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 $f : \mathbb{R}^L \to \mathbb{R}$, such that

$$\mathcal{L} f(Z_t) = \lim_{h \to 0^+} \frac{1}{h} [E[f(Z_{t+h})] - f(Z_t)] = \langle F(Z_t) \cdot \nabla + \frac{1}{2} \langle G(Z_t) G(Z_t)^T \rangle : \nabla \nabla^T \rangle f(Z_t),$$

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

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

$$\mathcal{L} \hat{\psi}(z_k) = \psi(z_k) - \tilde{\psi}$$ (12)

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

Assumption 1 $\tilde{\psi}$ exists, and its up to 4th-order derivatives, $\mathcal{D}^k \tilde{\psi}$, are bounded by a function $\mathcal{V}$, i.e., $\|D^k \tilde{\psi}\| \leq C_k \mathcal{V}(z)$ for $k = 0, 1, 2, 3, 4$, $C_k, p_k > 0$. Furthermore, the expectation of $\mathcal{V}$ on $(z_k)$ is bounded: $\sup \mathbb{E}[\mathcal{V}(z_k)] < \infty$, and $\mathcal{V}$ is smooth such that $\sup s \in (0, 1) \mathcal{V}(z + (1-s)y) \leq C \mathcal{V}(z) + \mathcal{V}(y), \forall z, y, p \leq \max \{2p_k\}$ 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 $\mathcal{F}(p)$ in (Jordan et al., 1998) is in the form of $\text{KL}(\rho||\rho_0(x, z))$ 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 $\rho_T$ in the limit of $h \to 0$.

Since this is true for each $k$ (thus each $t$ in $\rho_t$), we conclude that $\rho_k = \rho_{hk}$ 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 $\rho_T$ be the solution of the FP equation (3) at time $T$, $p_0(x, z)$ be the joint posterior distribution given $x$. Assume that $\int \rho_T(z)p_0^{-1}(x, z)d\mathcal{Z} < \infty$ and there exists a constant $\mathcal{C}$ such that $\frac{dW_2^2(\rho_T, p_0(x, z))}{dt} \geq \mathcal{C} W_2^2(\rho_T, p_0(x, z))$. Then

$$W_2(\rho_T, p(x, z)) \leq W_2(p_0, p(x, z)) e^{-\mathcal{C} T}.$$ (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 $\psi(z)$, defined as $\tilde{\psi} = \int \psi(z) p_0(x, z) dZ$. When using a numerical integrator to solve (2) to get samples $\{z_k\}_{k=1}^K$, the sample average $\tilde{\psi}_K = \frac{1}{K} \sum_{k=1}^K \psi(z_k)$ is used to approximate the posterior average. The accuracy is characterized by the mean square error (MSE) defined as: $\mathbb{E}[(\tilde{\psi}_K - \tilde{\psi})^2]$. Lemma 7 derives the bound for the MSE.

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

$$\mathbb{E}[(\tilde{\psi}_K - \tilde{\psi})^2] \leq C \left( \frac{1}{h K} + h^2 \right).$$

Furthermore, except for the 2nd-order Wasserstein distance defined in Lemma 1, we define the 1st-order Wasserstein distance between two probability measures $\mu_1$ and $\mu_2$ as

$$W_1(\mu_1, \mu_2) \triangleq \inf_{p \in P(\mu_1, \mu_2)} \int \|x - y\|_2 p(dx, dy).$$ (14)

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

$$W_1(\mu_1, \mu_2) = \sup_{f \in \mathcal{L}_1} \mathbb{E}_{\mu_1} [f(z)] - \mathbb{E}_{\mu_2} [f(z)],$$ (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 $\mu_1$ and $\mu_2$ that $W_1(\mu_1, \mu_2) \leq W_2(\mu_1, \mu_2)$.

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

**Proof [Proof of Proposition 4]** By defining \( D \) as standard Euclidean distance, the objective becomes:

\[
\phi' = \arg \min_{\phi} \frac{1}{S} \sum_{i=1}^{S} \left\| z_0^{(i)} - z_1^{(i)} \right\|^2,
\]

where \( \{z_0^{(i)}\}_{i=1}^{S} \) is a set of samples generated from \( q_{\phi'}(z_0 | x) \) via \( Q_{\phi}(\cdot) \), i.e.

\[
\omega^{(i)} \sim q_0(\omega), \quad \bar{z}_0^{(i)} = Q_{\phi}(\cdot | x, \omega^{(i)}),
\]

and \( \{z_1^{(i)}\}_{i=1}^{S} \) are samples drawn by

\[
\omega^{(i)} \sim q_0(\omega), \quad \bar{z}_0^{(i)} = Q_{\phi}(\cdot | x, \omega^{(i)}), \quad z_1^{(i)} \sim \mathcal{T}_1(\bar{z}_0^{(i)}).
\]

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

\[
\mathcal{T}_1(\bar{z}_0^{(i)}) = \bar{z}_0^{(i)} + \nabla \log p_\theta(x, \bar{z}_0^{(i)}) h + \sqrt{2h} \xi,
\]

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

\[
\hat{F} \triangleq \frac{1}{S} \sum_{i=1}^{S} \left\| Q_{\phi}(\cdot | x, \omega^{(i)}) - Q_{\phi}(\cdot | x, \omega^{(i)}) \right\| - \nabla \log p_\theta(x, \bar{z}_0^{(i)}) h + \sqrt{2h} \xi \|^2,
\]

