart},i}, K_i, t_{\text{min},i}, t_{\text{max},i})\}_{i=1}^{l}$
2	30, {(5, 2, 1, 9), (5, 2, 5, 10)}
3	30, {(2, 10, 0.1, 3)}
5	30, {(2, 10, 0.1, 2)}
8	30, {(2, 10, 0.1, 2)}



(a) FID versus CLIP score

(b) FID versus Aesthetic score

Figure 8: FID score versus **(a)** CLIP ViT-g/14 score and **(b)** Aesthetic score for text-to-image generation at 512×512 resolution, using Stable Diffusion v1.5 with varying classifier-free guidance weight $w = 2, 3, 5, 8$.

sampling and poor neural estimations at small t. Note that the Restart interval 0.3 is about twice the length of the one in Table 1 and Restart does not outperform the ODE baseline on EDM. This suggests that, as a rule of thumb, we should apply greater Restart strength (*e.g.*, larger K, $t_{\text{max}} - t_{\text{min}}$) for weaker or smaller architectures and vice versa.

E Extended Generated Images

In this section, we provide extended generated images by Restart, DDIM, Heun and DDPM on text-to-image Stable Diffusion v1.5 model [19]. We showcase the samples of four sets of text prompts in Fig. 10, Fig. 11, Fig. 12, Fig. 13, with a classifier-guidance weight $w = 8$.

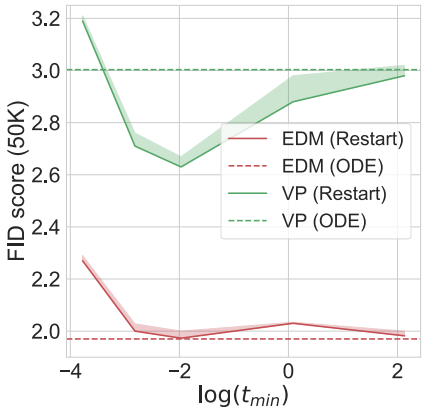


Figure 9: Adjusting t_{min} in Restart on VP/EDM

(a) Restart (Steps=66)

(b) DDIM (Steps=100)

(c) Heun (Steps=101)

(d) DDPM (Steps=100)

Figure 10: Generated images with text prompt="A photo of an astronaut riding a horse on mars" and $w = 8$.

(a) Restart (Steps=66)

(b) DDIM (Steps=100)

(c) Heun (Steps=101)

(d) DDPM (Steps=100)

Figure 11: Generated images with text prompt="A raccoon playing table tennis" and $w = 8$.

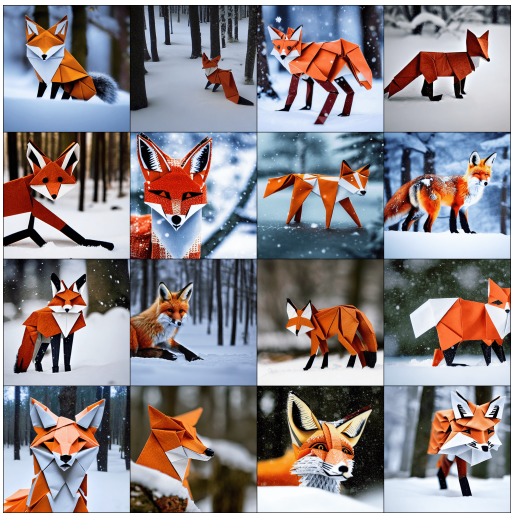
F Broader Impact

The field of deep generative models incorporating differential equations is rapidly evolving and holds significant potential to shape our society. Nowadays, a multitude of photo-realistic images generated by text-to-image Stable Diffusion models populate the internet. Our work introduces Restart, a novel sampling algorithm that outperforms previous samplers for diffusion models and PFGM++. With applications extending across diverse areas, the Restart sampling algorithm is especially suitable for generation tasks demanding high quality and rapid speed. Yet, it is crucial to recognize that the utilization of such algorithms can yield both positive and negative repercussions, contingent on their specific applications. On the one hand, Restart sampling can facilitate the generation of highly realistic images and audio samples, potentially advancing sectors such as entertainment, advertising, and education. On the other hand, it could also be misused in *deepfake* technology, potentially leading to social scams and misinformation. In light of these potential risks, further research is required to develop robustness guarantees for generative models, ensuring their use aligns with ethical guidelines and societal interests.

(a) Restart (Steps=66)

(b) DDIM (Steps=100)

(c) Heun (Steps=101)



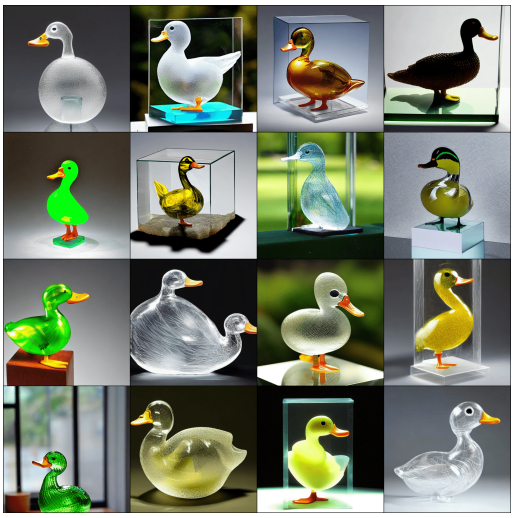
(d) DDPM (Steps=100)

Figure 12: Generated images with text prompt="Intricate origami of a fox in a snowy forest" and $w = 8$.

(a) Restart (Steps=66)

(b) DDIM (Steps=100)

(c) Heun (Steps=101)



(d) DDPM (Steps=100)

Figure 13: Generated images with text prompt="A transparent sculpture of a duck made out of glass" and $w = 8$.