Appendix

A. Training objective as an upper bound

Claim 1 *Let $\mathbb{P}_d$ and $\mathbb{P}_f$ be two distributions. Suppose that $\hat{\mathbb{P}}_d$ and $\hat{\mathbb{P}}_f$ are empirical measures of $\mathbb{P}_d$ and $\mathbb{P}_f$, induced by random sets (of n i.i.d samples) $\mathcal{D}$ and $\mathcal{F}$. Then*

$$\tilde{W}_2^2(\mathbb{P}_d, \mathbb{P}_f) \leq 16\mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)]. \tag{16}$$

Proof: Using the triangle inequality for the sliced Wasserstein distance, we have

$$\tilde{W}_2^2(\mathbb{P}_d, \mathbb{P}_f) \leq 2\tilde{W}_2^2(\mathbb{P}_d, \hat{\mathbb{P}}_d) + 2\tilde{W}_2^2(\mathbb{P}_f, \hat{\mathbb{P}}_d). \tag{17}$$

Using it again, we get

$$\tilde{W}_2^2(\mathbb{P}_d, \mathbb{P}_f) \leq 2\tilde{W}_2^2(\mathbb{P}_d, \hat{\mathbb{P}}_d) + 4\tilde{W}_2^2(\mathbb{P}_f, \hat{\mathbb{P}}_f) + 4\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f). \tag{18}$$

In the following we find upper bounds for $\tilde{W}_2^2(\mathbb{P}_f, \hat{\mathbb{P}}_f)$ in terms of $\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)$. In order to do this, we must deconstruct the sliced Wasserstein distance. By definition, we have

$$\tilde{W}_2^2(\mathbb{P}_f, \hat{\mathbb{P}}_f) = \int_{\omega \in \Omega} W_2^2(\mathbb{P}_f^\omega, \hat{\mathbb{P}}_f^\omega) d\omega. \tag{19}$$

Consider any one projection ω. We have a 1-d distribution $\mathbb{P}_f^\omega$, and its empirical measure $\hat{\mathbb{P}}_f^\omega$. Using Theorem 4.3 in [5]:

$$\mathbb{E}[W_2^2(\mathbb{P}_f^\omega, \hat{\mathbb{P}}_f^\omega)] \leq \mathbb{E}[W_2^2(\hat{\mathbb{P}}_f^\omega, \hat{\mathbb{P}}_f'^\omega)], \tag{20}$$

where $\hat{\mathbb{P}}_f'^\omega$ is an independent copy of $\hat{\mathbb{P}}_f^\omega$.

To bound $\mathbb{E}[W_2^2(\hat{\mathbb{P}}_f^\omega, \hat{\mathbb{P}}_f'^\omega)]$ in Eq. (20), we first see how the expectsed Wasserstein distance between two 1-d empirical measures $\hat{\mathbb{P}}_d^\omega$ and $\hat{\mathbb{P}}_f^\omega$ can be written in terms of the sets of samples $\mathcal{D}^\omega$ and $\mathcal{F}^\omega$ that they represent (i.e. are induced by). Note that $\mathcal{D}^\omega$ and $\mathcal{F}^\omega$ are obtained by simply projecting a the sets $\mathcal{D}$ and $\mathcal{F}$ onto the direction ω. If $\mathcal{D}^\omega_{\sigma_D(i)}$ and $\mathcal{F}^\omega_{\sigma_F(i)}$ denote the i-th smallest sample in $\mathcal{D}^\omega$ and $\mathcal{F}^\omega$,

$$\mathbb{E}[W_2^2(\hat{\mathbb{P}}_d^\omega, \hat{\mathbb{P}}_f^\omega)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\mathcal{D}^\omega_{\sigma_D(i)} - \mathcal{F}^\omega_{\sigma_F(i)})^2]. \tag{21}$$

$\mathcal{D}^\omega_{\sigma_D(i)}$ and $\mathcal{F}^\omega_{\sigma_F(i)}$ are infact the n sample order statistics of $\mathbb{P}_d^\omega$ and $\mathbb{P}_f^\omega$. For $\hat{\mathbb{P}}_f^\omega$ and $\hat{\mathbb{P}}_f'^\omega$, we can write this as

$$\mathbb{E}[W_2^2(\hat{\mathbb{P}}_f^\omega, \hat{\mathbb{P}}_f'^\omega)] = \frac{2}{n}\sum_{i=1}^n Var[\mathcal{F}^\omega_{\sigma_F(i)}]. \tag{22}$$

