Supplement to "Markovian Sliced Wasserstein Distances: Beyond Independent Projections"

In this supplementary material, we present additional materials in Appendix A. In particular, we provide additional background on sliced Wasserstein variants in Appendix A.1, background on von Mises-Fisher distribution in Appendix A.2, algorithms for computing Markovian sliced Wasserstein distances in Appendix A.3, additional information about burned thinned MSW in Appendix A.4, and discussion on related works in Appendix A.5. We then provide skipped proofs in the main paper in Appendix B. Additional experiments are presented in Appendix C.

A Additional Materials

A.1 Background on Sliced Wasserstein Variants

We review computational aspects of sliced Wasserstein variants.

Computation of Max sliced Wasserstein distance. We demonstrate the empirical estimation of Max-SW via projected sub-gradient ascent algorithm in Algorithm 1. The initialization step for $\hat{\theta}_0$ is rarely discussed in previous works. Normally, $\hat{\theta}_0$ is randomly initialized by drawing from the uniform distribution over the unit-hypersphere. Many previous works [25, 43, 44, 41] use Adam update instead of the standard gradient ascent update for Max-SW. In this work, we find out that using the standard gradient ascent update is more stable and effective.

Algorithm 1 Max sliced Wasserstein distance

Input. Probability measures μ, ν, learning rate η, the order p, and the number of iterations T.
Initialize $\hat{\theta}_0$.
for $t = 1$ to $T - 1$ **do**
  $\hat{\theta}_t = \hat{\theta}_{t-1} + \eta \cdot \nabla_{\hat{\theta}_{t-1}} \mathrm{W}_p(\hat{\theta}_{t-1}\sharp\mu, \hat{\theta}_{t-1}\sharp\nu)$
  $\hat{\theta}_t = \frac{\hat{\theta}_t}{||\hat{\theta}_t||_2}$
end for
Return. $\mathrm{W}_p(\hat{\theta}_T\sharp\mu, \hat{\theta}_T\sharp\nu)$

K sliced Wasserstein distance. We first review the Gram–Schmidt process in Algorithm 2. With the Gram–Schmidt process, the sampling from $\mathcal{U}(\mathbb{V}_K(\mathbb{R}^d))$ can be done by sampling $\theta_1, \ldots, \theta_k$ i.i.d from $\mathcal{N}(0, I_d)$ then applying the Gram-Schmidt process on them. Therefore, we present the computation of K sliced Wasserstein distance in Algorithm 3. We would like to recall that the original work of K-SW [49] uses only one set of orthogonal projecting directions. Here, we generalize the original work by using L sets of orthogonal projecting directions.

Algorithm 2 Gram–Schmidt process

Input. K vectors $\theta_1, \ldots, \theta_K$
$\theta_1 = \frac{\theta_1}{||\theta_1||_2}$
for $k = 2$ to K **do**
  for $i = 1$ to $k - 1$ **do**
    $\theta_k = \theta_k - \frac{\langle \theta_i, \theta_k \rangle}{\langle \theta_i, \theta_i \rangle} \theta_i$
  end for
  $\theta_k = \frac{\theta_k}{||\theta_k||_2}$
end for
Return. $\theta_1, \ldots, \theta_K$

Max K sliced Wasserstein distance. We now present the empirical estimation of Max-K-SW via projected sub-gradient ascent algorithm in Algorithm 4. This algorithm is first discussed in the original paper of Max-K-SW [12]. The optimization of Max-K-SW can be solved by using Riemannian optimization since the Stiefel manifold is a Riemannian manifold. However, to the best of our knowledge, Riemannian optimization has not been applied to Max-K-SW.

Algorithm 3 K sliced Wasserstein distance

Input. Probability measures μ, ν, the dimension d, the order p, the number of projections L, the number of orthogonal projections K.
for $l = 1$ to L **do**
  Draw $\theta_{l1}, \dots, \theta_{lK}$ i.i.d from $\mathcal{N}(0, I_d)$.
  $\theta_{l1}, \dots, \theta_{lK} =$ Gram–Schmidt$(\theta_{l1}, \dots, \theta_{lK})$
end for
Return. $\left(\frac{1}{LK}\sum_{l=1}^{L}\sum_{k=1}^{K} \mathrm{W}_p^p(\theta_{lk}\sharp\mu, \theta_{lk}\sharp\nu)\right)^{\frac{1}{p}}$

Algorithm 4 Max-K sliced Wasserstein distance

Input. Probability measures μ, ν, learning rate η, the dimension d, the order p, the number of iterations $T > 1$, and the number of orthogonal projections $K > 1$.
Initialize $\hat{\theta}_{01}, \dots, \hat{\theta}_{0K}$ to be orthogonal.
for $t = 1$ to $T - 1$ **do**
  for $k = 1$ to K **do**
    $\hat{\theta}_{tk} = \theta_{tk} + \eta \cdot \nabla_{\hat{\theta}_{t-1k}} \mathrm{W}_p(\hat{\theta}_{t-1k}\sharp\mu, \hat{\theta}_{t-1k}\sharp\nu)$
  end for
  $\hat{\theta}_{t1}, \dots, \hat{\theta}_{tK} =$ Gram-Schmidt$(\hat{\theta}_{t1}, \dots, \hat{\theta}_{tK})$
end for
Return. $\left(\frac{1}{K}\sum_{k=1}^{K} \mathrm{W}_p^p(\hat{\theta}_{Tk}\sharp\mu, \hat{\theta}_{Tk}\sharp\nu)\right)^{\frac{1}{p}}$

A.2 Von Mises-Fisher Distribution

We first start with the definition of von Mises-Fisher (vMF) distribution.

