A. Assumptions of Theorem 2

First, let us define the infinitesimal generator of the diffusion (2). Formally, the *generator* $\mathcal{L}$ of the diffusion (2) is defined for any compactly supported twice differentiable function $f : \mathbf{R}^L \to \mathbf{R}$, such that,

$$\mathcal{L}f(\mathbf{Z}_t) =\triangleq \lim_{h\to 0^+} \frac{\mathbb{E}\left[f(\mathbf{Z}_{t+h})\right] - f(\mathbf{Z}_t)}{h}$$
$$= \left(F(\mathbf{Z}_t)\cdot\nabla + \frac{1}{2}\left(G(\mathbf{Z}_t)G(\mathbf{Z}_t)^T\right) : \nabla\nabla^T\right) f(\mathbf{Z}_t)\ ,$$

where $\mathbf{a}\cdot\mathbf{b} \triangleq \mathbf{a}^T\,\mathbf{b}$, $\mathbf{A}:\mathbf{B} \triangleq \mathrm{tr}(\mathbf{A}^T\,\mathbf{B})$, $h \to 0^+$ means h approaches zero along the positive real axis.

Given an ergodic diffusion (2) with an invariant measure $\rho(\mathbf{Z})$, the posterior average is defined as: $\bar{\psi} \triangleq \int \psi(\mathbf{Z})\rho(\mathbf{Z})\mathrm{d}\,\mathbf{Z}$ for some test function $\psi(\mathbf{Z})$ of interest. For a given numerical method with generated samples $(\mathbf{z}_k)_{k=1}^K$, we use the *sample average* $\hat{\psi}$ defined as $\hat{\psi}_K = \frac{1}{K}\sum_{k=1}^K \psi(\mathbf{z}_k)$ to approximate $\bar{\psi}$. We define a functional $\tilde{\psi}$ that solves the following *Poisson Equation*:

$$\mathcal{L}\tilde{\psi}(\mathbf{z}_k) = \psi(\mathbf{z}_k) - \bar{\psi} \tag{12}$$

We make the following assumptions on $\tilde{\psi}$.

Assumption 1 *$\tilde{\psi}$ exists, and its up to 4rd-order derivatives, $\mathcal{D}^k\tilde{\psi}$, are bounded by a function $\mathcal{V}$, i.e., $\|\mathcal{D}^k\tilde{\psi}\| \le C_k\mathcal{V}^{p_k}$ for $k = (0,1,2,3,4)$, $C_k, p_k > 0$. Furthermore, the expectation of $\mathcal{V}$ on $\{\mathbf{z}_k\}$ is bounded: $\sup_l \mathbb{E}\mathcal{V}^p(\mathbf{z}_k) < \infty$, and $\mathcal{V}$ is smooth such that $\sup_{s\in(0,1)} \mathcal{V}^p\left(s\,\mathbf{z} + (1-s)\,\mathbf{y}\right) \le C\left(\mathcal{V}^p\left(\mathbf{z}\right) + \mathcal{V}^p\left(\mathbf{y}\right)\right)$, $\forall\,\mathbf{z},\mathbf{y}, p \le \max\{2p_k\}$ for some $C > 0$.*

B. Proofs for Section 3

Proof [Sketch Proof of Lemma 1] First note that (5) in Lemma 1 corresponds to eq.13 in (Jordan et al., 1998), where $F(p)$ in (Jordan et al., 1998) is in the form of $\mathrm{KL}(\rho\|p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}))$ in our setting.

Proposition 4.1 in (Jordan et al., 1998) then proves that (5) has a unique solution. Theorem 5.1 in (Jordan et al., 1998) then guarantees that the solution of (5) approach the solution of the Fokker-Planck equation in (3), which is ρ_T in the limit of $h \to 0$.

Since this is true for each k (thus each t in ρ_t), we conclude that $\tilde{\rho}_k = \rho_{hk}$ in the limit of $h \to 0$. ■

To prove Theorem 2, we first need a convergence result about convergence to equilibrium in Wasserstein distance for Fokker-Planck equations, which is presented in (Bolley et al., 2012). Putting in our setting, we can get the following lemma based on Corollary 2.4 in (Bolley et al., 2012).