The RHS of Eq. (21) can be decomposed as

$$\begin{aligned}
&\mathbb{E}[(\mathcal{D}^\omega_{\sigma_D(i)} - \mathcal{F}^\omega_{\sigma_F(i)})^2] \\
&= \mathbb{E}[(\mathcal{D}^\omega_{\sigma_D(i)} - \mathbb{E}[\mathcal{F}^\omega_{\sigma_F(i)}] + \mathbb{E}[\mathcal{F}^\omega_{\sigma_F(i)}] - \mathcal{F}^\omega_{\sigma_F(i)})^2] \\
&= Var[\mathcal{F}^\omega_{\sigma_F(i)}] + E[(\mathcal{D}^\omega_{\sigma_D(i)} - \mathbb{E}[\mathcal{F}^\omega_{\sigma_F(i)}])^2] \\
&\geq Var[\mathcal{F}^\omega_{\sigma_F(i)}],
\end{aligned}$$

hence

$$\frac{1}{n}\sum_{i=1}^n Var[\mathcal{F}^\omega_{\sigma(i)}] \leq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(\mathcal{D}^\omega_{\sigma_D(i)} - \mathcal{F}^\omega_{\sigma_F(i)})^2].$$

Combining this result with Eq. (21) and Eq. (22) yields

$$\mathbb{E}[W_2^2(\hat{\mathbb{P}}_f^\omega, \hat{\mathbb{P}}_f'^\omega)] \leq 2\mathbb{E}[W_2^2(\hat{\mathbb{P}}_d^\omega, \hat{\mathbb{P}}_f^\omega)],$$

which, when combined with Eq. (20), results in

$$\mathbb{E}[W_2^2(\mathbb{P}_f^\omega, \hat{\mathbb{P}}_f^\omega)] \leq 2\mathbb{E}[W_2^2(\hat{\mathbb{P}}_d^\omega, \hat{\mathbb{P}}_f^\omega)]. \tag{23}$$

Applying the expectation operator on Eq. (19) and using Eq. (23),

$$\begin{aligned}\mathbb{E}[\tilde{W}_2^2(\mathbb{P}_f, \hat{\mathbb{P}}_f)] &\leq 2\int_{\omega\in\Omega} \mathbb{E}[W_2^2(\hat{\mathbb{P}}_d^\omega, \hat{\mathbb{P}}_f^\omega)]d\omega \\ &= 2\mathbb{E}[\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)].\end{aligned} \tag{24}$$

The same bound holds for $\mathbb{E}[\tilde{W}_2^2(\mathbb{P}_d, \hat{\mathbb{P}}_d)]$.

Substituting from Eq. (24) in Eq. (18) and applying the expectation operator, we get

$$\tilde{W}_2^2(\mathbb{P}_d, \mathbb{P}_f) \leq 16\mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)], \tag{25}$$

which completes the proof. ■

B. Bounds for generated distribution

Corollary 1 *Let $\mathbb{P}_d$ and $\mathbb{P}_f$ be two distributions. Suppose that $\hat{\mathbb{P}}_d$ and $\hat{\mathbb{P}}_f$ are (n-sample) empirical measures of $\mathbb{P}_d$ and $\mathbb{P}_f$, and let $\hat{\mathbb{P}}'_d$ be an independent copy of $\hat{\mathbb{P}}_d$. For $\mathbb{P}_f^*$ defined by $\mathbb{P}_f^* = \operatorname{argmin}_{\mathbb{P}_f}\ \mathbb{E}[\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)]$, the following holds:*

$$\tilde{W}_2(\mathbb{P}_d, \mathbb{P}_f^*) \leq 14\mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}'_d)]. \tag{26}$$

Proof: This follows easily from Claim 1. Using Eq. (20), we can show that

$$\mathbb{E}[\tilde{W}_2^2(\mathbb{P}_d, \hat{\mathbb{P}}_d)] \leq \mathbb{E}[\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}'_d)], \tag{27}$$

and therefore we can rewrite (18) as:

$$\tilde{W}_2(\mathbb{P}_d, \mathbb{P}_f) \leq 2\mathbb{E}[\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}'_d)] + 12\mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)]. \tag{28}$$

Since $\mathbb{P}_f^*$ minimizes $\mathbb{E}[\tilde{W}_2^2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f)]$ over all $\mathbb{P}_f$,

$$\mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}_f^*)] \leq \mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}'_d)]. \tag{29}$$

Therefore,

$$\tilde{W}_2(\mathbb{P}_d, \mathbb{P}_f^*) \leq 14\mathbb{E}[\tilde{W}_2(\hat{\mathbb{P}}_d, \hat{\mathbb{P}}'_d)]. \tag{30}$$

■

C. Discriminator update frequency experiments

We tested different discriminator update schemes (*i.e.*, number of generator updates per discriminator updates, and number of iterations of discriminator updates). In Tab. 4 we show samples after 40 epochs of training on the LSUN dataset with these different schemes for two discriminator configurations. The generator architecture for both is the DCGAN.