Algorithm 5 Sampling from vMF distribution

Input. location ϵ, concentration κ, dimension d, unit vector $e_1 = (1, 0, .., 0)$
Draw $v \sim \mathcal{U}(\mathbb{S}^{d-2})$
$b \leftarrow \frac{-2\kappa+\sqrt{4\kappa^2+(d-1)^2}}{d-1}$, $a \leftarrow \frac{(d-1)+2\kappa+\sqrt{4\kappa^2+(d-1)^2}}{4}$, $m \leftarrow \frac{4ab}{(1+b)} - (d-1)\log(d-1)$
repeat
  Draw $\psi \sim \mathrm{Beta}\left(\frac{1}{2}(d-1), \frac{1}{2}(d-1)\right)$
  $\omega \leftarrow h(\psi, \kappa) = \frac{1-(1+b)\psi}{1-(1-b)\psi}$
  $t \leftarrow \frac{2ab}{1-(1-b)\psi}$
  Draw $u \sim \mathcal{U}([0, 1])$
until $(d-1)\log(t) - t + m \geq \log(u)$
$h_1 \leftarrow (\omega, \sqrt{1-\omega^2}v^\top)^\top$
$\epsilon' \leftarrow e_1 - \epsilon$
$u = \frac{\epsilon'}{||\epsilon'||_2}$
$U = I - 2uu^\top$
Output. Uh_1

Definition 3. *The von Mises–Fisher distribution (*vMF*)[22] is a probability distribution on the unit hypersphere $\mathbb{S}^{d-1}$ with the density function be:*

$$f(x|\epsilon, \kappa) := C_d(\kappa)\exp(\kappa\epsilon^\top x), \tag{2}$$

where $\epsilon \in \mathbb{S}^{d-1}$ is the location vector, $\kappa \geq 0$ is the concentration parameter, and $C_d(\kappa) := \frac{\kappa^{d/2-1}}{(2\pi)^{d/2} I_{d/2-1}(\kappa)}$ is the normalization constant. Here, I_v is the modified Bessel function of the first kind at order v [55].

The vMF distribution is a continuous distribution, its mass concentrates around the mean ϵ, and its density decrease when x goes away from ϵ. When $\kappa \to 0$, vMF converges in distribution to $\mathcal{U}(\mathbb{S}^{d-1})$, and when $\kappa \to \infty$, vMF converges in distribution to the Dirac distribution centered at ϵ [54].

Sampling: We review the sampling process in Algorithm 5 [13, 44]. The sampling process of vMF distribution is based on the rejection sampling procedure. It is worth noting that the sampling algorithm is doing reparameterization implicitly. However, we only use the algorithm to obtain random samples without estimating stochastic gradients.

A.3 Algorithms for Computing Markovian Sliced Wasserstein Distances

We first start with the general computation of MSW in Algorithm 6. For the orthogonal-based transition in oMSW, we use $\theta_{lt} \sim \mathcal{U}(\mathcal{S}^{d-1}_{\theta_{lt-1}})$ by first sampling $\theta'_{lt} \sim \mathcal{U}(\mathbb{S}^{d-1})$ then set $\theta_{lt} = \theta_{lt} - \frac{\langle \theta'_{lt}, \theta_{lt} \rangle}{\langle \theta'_{lt}, \theta'_{lt} \rangle} \theta'_{lt}$ then normalize $\theta_{lt} = \frac{\theta_{lt}}{||\theta_{lt}||_2}$. For deterministic input-awared transition, iMSW, we set $\theta_{lt} = \theta_{lt-1} + \eta \nabla_{\theta_{lt-1}} W_p(\theta_{lt-1}\sharp\mu, \theta_{lt-1}\sharp\nu)$ then normalize $\theta_{lt} = \frac{\theta_{lt}}{||\theta_{lt}||_2}$. For probabilistic input-awared transition, viMSW, $\theta_{lt} \sim \text{vMF}(\theta_t | \epsilon = \text{Prod}_{\mathbb{S}^{d-1}} \theta'_{lt}, \kappa)$ with $\theta'_{lt} = \theta_{lt-1} + \eta \nabla_{\theta_{lt-1}} W_p(\theta_{lt-1}\sharp\mu, \theta_{lt-1}\sharp\nu)$.

Algorithm 6 Markovian sliced Wasserstein distance

Input. Probability measures μ, ν, the dimension d, the order p, the number of projections L, and the number of timesteps T.
for $l = 1$ to L **do**
  Draw $\theta_{l0} \sim \sigma(\theta_0)$
  for $t = 1$ to $T - 1$ **do**
    Draw $\theta_{lt} \sim \sigma_t(\theta_t | \theta_{lt-1})$
  end for
end for
Return. $\left(\frac{1}{LT} \sum_{l=1}^{L} \sum_{t=1}^{T} \text{W}_p^p(\theta_{lt}\sharp\mu, \theta_{lt}\sharp\nu) \right)^{\frac{1}{p}}$

A.4 Burned Thinned Markovian Sliced Wasserstein Distance

We continue the discussion on burned thinned MSW in Section 3.3. We first start with the Monte Carlo estimation of burned thinned MSW.