Lemma 6 ((Bolley et al., 2012)) *Let ρ_T be the solution of the FP equation (3) at time T, $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$ be the joint posterior distribution given $\mathbf{x}$. Assume that $\int \rho_T(\mathbf{z})p_{\boldsymbol{\theta}}^{-1}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z} < \infty$ and there exists a constant C such that $\frac{\mathrm{d}W_2^2(\rho_T, p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}))}{\mathrm{d}t} \ge CW_2^2\left(\rho_T, p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\right)$. Then*

$$W_2\left(\rho_T, p(\mathbf{x},\mathbf{z})\right) \le W_2\left(\rho_0, p(\mathbf{x},\mathbf{z})\right)e^{-CT}\ . \tag{13}$$

We further need to borrow convergence results from (Mattingly et al., 2010; Vollmer et al., 2016; Chen et al., 2015) to characterize error bounds of a numerical integrator for the diffusion (2). Specifically, the goal is to evaluate the posterior average of a test function $\psi(\mathbf{z})$, defined as $\bar{\psi} \triangleq \int \psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}$. When using a numerical integrator to solve (2) to get samples $\{\mathbf{z}_k\}_{k=1}^K$, the sample average $\hat{\psi}_K \triangleq \frac{1}{K}\sum_{k=1}^K \psi(\mathbf{z}_k)$ is used to approximate the posterior average. The accuracy is characterized by the mean square error (MSE) defined as: $\mathbb{E}\left(\hat{\psi}_K - \bar{\psi}\right)^2$. Lemma 7 derives the bound for the MSE.

Lemma 7 ((Vollmer et al., 2016)) *Under Assumption 1, and for a 1st-order numerical intergrator, the MSE is bounded, for a constant C independent of h and K, by*

$$\mathbb{E}\left(\hat{\psi}_K - \bar{\psi}\right)^2 \le C\left(\frac{1}{hK} + h^2\right)\ .$$

Furthermore, except for the 2nd-order Wasserstein distance defined in Lemma 1, we define the 1st-order Wasserstein distance between two probability measures μ_1 and μ_2 as

$$W_1\left(\mu_1,\mu_2\right) \triangleq \inf_{p\in\mathcal{P}(\mu_1,\mu_2)} \int \|\mathbf{x}-\mathbf{y}\|_2\, p(\mathrm{d}\,\mathbf{x},\mathrm{d}\,\mathbf{y})\ . \tag{14}$$

According to the Kantorovich-Rubinstein duality (Arjovsky et al., 2017), $W_1(\mu_1,\mu_2)$ is equivalently represented as

$$W_1\left(\mu_1,\mu_2\right) = \sup_{f\in\mathcal{L}_1} \mathbb{E}_{\mathbf{z}\sim\mu_1}\left[f(\mathbf{z})\right] - \mathbb{E}_{\mathbf{z}\sim\mu_2}\left[f(\mathbf{z})\right]\ , \tag{15}$$

where $\mathcal{L}_1$ is the space of 1-Lipschitz functions $f : \mathbb{R}^L \to \mathbb{R}$.

We have the following relation between $W_1(\mu_1,\mu_2)$ and $W_2(\mu_1,\mu_2)$.

Lemma 8 ((Givens & Shortt, 1984)) *We have for any two distributions μ_1 and μ_2 that $W_1(\mu_1,\mu_2) \le W_2(\mu_1,\mu_2)$.*

Now it is ready to prove Theorem 2.

Proof [Proof of Theorem 2] The idea is to simply decompose the MSE into two parts, with one part charactering the

MSE of the numerical method, the other part charactering the MSE of ρ_T and $p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})$, which consequentially can be bounded using Lemma 6 above.

Specifically, we have

$$
\begin{aligned}
&\mathrm{MSE}(\bar{\rho}_T,\rho_T;\psi) \triangleq \mathbb{E}\left(\int \psi(\mathbf{z})(\tilde{\rho}_T-\rho_T)(\mathbf{z})\mathrm{d}\,\mathbf{z}\right)^2\\
=&\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{z}_k)-\int\psi(\mathbf{z})\rho_T(\mathbf{z})\mathrm{d}\,\mathbf{z}\right)^2\\
=&\mathbb{E}\left(\left(\frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{z}_k)-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)\right.\\
&\left.-\left(\int\psi(\mathbf{z})\rho_T(\mathbf{z})\mathrm{d}\,\mathbf{z}-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)\right)^2\\
\overset{(1)}{=}&\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{z}_k)-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)^2\\
&+\left(\int\psi(\mathbf{z})\rho_T(\mathbf{z})\mathrm{d}\,\mathbf{z}-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)^2\\
\overset{(2)}{\leq}&\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{z}_k)-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)^2+W_1^2(\rho_T,p_{\boldsymbol{\theta}})\\
\overset{(3)}{\leq}&\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{z}_k)-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)^2+W_2^2(\rho_T,p_{\boldsymbol{\theta}})\\
\overset{(4)}{\leq}&C_1\left(\frac{1}{hK}+h^2\right)+W_2^2\left(\rho_0,p(\mathbf{x},\mathbf{z})\right)e^{-2CT}\\
=&O\left(\frac{1}{hK}+h^2+e^{-2ChK}\right),
\end{aligned}
$$

where "(1)" follows by the fact that $\mathbb{E}\left(\frac{1}{K}\sum_{k=1}^{K}\psi(\mathbf{z}_k)-\int\psi(\mathbf{z})p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})\mathrm{d}\,\mathbf{z}\right)=0$ (Chen et al., 2015); "(2)" follows by the definition of $W_1(\mu_1,\mu_2)$ in (14) and the 1-Lipschitz assumption of the test function ψ; "(3)" follows by Lemma 8; "(4)" follows by Lemma 6 and Lemma 7. ■

C. Sample Distance $\mathcal{D}$ Implemented as a Discriminator in the GAN Framework

We first prove Proposition 4, and then describe our implementation for the Wasserstein distance $\mathcal{D}$ in (8).