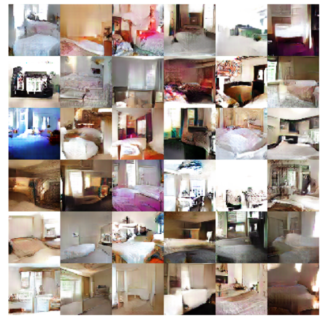
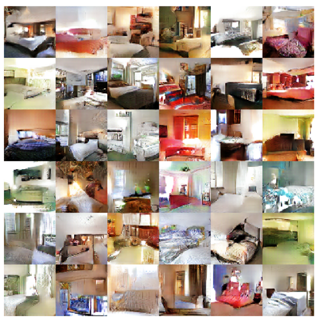
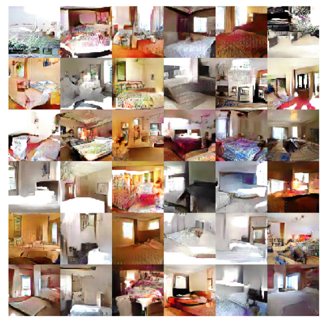
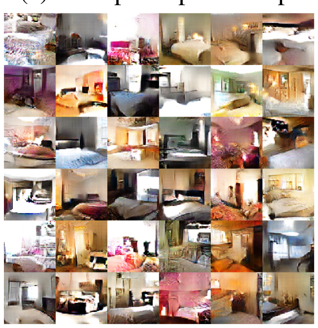
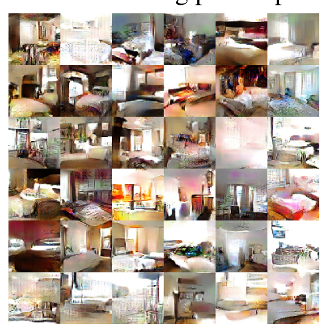
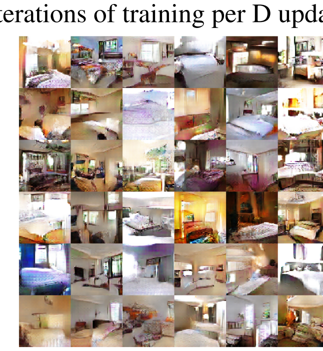
Discriminator:**DCGAN**	**DCGAN with 64 filters in each layer**
(a) 1 D update per G update, 1 iteration of training per D update	
	
(b) 1 D update per G update, 5 iterations of training per D update	
	
(c) 1 D update per 5 G updates, 1 iteration of training per D update	
	
(d) 1 D update per 5 G updates, 5 iterations of training per D update	
	

Table 4. The SWG is robust to different discriminator update schemes. Tested for two discriminator architectures (columns). Sample size $= 64$, learning rate $= 0.0005$, Adam optimizer, 40 epochs.

D. Network architectures for experiments on MNIST

Here we summarize the different network architectures used for experiments with the MNIST dataset presented in Sec. 4.2.

Generator (Fully Connected)	Generator (Conv & Deconv)	Discriminator
output: 784-d sample	**output:** 784-d sample	**output:** scalar
fc-784, sigmoid	conv2d-1-3-1, sigmoid	2× fc-256, relu
7× fc-512, relu	deconv2d-16-3-2, (bn), relu	**input:** 784-d sample
input: 32-d random noise	conv2d-32-3-1, (bn), relu	
	deconv2d-32-3-2, (bn), relu	
	conv2d-64-3-1, (bn), relu	
	deconv2d-64-3-2, (bn), relu	
	fc-1024	
	input: 32-d random noise	

Table 5. Generator and discriminator for MNIST. "fc-n" means applying a fully connected layer with n output units. Both "conv2d-c-k-s" and "deconv2d-c-k-s" mean applying c convolutional filters of size k by k with stride s by s. "bn" means batch normalization.