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

\[
\hat{F} \triangleq \mathbb{E}_{\omega \sim p_0(\omega), \xi} \left\| Q_{\phi}(\cdot | x, \omega^{(i)}) - Q_{\phi}(\cdot | x, \omega^{(i)}) \right\| - \nabla \log p_\theta(x, \bar{z}_0^{(i)}) h + \sqrt{2h} \xi \|^2.
\]

(17)

There are two cases related to \( \omega \) and \( \omega' \). If \( \omega \) is restricted to be equal to \( \omega' \), 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. If \( \Omega \) and \( \Omega' \) 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:

\[
F = \mathbb{E}_{\omega \sim p_0(\omega)} \left\| \nabla \log p_\theta(x, \bar{z}_0^{(i)}) h + \sqrt{2h} \xi \|^2.
\]

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

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

\[
\hat{F} \leq \mathbb{E}_{\omega \sim p_0(\omega)} \left\| Q_{\phi}(\cdot | x, \omega^{(i)}) - Q_{\phi}(\cdot | x, \omega^{(i)}) \right\|^2 + \mathbb{E}_{\omega \sim p_0(\omega)} \left\| \nabla \log p_\theta(x, \bar{z}_0^{(i)}) h + \sqrt{2h} \xi \|^2.
\]

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

To sum up, by defining \( D \) to be standard Euclidean distance, \( Q_{\phi} \) would generate samples from local modes of \( p_\theta(z | x) \).
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 $\mathcal{D}$ as Wasserstein within a GAN framework. Following (Li et al., 2017a), we define a discriminator to match the joint distributions $p(x, z_{ctf})$ (an implicit distribution) and $q_\phi(x, \tilde{z})$, where

$$q_\phi(x, \tilde{z}) \triangleq q(z)|q_\phi(\tilde{z} \mid x)$$

$(x, z_{ctf}) \sim p(x, z)$, with $z_{ctf} = T_1(\tilde{z})$.

The graphical structure is defined in Figure 8.

D. Two 2D Distributions

$z = \{z_1, z_2\}$; $p(z) \propto e^{-U(z)}$.

The first distribution is

$$U(z) \triangleq \frac{1}{2} \left( \frac{\|z\|}{0.4} - 2 \right)^2 - \ln(e^{-\frac{1}{2}(\frac{z_2}{0.4})^2} + e^{-\frac{1}{2}(\frac{z_1}{0.6})^2})$$

The second distribution is

$$U(z) \triangleq \ln(e^{-\frac{1}{2}(\frac{z_2 - \frac{1}{2}(z_1)}{0.4})^2} + e^{-\frac{1}{2}(\frac{z_1 - \frac{1}{2}(z_2)}{0.6})^2})$$

where

$$w_2(z) = \sin(\frac{2\pi a_1}{4}), \quad w_2(z) = 3 \exp(\frac{1}{2} \left( \frac{z_2 - 1}{0.6} \right)^2)$$

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

$$\max \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i)$$

$$= \max \frac{1}{N} \sum_{i=1}^{N} \left( U(x_i; \theta) - \log \int e^{U(x; \theta)} \, dx \right)$$

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

$$\max \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(x_i)$$

$$\leq \max E_{\sim p_r}[U(x; \theta)] - \log \int \frac{e^{U(x; \theta)}}{q_\phi(x; \omega)} q_\phi(x; \omega) \, dx$$

$$\leq \max E_{\sim p_r}[U(x; \theta)] - E_{\sim q_\phi(x; \omega)} \left[ \log \frac{e^{U(x; \theta)}}{q_\phi(x; \omega)} \right] \quad (19)$$

$$= E_{\sim q_\phi(x; \omega)} \left[ U(x; \theta) \right] - E_{\sim q_\phi(x; \omega)} \left[ U(x; \theta) \right]$$

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

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

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

From the description in Section 4 and (18), we know that $\theta$ and $\phi$ share the same objective function, which is an upper bound of the MLE in (18).
Furthermore, based on the property of continuous-time flows (or formally Theorem 2), we know that $q_\phi$ is learned such that $q_\phi \to p_\theta$ 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 $q_\phi = p_\theta$, 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 $x_*$. First, after distilling the CTF into the inference network $q_\phi$, we have that the ELBO can be represented as

$$\log p(x_*) \geq \mathbb{E}_{q_\phi} [\log p_\theta(x_*, z_*)] - \mathbb{E}_{q_\phi} [\log q_\phi] .$$

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

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 we 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>
<thead>
<tr>
<th>Output Size</th>
<th>Architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>$100 \times 1$</td>
<td>$100 \times 10$ Linear, BN, ReLU</td>
</tr>
<tr>
<td>$256 \times 8 \times 8$</td>
<td>$512 \times 4 \times 4$ deconv, $256 \times 5 \times 5$ kernels, ReLU, strike 2, BN</td>
</tr>
<tr>
<td>$128 \times 16 \times 16$</td>
<td>$256 \times 8 \times 8$ deconv, $128 \times 5 \times 5$ kernels, ReLU, strike 2, BN</td>
</tr>
<tr>
<td>$3 \times 32 \times 32$</td>
<td>$128 \times 16 \times 16$ deconv, $3 \times 5 \times 5$ kernels, Tanh, strike 2</td>
</tr>
</tbody>
</table>

*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 $\omega$ space for CelebA datasets with MacGAN, $\omega_t = \omega_{t-1} + 0.02 \times \text{rand}([-1, 1])$. 

Continuous-Time Flows