Proof [Proof of Proposition 4] By defining $\mathcal{D}$ as standard Euclidean distance, the objective becomes:

$$
\boldsymbol{\phi}'=\arg\min_{\boldsymbol{\phi}}\frac{1}{S}\sum_{i=1}^{S}\left\|\mathbf{z}_0'^{(i)}-\mathbf{z}_1^{(i)}\right\|^2,
$$

where $\{\mathbf{z}_0'^{(i)}\}_{i=1}^S$ are a set of samples generated from $q_{\boldsymbol{\phi}'}(\mathbf{z}_0'\,|\,\mathbf{x})$ via $Q_\phi(\cdot)$, *i.e.*

$$
\omega'^i\sim q_0(\omega),\;\;\tilde{\mathbf{z}}_0'^i=Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega'^i),
$$

and $\{\mathbf{z}_1^{(i)}\}_{i=1}^S$ are samples drawn by

$$
\omega^i\sim q_0(\omega),\;\;\tilde{\mathbf{z}}_0^i=Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega^i),\;\;\mathbf{z}_1^{(i)}\sim\mathcal{T}_1(\tilde{\mathbf{z}}_0^i).
$$

For simplicity, we consider $\mathcal{T}_1$ as one discretized step for Langevin dynamics, *i.e.*,

$$
\mathcal{T}_1(\tilde{\mathbf{z}}_0^i)=\tilde{\mathbf{z}}_0^i+\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}}_0^i)h+\sqrt{2h}\xi,
$$

where $\xi\sim\mathcal{N}(0,\mathbf{I})$. Consequently, the objective becomes

$$
\begin{aligned}
\tilde{F}\triangleq\frac{1}{S}\sum_{i=1}^{S}\Big\|&Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega'^i)-Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega^i)\\
&-\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}}_0^i)h+\sqrt{2h}\xi\Big\|^2,
\end{aligned}
\tag{16}
$$

(16) is a stochastic version of the following equivalent objective:

$$
\begin{aligned}
F\triangleq\mathbb{E}_{\omega',\omega\sim p_0(\omega),\xi}\Big\|&Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega'^i)-Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega^i)\\
&-\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}}_0^i)h+\sqrt{2h}\xi\Big\|^2.
\end{aligned}
\tag{17}
$$

There are two cases related to ω and ω'. *i*) If ω is restricted to be equal to ω', *e.g.*, they share the same random seed, this is the case in amortized SVGD (Wang & Liu, 2017) or amortized MCMC (Li et al., 2017b), as well as in Proposition 4 where Euclidean distance is adopted. *ii*) If Ω and Ω' do not share the same random seed, this is a more general case, which we also want to show that it can not learn a good generator.

For case *i*), F is simplified as:

$$
F=\mathbb{E}_{\omega\sim p_0(\omega)}\left\|\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}}_0^i)\right\|^2h^2+\sqrt{2h}\mathbb{E}_\xi\|\xi\|^2.
$$

Thus the minimum value corresponds to $\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}}_0^i)=0$, *i.e.*, $\boldsymbol{\phi}$ is updated so that $\tilde{\mathbf{z}}_0^i$ falls in one of the local modes of $p_{\boldsymbol{\theta}}(\mathbf{z}\,|\,\mathbf{x})$. Proposition 4 is proved.

We also want to consider case *ii*). In this case, F is bounded by

$$
\begin{aligned}
F\leq&\,\mathbb{E}_{\omega',\omega\sim p_0(\omega)}\left\|Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega'^i)-Q_{\boldsymbol{\phi}}(\cdot\,|\,\mathbf{x},\omega^i)\right\|^2\\
&+\mathbb{E}_{\omega\sim p_0(\omega)}\left\|\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}}_0^i)\right\|^2h^2+2h\mathbb{E}_\xi\|\xi\|^2
\end{aligned}
$$

The minimum possible value of the upper bound of F is achieved when $Q_{\boldsymbol{\phi}}$ matches all $\omega\sim p_0(\omega)$ to a fixed point $\tilde{\mathbf{z}}$, and also $\nabla_{\mathbf{z}}\log p_{\boldsymbol{\theta}}(\mathbf{x},\tilde{\mathbf{z}})=0$. This is also a special mode of $p_{\boldsymbol{\theta}}(\mathbf{z}\,|\,\mathbf{x})$ if exists.