Monte Carlo Estimation: We samples $\theta_{11}, \ldots, \theta_{L1} \sim \sigma_1(\theta_1)$ for $L \geq 1$, then we samples $\theta_{lt} \sim \sigma_t(\theta_t | \theta_{lt-1})$ for $t = 1, \ldots, T$ and $l = 1, \ldots, L$. We then obtain samples θ'_{lt} by filtering out $t < M$ and $t\%N \neq 0$ from the set $\{\theta_{lt}\}$ for $l = 1, \ldots, L$ and $t = 1, \ldots, T$. The Monte Carlo approximation of the burned-thinned Markovian sliced Wasserstein distance is:

$$\widehat{\text{MSW}}_{p,T,N,M}(\mu, \nu) = \left(\frac{N}{L(T-M)} \sum_{l=1}^{L} \sum_{t=1}^{(T-M)/N} W_p^p \left(\theta'_{lt}\sharp\mu, \theta'_{lt}\sharp\nu \right) \right)^{\frac{1}{p}}. \tag{3}$$

Theoretical properties. We first state the following assumption: **A2.** Given $T > M \geq 0$, $N \geq 1$, the prior distribution $\sigma_1(\theta_1)$ and the transition distribution $\sigma_t(\theta_t | \theta_{t-1})$ are chosen such that there exists marginals $\sigma_t(\theta_t) = \int_{t^-} \sigma(\theta_1, \ldots, \theta_t) dt^-$ with $t \geq M$ and $t\%N = 0$, $t^- = \{t' = 1, \ldots, T | t' \neq t\}$.

The assumption **A2** can be easily obtained by using vMF transition, e.g., in probabilistic input-awared transition. From this assumption, we can derive theoretical properties of burned-thinned MSW including topological properties and statistical complexity.

Proposition 4. *For any $p \geq 1$, $T \geq 1$, $M \geq 0$, $N \geq 1$, and dimension $d \geq 1$, if **A2** holds, the burned thinned Markovian sliced Wasserstein distance $\text{MSW}_{p,T,N,M}(\cdot, \cdot)$ is a valid metric on the space of probability measures $\mathcal{P}_p(\mathbb{R}^d)$, namely, it satisfies the (i) non-negativity, (ii) symmetry, (iii) triangle inequality, and (iv) identity.*

The proof of Proposition 4 follows directly the proof of Theorem 1 in Appendix B.1.

Proposition 5 (Weak Convergence)**.** *For any $p \geq 1$, $T \geq 1$, $M \geq 0$, $N \geq 1$, and dimension $d \geq 1$, if **A2** holds, the convergence of probability measures in $\mathcal{P}_p(\mathbb{R}^d)$ under the burned thinned Markovian sliced Wasserstein distance $\text{MSW}_{p,T,N,M}(\cdot, \cdot)$ implies weak convergence of probability measures and vice versa.*

The proof of Proposition 5 follows directly the proof of Theorem 2 in Appendix B.2.

Proposition 6. *For any $p \geq 1$ and dimension $d \geq 1$, for any $T \geq 1$, $M \geq 0$, $N \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, $MSW_{p,T,N,M}(\mu,\nu) \leq$ Max-$SW_p(\mu,\nu) \leq W_p(\mu,\nu)$.*

The proof of Proposition 6 follows directly the proof of Proposition 1 in Appendix B.3.

Proposition 7 (Sample Complexity)**.** *Let $X_1, X_2, \ldots, X_n$ be i.i.d. samples from the probability measure μ being supported on compact set of $\mathbb{R}^d$. We denote the empirical measure $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. Then, for any $p \geq 1$ and $T \geq 1$, $M \geq 0$, $N \geq 1$, there exists a universal constant $C > 0$ such that*

$$\mathbb{E}[MSW_{p,T,N,M}(\mu_n,\mu)] \leq C\sqrt{(d+1)\log n/n},$$

where the outer expectation is taken with respect to the data $X_1, X_2, \ldots, X_n$.

The proof of Proposition 7 follows directly the proof of Proposition 2 in Appendix B.4.

Proposition 8 (Monte Carlo error)**.** *For any $p \geq 1$, $T \geq 1$, $M \geq 0$, $N \geq 1$, dimension $d \geq 1$, and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, we have:*

$$\begin{aligned}
&\mathbb{E}|\widehat{MSW}^p_{p,T,N,M}(\mu,\nu) - MSW^p_{p,T,N,M}(\mu,\nu)| \\
&\leq \frac{N}{\sqrt{L}T(T-M)} Var\left[\sum_{t=1}^{(T-M)/N} W_p^p\left(\theta'_t\sharp\mu, \theta'_t\sharp\nu\right)\right]^{\frac{1}{2}},
\end{aligned}$$

where the variance is with respect to $\sigma(\theta'_1, \ldots, \theta'_{(T-M)/N})$.

The proof of Proposition 8 follows directly the proof of Proposition 3 in Appendix B.5.