To sum up, by defining $\mathcal{D}$ to be standard Euclidean distance, $Q_{\boldsymbol{\phi}}$ would generate samples from local modes of $p_{\boldsymbol{\theta}}(\mathbf{z}\,|\,\mathbf{x})$.

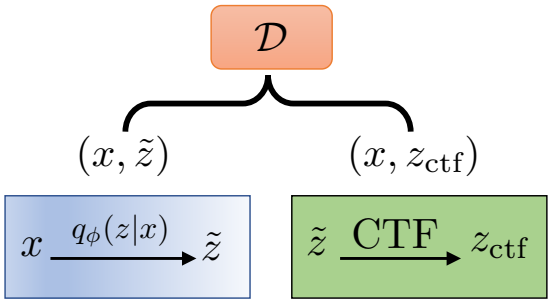

Figure 8. Implementation of $\mathcal{D}$ defined in (8) for distribution matching with the ALICE framework (Li et al., 2017a).

■

Now we describe how to define $\mathbb{D}$ as Wasserstein within a GAN framework. Following (Li et al., 2017a), we define a discriminator to match the joint distributions $p(\mathbf{x}, \mathbf{z}_{\text{ctf}})$ (an implicit distribution) and $q_\phi(\mathbf{x}, \tilde{\mathbf{z}})$, where

$$q_\phi(\mathbf{x}, \tilde{\mathbf{z}}) \triangleq q(\mathbf{x}) q_\phi(\tilde{\mathbf{z}} | \mathbf{x})$$
$$(\mathbf{x}, \mathbf{z}_{\text{ctf}}) \sim p(\tilde{\mathbf{x}}, \mathbf{z}), \text{ with } \mathbf{z}_{\text{ctf}} = \mathcal{T}_1(\tilde{\mathbf{z}}) .$$

The graphical structure is defined in Figure 8.

D. Two 2D Distributions

$\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2\}$: $p(\mathbf{z}) \propto e^{-U(\mathbf{z})}$.

The first distribution is

$$U(\mathbf{z}) \triangleq \frac{1}{2}(\frac{\|\mathbf{z}\| - 2}{0.4})^2 - \ln(e^{-\frac{1}{2}[\frac{\mathbf{z}_2 - 4}{2.0}]^2} + e^{-\frac{1}{2}[\frac{\mathbf{z}_2 + 2}{0.2}]^2})$$

The second distribution is

$$U(\mathbf{z}) \triangleq -\ln(e^{-\frac{1}{2}[\frac{\mathbf{z}_2 - w_1(\mathbf{z})}{0.35}]^2} + e^{-\frac{1}{2}[\frac{\mathbf{z}_2 - w_1(\mathbf{z}) + w_2(\mathbf{z})}{0.35}]^2})$$

where

$$w_2(\mathbf{z}) = \sin(\frac{2\pi\, \mathbf{a}_1}{4}), \text{ and } w_2(\mathbf{z}) = 3\exp(\frac{1}{2}\left[\frac{\mathbf{z}_1 - 1}{0.6}\right]^2)$$

E. Algorithm for Density Estimation with CTFs

Algorithm 1 illustrates the details updates for MacGAN.

F. Connection to WGAN

We derive the upper bound of the maximum likelihood estimator, which connects MacGAN to WGAN. Let p_r be the data distribution, rewrite our maximum likelihood objective as

$$\max \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i)$$
$$= \max \frac{1}{N} \sum_{i=1}^{N} \left(U(\mathbf{x}_i; \boldsymbol{\theta}) - \log \int e^{U(\mathbf{x}; \boldsymbol{\theta})} \mathrm{d}\, \mathbf{x} \right) .$$

The above maximum likelihood estimator can be bounded with Jensen's inequality as:

$$\max \frac{1}{N} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}_i) \tag{18}$$
$$\leq \max \mathbb{E}_{\mathbf{x} \sim p_r} [U(\mathbf{x}; \boldsymbol{\theta})] - \log \int \frac{e^{U(\mathbf{x}; \boldsymbol{\theta})}}{q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)} q_{\boldsymbol{\phi}}(\mathbf{x}; \omega) \mathrm{d}\, \mathbf{x}$$
$$\leq \max \mathbb{E}_{\mathbf{x} \sim p_r} [U(\mathbf{x}; \boldsymbol{\theta})] - \mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)} \left[\log \frac{e^{U(\mathbf{x}; \boldsymbol{\theta})}}{q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)} \right]$$
$$= \max \mathbb{E}_{\mathbf{x} \sim p_r} [U(\mathbf{x}; \boldsymbol{\theta})] - \mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)} [U(\mathbf{x}; \boldsymbol{\theta})] \tag{19}$$
$$- \mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)} [\log q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)] . \tag{20}$$

This results in the same objective form as WGAN except that our model does not restrict $U(\mathbf{x}; \boldsymbol{\theta})$ to be 1-Lipschitz functions and the objective has an extra constant term $\mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)} [\log q_{\boldsymbol{\phi}}(\mathbf{x}; \omega)]$ w.r.t. $\boldsymbol{\theta}$.

Now we prove Proposition 5.

Proof [Proof of Proposition 5] First it is clear that the equality in (18) is achieved if and only if

$$q_{\boldsymbol{\phi}}(\mathbf{x}; \omega) = p_{\boldsymbol{\theta}}(\mathbf{x}) \propto e^{U(\mathbf{x}; \boldsymbol{\theta})} .$$

From the description in Section 4 and (18), we know that $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ share the same objective function, which is an upper bound of the MLE in (18).

Algorithm 1 CTFs for generative models at the k-th iteration. $\mathcal{D}(\cdot, \cdot)$ is the same as (8).

Input: parameters from last step $\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\phi}^{(k-1)}$
Output: updated parameters $\boldsymbol{\theta}^{(k)}, \boldsymbol{\phi}^{(k)}$
1. Generate samples $\{\mathbf{x}_{1,s}\}_{s=1}^{S}$ via a discretized CTF: $\mathbf{x}_{0,s} \sim q_{\boldsymbol{\phi}^{(k-1)}}(\mathbf{x}_0), \mathbf{x}_{1,s} \sim \mathcal{T}_1(\mathbf{x}_{0,s})$;
2. Update the generator by minimizing ($\{\mathbf{x}'_{0,s}\}_{s=1}^{S}$ are generated with the updated parameter $\boldsymbol{\phi}^{(k)}$):

$$\boldsymbol{\phi}^{(k)} = \arg\min_{\boldsymbol{\phi}} \mathcal{D}\left(\{\mathbf{x}_{1,s}\}, \{\mathbf{x}'_{0,s}\}\right) .$$

3. Update the energy-based model $\boldsymbol{\theta}^k$ by maximum likelihood, with gradient as (9) except replacing $\mathbb{E}_{\mathbf{x} \sim p_{\boldsymbol{\theta}}(\mathbf{x})}$ with $\mathbb{E}_{\mathbf{x} \sim q_{\boldsymbol{\phi}}(\mathbf{x})}$;

Furthermore, based on the property of continuous-time flows (or formally Theorem 2), we know that $q_{\boldsymbol{\phi}}$ is learned such that $q_{\boldsymbol{\phi}} \to p_{\boldsymbol{\theta}}$ in the limit of $h \to 0$ (or alternatively, we could achieve this by using a decreasing-step-size sequence in a numerical method, as proved in (Chen et al., 2015)). When $q_{\boldsymbol{\phi}} = p_{\boldsymbol{\theta}}$, the equality in (18) is achieved, leading to the MLE. ■

G. Additional Experiments

G.1. Calculating the testing ELBO for MacVAE

We follow the method in (Pu et al., 2017) for calculating the ELBO for a test data $\mathbf{x}_*$. First, after distilling the CTF into the inference network $q_{\boldsymbol{\phi}}$, we have that the ELBO can be represented as

$$\log p(\mathbf{x}_*) \geq \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}_*, \mathbf{z}_*)\right] - \mathbb{E}_{q_{\boldsymbol{\phi}}}\left[\log q_{\boldsymbol{\phi}}\right] \ .$$

The expectation is approximated with samples $\{\mathbf{z}_{*j}\}_{j=1}^M$ with $\mathbf{z}_{*j} = f_{\boldsymbol{\phi}}(\mathbf{x}_*, \boldsymbol{\zeta}_j)$, and $\boldsymbol{\zeta}_j \sim q_0(\boldsymbol{\zeta})$ the standard isotropic normal. Here $f_{\boldsymbol{\phi}}$ represents the deep neural network in the inference network. Note $q_{\boldsymbol{\phi}}(\mathbf{z}_*)$ is not readily obtained. To evaluate it, we use the density transformation formula: $q_{\boldsymbol{\phi}}(\mathbf{z}_*) = q_0(\boldsymbol{\zeta}) \left|\det \frac{\partial f_{\boldsymbol{\phi}}(\mathbf{x}_*, \boldsymbol{\zeta})}{\partial \boldsymbol{\zeta}}\right|^{-1}$.

G.2. Network architecture

The architecture of the generator of MacGAN is given in Table 1.

G.3. Additional results

Additional experimental results are given in Figure 9 – 14.