A.5 Discussions on Related Works

K-SW is autoregressive decomposition. In MSW, we assume that the joint distribution over projecting directions has the first-order Markov structure: $\sigma(\theta_1, \ldots, \theta_T) = \sigma_1(\theta_1)\prod_{t=2}^T \sigma_t(\theta_t|\theta_{t-1})$. However, we can consider the full autoregressive decomposition $\sigma(\theta_1, \ldots, \theta_T) = \sigma_1(\theta_1)\prod_{t=2}^T \sigma_t(\theta_t|\theta_1, \ldots, \theta_{t-1})$. Let $T = K$ in K-SW, hence the transition distribution that is used in K-SW is: $\sigma_t(\theta_t|\theta_1, \ldots, \theta_{t-1}) = \text{Gram-Schmidt}_{\theta_1,\ldots,\theta_{t-1}}\sharp\mathcal{U}(\mathbb{S}^{d-1})$, where $\text{Gram-Schmidt}_{\theta_1,\ldots,\theta_{t-1}}(\theta_t)$ denotes the Gram-Schmidt process update that applies on θ_t.

Generalization of Max-K-SW. Similar to Max-SW, we can derive a Markovian-based K-sliced Wasserstein distance that generalizes the idea of the projected gradient ascent update in Max-K-SW. However, the distance considers the transition on the Stiefel manifold instead of the unit hypersphere, hence, it will be more computationally expensive. Moreover, orthogonality might not be a good constraint. Therefore, the generalization of Max-K-SW might not have many advantages.

Beyond the projected sub-gradient ascent update. In the input-awared transition for MSW, we utilize the projected sub-gradient update as the transition function to create a new projecting direction. Therefore, we could other optimization techniques such as momentum, adaptive stepsize, and so on to create the transition function. We will leave the investigation about this direction to future work.

Applications to other sliced Wasserstein variants. The Markovian approach can be applied to other variants of sliced Wasserstein distances e.g., generalized sliced Wasserstein [25], augmented sliced Wasserstein distance [10], projected robust Wasserstein (PRW) [46, 31, 21] ($k > 1$ dimensional projection), convolution sliced Wasserstein [42], sliced partial optimal transport [6, 2], and so on.

Markovian sliced Wasserstein distances in other applications. We can apply MSW to the setting in [30] which is an implementation technique that utilizes both RAM and GPUs' memory for training sliced Wasserstein generative models. MSW can also replace sliced Wasserstein distance in pooling in [37]. Similarly, MSW can be used in applications that exist sliced Wasserstein distance e.g., clustering [27], Bayesian inference [38, 60], domain adaptation [59], and so on.

B Proofs

B.1 Proof of Theorem 1

(i), (ii): the MSW is an expectation of the one-dimensional Wasserstein distance hence the non-negativity and symmetry properties of the MSW follow directly by the non-negativity and symmetry of the Wasserstein distance.

(iii) From the definition of MSW in Definition 1, given three probability measures $\mu_1, \mu_2, \mu_3 \in \mathcal{P}_p(\mathbb{R}^d)$ we have:

$$\begin{aligned}
\text{MSW}_{p,T}(\mu_1, \mu_3) &= \left(\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta_t\sharp\mu_1, \theta_t\sharp\mu_3\right)\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} \left(W_p\left(\theta_t\sharp\mu_1, \theta_t\sharp\mu_2\right) + W_p\left(\theta_t\sharp\mu_2, \theta_t\sharp\mu_3\right)\right)^p\right]\right)^{\frac{1}{p}} \\
&\leq \left(\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta_t\sharp\mu_1, \theta_t\sharp\mu_2\right)\right]\right)^{\frac{1}{p}} \\
&\quad + \left(\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta_t\sharp\mu_2, \theta_t\sharp\mu_3\right)\right]\right)^{\frac{1}{p}} \\
&= \text{MSW}_{p,T}(\mu_1, \mu_2) + \text{MSW}_{p,T}(\mu_2, \mu_3),
\end{aligned}$$

where the first inequality is due to the triangle inequality of Wasserstein distance and the second inequality is due to the Minkowski inequality. We complete the triangle inequality proof.

(iv) We need to show that $\text{MSW}_{p,T}(\mu, \nu) = 0$ if and only if $\mu = \nu$. First, from the definition of MSW, we obtain directly $\mu = \nu$ implies $\text{MSW}_{p,T}(\mu, \nu) = 0$. For the reverse direction, we use the same proof technique in [8]. If $\text{MSW}_{p,T}(\mu, \nu) = 0$, we have $\int_{\mathbb{S}^{(d-1)\otimes T}} \frac{1}{T}\sum_{t=1}^{T} \text{W}_p\left(\theta_t\sharp\mu, \theta_t\sharp\nu\right) \text{d}\sigma(\theta_{1:T}) = 0$. If **A1** holds, namely, the prior distribution $\sigma_1(\theta_1)$ is supported on all the unit-hypersphere or exists a transition distribution $\sigma_t(\theta_t|\theta_{t-1})$ is supported on all the unit-hypersphere, we have $W_p(\theta\sharp\mu, \theta\sharp\nu) = 0$ for all $\theta \in \mathbb{S}^{d-1}$ where $\boldsymbol{\sigma}$ denotes the prior or the transition distribution that satisfies the assumption **A1**. From the identity property of the Wasserstein distance, we obtain $\theta\sharp\mu = \theta\sharp\nu$ for $\boldsymbol{\sigma}$-a.e $\theta \in \mathbb{S}^{d-1}$. Therefore, for any $t \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$, we have:

$$\begin{aligned}
\mathcal{F}[\mu](t\theta) &= \int_{\mathbb{R}^d} e^{-it\langle\theta,x\rangle} d\mu(x) = \int_{\mathbb{R}} e^{-itz} d\theta\sharp\mu(z) = \mathcal{F}[\theta\sharp\mu](t) \\
&= \mathcal{F}[\theta\sharp\nu](t) = \int_{\mathbb{R}} e^{-itz} d\theta\sharp\nu(z) = \int_{\mathbb{R}^d} e^{-it\langle\theta,x\rangle} d\nu(x) = \mathcal{F}[\nu](t\theta),
\end{aligned}$$

where $\mathcal{F}[\gamma](w) = \int_{\mathbb{R}^{d'}} e^{-i\langle w,x\rangle} d\gamma(x)$ denotes the Fourier transform of $\gamma \in \mathcal{P}(\mathbb{R}^{d'})$. By the injectivity of the Fourier transform, we obtain $\mu = \nu$ which concludes the proof.

B.2 Proof of Theorem 2

Our goal is to show that for any sequence of probability measures $(\mu_k)_{k\in\mathbb{N}}$ and μ in $\mathcal{P}_p(\mathbb{R}^d)$, $\lim_{k\to+\infty} \text{MSW}_{p,T}(\mu_k, \mu) = 0$ if and only if for any continuous and bounded function $f : \mathbb{R}^d \to \mathbb{R}$, $\lim_{k\to+\infty} \int f \, \text{d}\mu_k = \int f \, \text{d}\mu$. The proof follows the techniques in [40]. We first state the following lemma.

Lemma 1. *For any $p \geq 1$, $T \geq 1$, and dimension $d \geq 1$, if **A1** holds and a sequence of probability measures $(\mu_k)_{k\in\mathbb{N}}$ satisfies $\lim_{k\to+\infty} MSW_{p,T}(\mu_k, \mu) = 0$ with μ in $\mathcal{P}_p(\mathbb{R}^d)$, there exists an increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that the subsequence $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ converges weakly to μ.*

Proof. We are given that $\lim_{k\to+\infty} \text{MSW}_{p,T}(\mu_k, \mu) = 0$, therefore $\lim_{k\to\infty} \int_{\mathbb{S}^{(d-1)\otimes T}} \frac{1}{T}\sum_{t=1}^{T} \text{W}_p\left(\theta_t\sharp\mu_k, \theta_t\sharp\mu\right) \text{d}\sigma(\theta_{1:T}) = 0$. If **A1** holds, namely, the prior

distribution $\sigma_1(\theta_1)$ is supported on all the unit-hypersphere or exists a transition distribution $\sigma_t(\theta_t|\theta_{t-1})$ is supported on all the unit-hypersphere, we have

$$\lim_{k\to\infty}\int_{\mathbb{S}^{d-1}} \mathrm{W}_p\left(\theta\sharp\mu_k, \theta\sharp\mu\right)\mathrm{d}\boldsymbol{\sigma}(\theta) = 0,$$

where $\boldsymbol{\sigma}$ denotes the prior or the transition distribution that satisfies the assumption **A1**. From Theorem 2.2.5 in [3], there exists an increasing function $\phi : \mathbb{N} \to \mathbb{N}$ such that $\lim_{k\to\infty} \mathrm{W}_p(\theta\sharp\mu_{\phi(k)}, \theta\sharp\nu) = 0$ for $\boldsymbol{\sigma}$-a.e $\theta \in \mathbb{S}^{d-1}$. Since the Wasserstein distance of order p implies weak convergence in $\mathcal{P}_p(\mathbb{R}^d)$ [57], $\left(\theta\sharp\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ converges weakly to $\theta\sharp\mu$ for $\boldsymbol{\sigma}$-a.e $\theta \in \mathbb{S}^{d-1}$.

Let $\Phi_\mu = \int_{\mathbb{R}^d} \mathrm{e}^{\mathrm{i}\langle v,w\rangle}\mathrm{d}\mu(w)$ be the characteristic function of $\mu \in \mathcal{P}_p(\mathbb{R}^d)$, we have the weak convergence implies the convergence of characteristic function (Theorem 4.3 [23]): $\lim_{k\to\infty} \Phi_{\theta\sharp\mu_{\phi(k)}}(s) = \Phi_{\theta\sharp\mu}(s), \quad \forall s \in \mathbb{R}$, for $\boldsymbol{\sigma}$-a.e $\theta \in \mathbb{S}^{d-1}$. Therefore, $\lim_{k\to\infty} \Phi_{\mu_{\phi(k)}}(z) = \Phi_\mu(z)$, for almost most every $z \in \mathbb{R}^d$.

For any $\gamma > 0$ and a continuous function $f : \mathbb{R}^d \to \mathbb{R}$ with compact support, we denote $f_\gamma(x) = f * g_\gamma(x) = \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d} f(x-z)\exp\left(-\|z\|^2/\left(2\gamma^2\right)\right)\mathrm{d}z$ where g_γ is the density function of $\mathcal{N}(0, \gamma I_d)$. We have:

$$\begin{aligned}
\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu_{\phi(k)}(z) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)g_\gamma(z-w)\mathrm{d}w\,\mathrm{d}\mu_{\phi(k)}(z)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\left(2\pi\gamma^2\right)^{-d/2}\exp(-||z-w||^2/(2\gamma^2))\mathrm{d}w\,\mathrm{d}\mu_{\phi(k)}(z)\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\int_{\mathbb{R}^d} e^{\mathrm{i}\langle z-w,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\,\mathrm{d}w\,\mathrm{d}\mu_{\phi(k)}(z)\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\int_{\mathbb{R}^d} e^{-\mathrm{i}\langle w,x\rangle}e^{\mathrm{i}\langle z,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\,\mathrm{d}w\,\mathrm{d}\mu_{\phi(k)}(z)\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)e^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\int_{\mathbb{R}^d} e^{\mathrm{i}\langle z,x\rangle}\,\mathrm{d}\mu_{\phi(k)}(z)\mathrm{d}x\,\mathrm{d}w\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\mathrm{e}^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x\,\mathrm{d}w\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x,
\end{aligned}$$

where the third equality is due to the fact that $\int_{\mathbb{R}^d} e^{\mathrm{i}\langle z-w,x\rangle}g_{1/\gamma}(x)\mathrm{d}x = \exp(-||z-w||^2/(2\gamma^2))$ and $\mathcal{F}[f](w) = \int_{\mathbb{R}^{d'}} f(x)e^{-i\langle w,x\rangle}dx$ denotes the Fourier transform of the bounded function f. Similarly, we have

$$\begin{aligned}
\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu(z) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)g_\gamma(z-w)\mathrm{d}w\,\mathrm{d}\mu(z)\\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\left(2\pi\gamma^2\right)^{-d/2}\exp(-||z-w||^2/(2\gamma^2))\mathrm{d}w\,\mathrm{d}\mu(z)\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\int_{\mathbb{R}^d} e^{\mathrm{i}\langle z-w,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\,\mathrm{d}w\,\mathrm{d}\mu(z)\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\int_{\mathbb{R}^d} e^{-\mathrm{i}\langle w,x\rangle}e^{\mathrm{i}\langle z,x\rangle}g_{1/\gamma}(x)\mathrm{d}x\,\mathrm{d}w\,\mathrm{d}\mu(z)\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)e^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\int_{\mathbb{R}^d} e^{\mathrm{i}\langle z,x\rangle}\,\mathrm{d}\mu(z)\mathrm{d}x\,\mathrm{d}w\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(w)\mathrm{e}^{-\mathrm{i}\langle w,x\rangle}g_{1/\gamma}(x)\Phi_{\mu}(x)\mathrm{d}x\,\mathrm{d}w\\
&= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu}(x)\mathrm{d}x.
\end{aligned}$$

Since f is assumed to have compact support, $\mathcal{F}[f]$ exists and is bounded by $\int_{\mathbb{R}^d}|f(w)|\mathrm{d}w < +\infty$. Hence, for any $k \in \mathbb{R}$ and $x \in \mathbb{R}^d$, we have $\left|\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\right| \leq$

$g_{1/\gamma}(x)\int_{\mathbb{R}^d}|f(w)|\mathrm{d}w$ and $\left|\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_\mu(x)\right| \leq g_{1/\gamma}(x)\int_{\mathbb{R}^d}|f(w)|\mathrm{d}w$. Using the proved result of $\lim_{k\to\infty}\Phi_{\mu_{\phi(k)}}(z)=\Phi_\mu(z)$ and Lebesgue's Dominated Convergence Therefore, we obtain

$$\begin{aligned}\lim_{k\to\infty}\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu_{\phi(k)}(z) &= \lim_{k\to\infty}\left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x\\ &= \left(2\pi\gamma^2\right)^{-d/2}\int_{\mathbb{R}^d}\mathcal{F}[f](x)g_{1/\gamma}(x)\Phi_{\mu_{\phi(k)}}(x)\mathrm{d}x\\ &= \int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu(z).\end{aligned}$$

Moreover, we have:

$$\begin{aligned}&\lim_{\gamma\to 0}\limsup_{k\to+\infty}\left|\int_{\mathbb{R}^d} f(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d} f(z)\mathrm{d}\mu(z)\right|\\ &\leq \lim_{\gamma\to 0}\limsup_{k\to+\infty}\left[2\sup_{z\in\mathbb{R}^d}|f(z)-f_\gamma(z)|+\left|\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu_{\phi(k)}(z)-\int_{\mathbb{R}^d} f_\gamma(z)\mathrm{d}\mu(z)\right|\right]\\ &= \lim_{\gamma\to 0} 2\sup_{z\in\mathbb{R}^d}|f(z)-f_\gamma(z)|=0,\end{aligned}$$

which implies $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ converges weakly to μ. □

We now continue the proof of Theorem 2. We first show that if $\lim_{k\to\infty}\mathrm{MSW}_{p,T}(\mu_k,\mu)=0$, $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ. We consider a sequence $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\mathrm{MSW}_{p,T}(\mu_k,\mu)=0$ and we suppose $\left(\mu_{\phi(k)}\right)_{k\in\mathbb{N}}$ does not converge weakly to μ. Therefore, let $d_{\mathcal{P}}$ be the Lévy-Prokhorov metric, $\lim_{k\to\infty} d_{\mathcal{P}(\mu_k,\mu)}\neq 0$ that implies there exists $\varepsilon>0$ and a subsequence $\left(\mu_{\psi(k)}\right)_{k\in\mathbb{N}}$ with an increasing function $\psi:\mathbb{N}\to\mathbb{N}$ such that for any $k\in\mathbb{N}$: $d_{\mathcal{P}}(\mu_{\psi(k)},\mu)\geq\varepsilon$. However, we have