G.4. Robustness of the discretization stepsize

To test the impact of the discretization stepsize h in (6), following SteinGAN (Feng et al., 2017), we test MacGAN on the MNIST dataset, where ee use a simple Gaussian-Bernoulli Restricted Boltzmann Machines as the energy-based model. We adopt the annealed importance sampling method to evaluate log-likelihoods (Feng et al., 2017). We vary h in $\{6e-4, 2.4e-3, 3.6e-3, 6e-3, 1e-2, 1.5e-2\}$. The trend of log-likelihoods is plotted in Figure 15. We can see that log-likelihoods do not change a lot within the chosen stepsize interval, demonstrating the robustness of h.

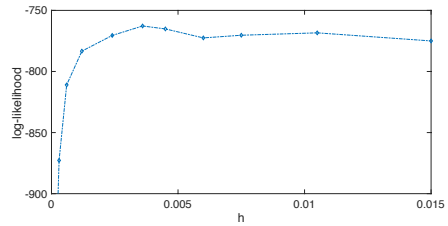

Figure 15. Log-likelihoods vs discretization stepsize for MacGAN on MNIST.

Table 1. Architecture of generator in MacGAN

Output Size	Architecture
100×1	100×10 Linear, BN, ReLU
$256 \times 8 \times 8$	$512 \times 4 \times 4$ deconv, 256 5×5 kernels, ReLU, strike 2, BN
$128 \times 16 \times 16$	$256 \times 8 \times 8$ deconv, 128 5×5 kernels, ReLU, strike 2, BN
$3 \times 32 \times 32$	$128 \times 16 \times 16$ deconv, 3 5×5 kernels, Tanh, strike 2

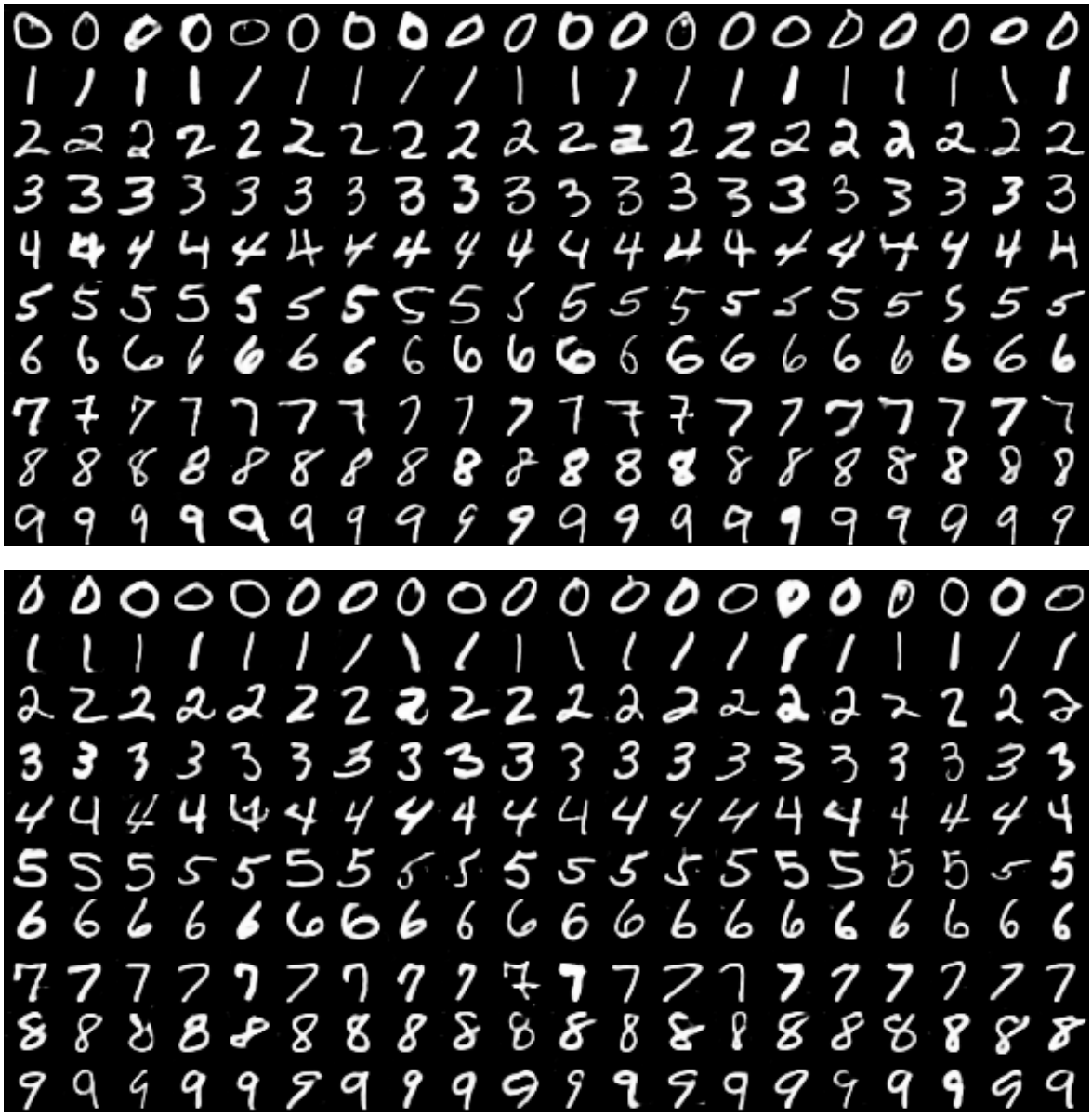

Figure 9. Generated images for MNIST datasets with MacGAN (top) and SteinGAN (bottom).