$$\begin{aligned}\mathrm{MSW}_{p,T}(\mu,\nu) &= \left(\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta_t\sharp\mu,\theta_t\sharp\nu\right)\right]\right)^{\frac{1}{p}}\\ &\geq \mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p\left(\theta_t\sharp\mu,\theta_t\sharp\nu\right)\right]\\ &\geq \mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_1\left(\theta_t\sharp\mu,\theta_t\sharp\nu\right)\right]=\mathrm{MSW}_{1,T}(\mu,\nu),\end{aligned}$$

by the Holder inequality with $\mu,\nu\in\mathbb{P}_p(\mathbb{R}^d)$. Therefore, $\lim_{k\to\infty}\mathrm{MSW}_{1,T}(\mu_{\psi(k)},\mu)=0$ which implies that there exists s a subsequence $\left(\mu_{\phi(\psi(k))}\right)_{k\in\mathbb{N}}$ with an increasing function $\phi:\mathbb{N}\to\mathbb{N}$ such that $\left(\mu_{\phi(\psi(k))}\right)_{k\in\mathbb{N}}$ converges weakly to μ by Lemma 1. Hence, $\lim_{k\to\infty} d_{\mathcal{P}}\left(\mu_{\phi(\psi(k))},\mu\right)=0$ which contradicts our assumption. We conclude that if $\lim_{k\to\infty}\mathrm{MSW}_{p,T}(\mu_k,\mu)=0$, $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ.

Now, we show that if $(\mu_k)_{k\in\mathbb{N}}$ converges weakly to μ, we have $\lim_{k\to\infty}\mathrm{MSW}_{p,T}(\mu_k,\mu)=0$. By the continuous mapping theorem, we obtain $(\theta\sharp\mu_k)_{k\in\mathbb{N}}$ converges weakly to $\theta\sharp\mu$ for any $\theta\in\mathbb{S}^{d-1}$. Since the weak convergence implies the convergence under the Wasserstein distance [57], we obtain $\lim_{k\to\infty} W_p(\theta\sharp\mu_k,\mu)=0$. Moreover, the Wasserstein distance is also bounded, hence the bounded convergence theorem:

$$\begin{aligned}\lim_{k\to\infty}\mathrm{MSW}_{p,T}^p(\mu_k,\mu) &= \mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta_t\sharp\mu_k,\theta_t\sharp\mu\right)\right]\\ &= \mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} 0\right]=0.\end{aligned}$$

By the continuous mapping theorem with function $x\to x^{1/p}$, we obtain $\lim_{k\to\infty}\mathrm{MSW}_{p,T}(\mu_k,\mu)\to 0$ which completes the proof.

B.3 Proof of Proposition 1

(i) We recall the definition of Max-SW:

$$\text{Max-SW}_p(\mu,\nu) = \max_{\theta\in\mathbb{S}^{d-1}} W_p(\theta\sharp\mu,\theta\sharp\nu).$$

Let $\theta^* = \text{argmax}_{\theta\in\mathbb{S}^{d-1}} W_p(\theta\sharp\mu,\theta\sharp\nu)$, from Definition 1, for any $p\geq 1$, $T\geq 1$, dimension $d\geq 1$, and $\mu,\nu\in\mathcal{P}_p(\mathbb{R}^d)$ we have:

$$\begin{aligned}\text{MSW}_{p,T}(\mu,\nu) &= \left(\mathbb{E}_{(\theta_{1:T})\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta_t\sharp\mu,\theta_t\sharp\nu\right)\right]\right)^{\frac{1}{p}}\\ &\leq \left(\frac{1}{T}\sum_{t=1}^{T} W_p^p\left(\theta^*\sharp\mu,\theta^*\sharp\nu\right)\right)^{\frac{1}{p}} = W_p\left(\theta^*\sharp\mu,\theta^*\sharp\nu\right) = \text{Max-SW}_p(\mu,\nu).\end{aligned}$$

Furthermore, by applying the Cauchy-Schwartz inequality, we have:

$$\begin{aligned}\text{Max-SW}_p^p(\mu,\nu) &= \max_{\theta\in\mathbb{S}^{d-1}}\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d}\left|\theta^\top x-\theta^\top y\right|^p d\pi(x,y)\right)\\ &\leq \max_{\theta\in\mathbb{S}^{d-1}}\left(\inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\theta\|^p\|x-y\|^p d\pi(x,y)\right)\\ &= \inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|\theta\|^p\|x-y\|^p d\pi(x,y)\\ &= \inf_{\pi\in\Pi(\mu,\nu)}\int_{\mathbb{R}^d\times\mathbb{R}^d}\|x-y\|^p d\pi(x,y)\\ &= W_p^p(\mu,\nu),\end{aligned}$$

which completes the proof.

(ii) This result can be directly obtained from the definitions of MSW and SW.