Figure 10. Generated images for CelebA datasets with MacGAN.

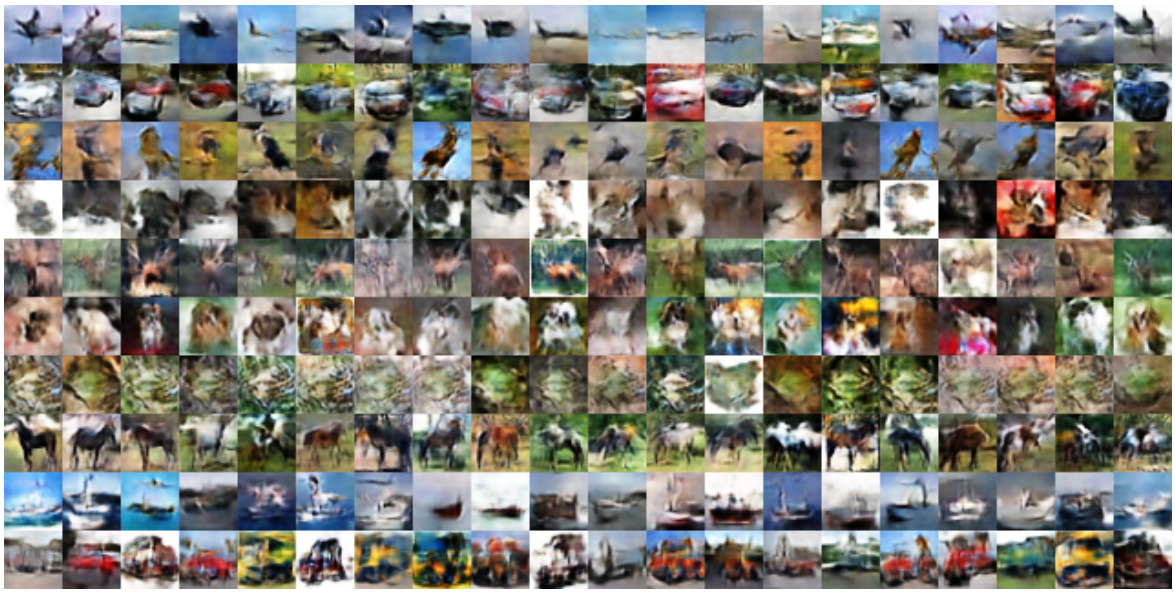

Figure 11. Generated images for CIFAR-10 datasets with MacGAN.

Figure 12. Generated images for CelebA datasets with SteinGAN.

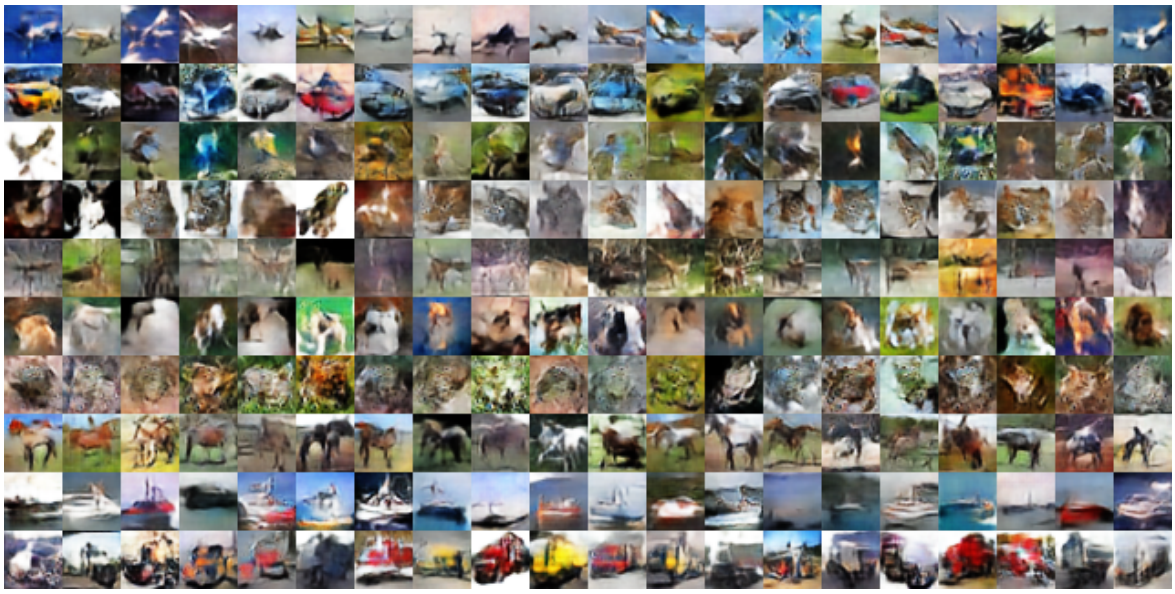

Figure 13. Generated images for CIFAR-10 datasets with SteinGAN.

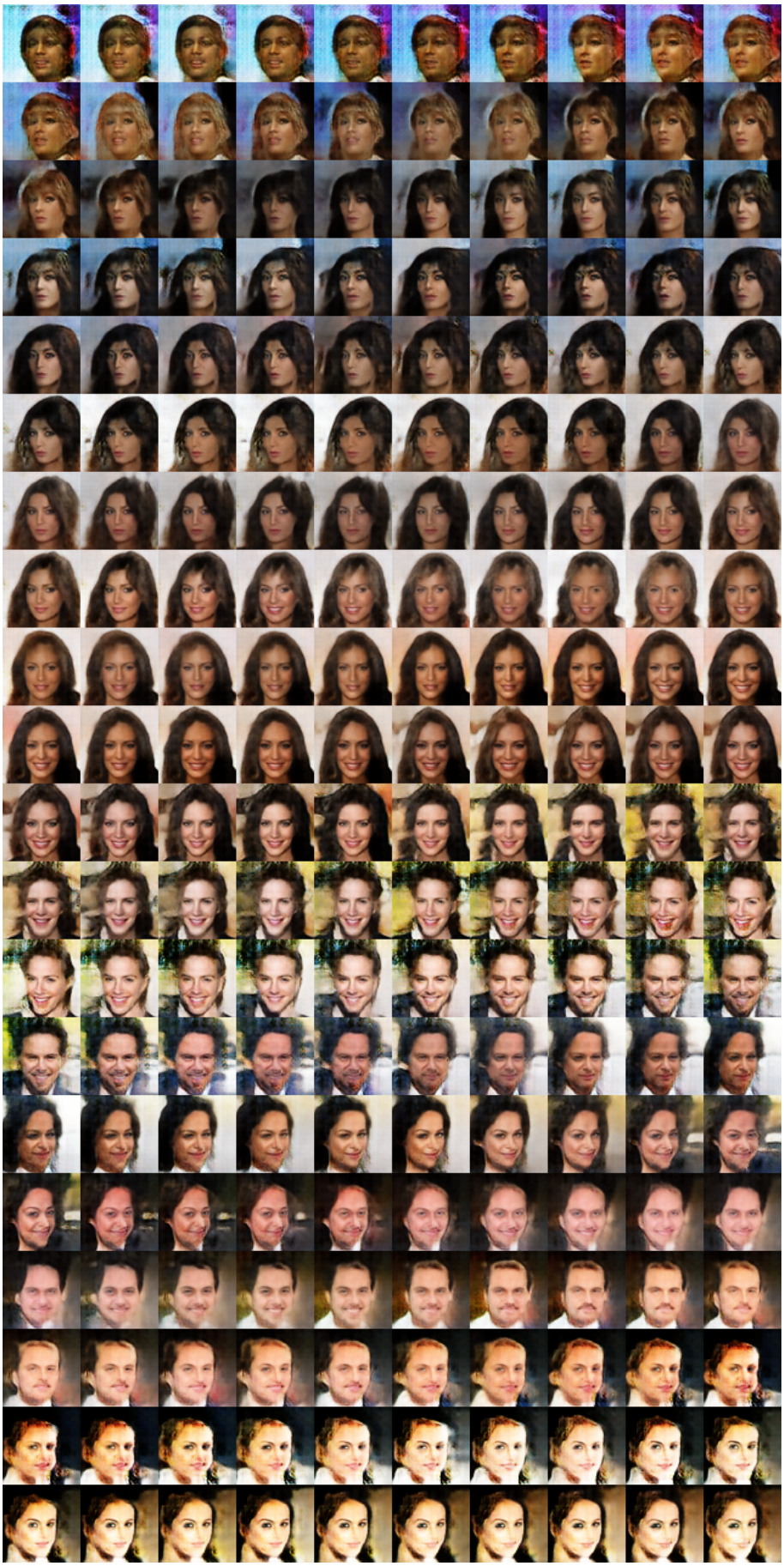

Figure 14. Generated images with a random walk on the ω space for CelebA datasets with MacGAN, $\omega_t = \omega_{t-1} + 0.02 \times \text{rand}([-1, 1])$.