B.4 Proof of Proposition 2

In this proof, we denote $\Theta\subset\mathbb{R}^d$ as the compact set of the probability measure μ. From Proposition 1, we find that

$$\mathbb{E}[\text{MSW}_{p,T}(\mu_n,\mu)] \leq \mathbb{E}\left[\text{Max-SW}_p(\mu_n,\mu)\right].$$

Therefore, the proposition follows as long as we can demonstrate that

$$\mathbb{E}[\text{Max-SW}_p(\mu_n,\mu)] \leq C\sqrt{(d+1)\log_2 n/n}$$

where $C>0$ is some universal constant and the outer expectation is taken with respect to the data. The proof for this result follows from the proof of Proposition 3 in [42]. Here, we provide the proof for the completeness. By defining $F_{n,\theta}$ and F_θ as the cumulative distributions of $\theta\sharp\mu_n$ and $\theta\sharp\mu$, the closed-form expression of the Wasserstein distance in one dimension leads to the following equations and inequalities:

$$\begin{aligned}\text{Max-SW}_p^p(\mu_n,\mu) &= \max_{\theta\in\mathbb{S}^{d-1}}\int_0^1 |F_{n,\theta}^{-1}(u)-F_\theta^{-1}(u)|^p du\\ &= \max_{\theta\in\mathbb{R}^d:\|\theta\|=1}\int_0^1 |F_{n,\theta}^{-1}(u)-F_\theta^{-1}(u)|^p du\\ &\leq \text{diam}(\Theta)\max_{\theta\in\mathbb{R}^d:\|\theta\|\leq 1}|F_{n,\theta}(x)-F_\theta(x)|^p.\end{aligned}$$

We can check that

$$\max_{\theta\in\mathbb{R}^d:\|\theta\|\leq 1}|F_{n,\theta}(x)-F_\theta(x)| = \sup_{B\in\mathcal{B}}|\mu_n(B)-\mu(B)|,$$

Algorithm 7 Gradient flow with the Euler scheme

Input. the start distribution $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$, the target distribution $\nu = \frac{1}{n}\sum_{i=1}^n \delta_{Y_i}$, number of Euler iterations T (abuse of notation), Euler step size η (abuse of notation), a metric $\mathcal{D}$.
for $t = 1$ to T **do**
  $X = X - n \cdot \eta \nabla_X \mathcal{D}(P_X, P_Y)$
end for
Output. $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$

where $\mathcal{B}$ is the set of half-spaces $\{z \in \mathbb{R}^d : \theta^\top z \leq x\}$ for all $\theta \in \mathbb{R}^d$ such that $\|\theta\| \leq 1$. From [58], we can show that the Vapnik-Chervonenkis (VC) dimension of $\mathcal{B}$ is at most $d+1$. Therefore, the following inequality holds:

$$\sup_{B \in \mathcal{B}} |\mu_n(B) - \mu(B)| \leq \sqrt{\frac{32}{n}[(d+1)\log_2(n+1) + \log_2(8/\delta)]}$$

with probability at least $1-\delta$. Putting the above results together leads to

$$\mathbb{E}[\text{Max-SW}_p(\mu_n, \mu)] \leq C\sqrt{(d+1)\log_2 n / n},$$

where $C > 0$ is some universal constant. As a consequence, we obtain the conclusion of the proposition.

B.5 Proof of Proposition 3

For any $p \geq 1$, $T \geq 1$, dimension $d \geq 1$, and $\mu, \nu \in \mathcal{P}_p(\mathbb{R}^d)$, using the Holder's inequality, we have:

$$\begin{aligned}
&\mathbb{E}|\widehat{\text{MSW}}^p_{p,T}(\mu,\nu) - \text{MSW}^p_{p,T}(\mu,\nu)| \\
&\quad\leq \left(\mathbb{E}|\widehat{\text{MSW}}^p_{p,k}(\mu,\nu) - \text{MSW}^p_{p,k}(\mu,\nu)|^2\right)^{\frac{1}{2}} \\
&\quad= \left(\mathbb{E}\left|\frac{1}{TL}\sum_{t=1}^T\sum_{l=1}^L \text{W}^p_p(\theta_{tl}\sharp\mu, \theta_{tl}\sharp\nu) - \mathbb{E}_{\theta_{1:T}\sim\sigma(\theta_{1:T})}\left[\frac{1}{T}\sum_{t=1}^T W^p_p\left(\theta_t\sharp\mu, \theta_t\sharp\nu\right)\right]\right|^2\right)^{\frac{1}{2}} \\
&\quad= \left(Var\left[\frac{1}{TL}\sum_{t=1}^T\sum_{l=1}^L W^p_p\left(\theta_t\sharp\mu, \theta_t\sharp\nu\right)\right]\right)^{\frac{1}{2}} \\
&\quad= \frac{1}{T\sqrt{L}} Var\left[\sum_{t=1}^T W^p_p\left(\theta_t\sharp\mu, \theta_t\sharp\nu\right)\right]^{\frac{1}{2}},
\end{aligned}$$

which completes the proof.

C Additional Experiments

In this section, we present the detail of experimental frameworks and additional experiments on gradient flows, color transfer, and deep generative modeling which are not in the main paper.

C.1 Gradient Flows

Framework. We have discussed in detail the framework of gradient flow in Section 4.1 in the main paper. Here, we summarize the Euler scheme for solving the gradient flow in Algorithm 7.

Visualization of gradient flows. We show the visualization of gradient flows from all distances (Table 1) in Figure 4. Overall, we observe that the quality of the flows is consistent with the quantitative Wasserstein-2 score which is computed using [17]. From the figures, we see that iMSW and viMSW help the flows converge very fast. Namely, Wasserstein-2 scores at steps 200 of iMSW and viMSW are much lower than other distances. For oMSW, with $L = 5, T = 2$, it achieves a comparable result to SW, K-SW, and Max-SW while being faster.

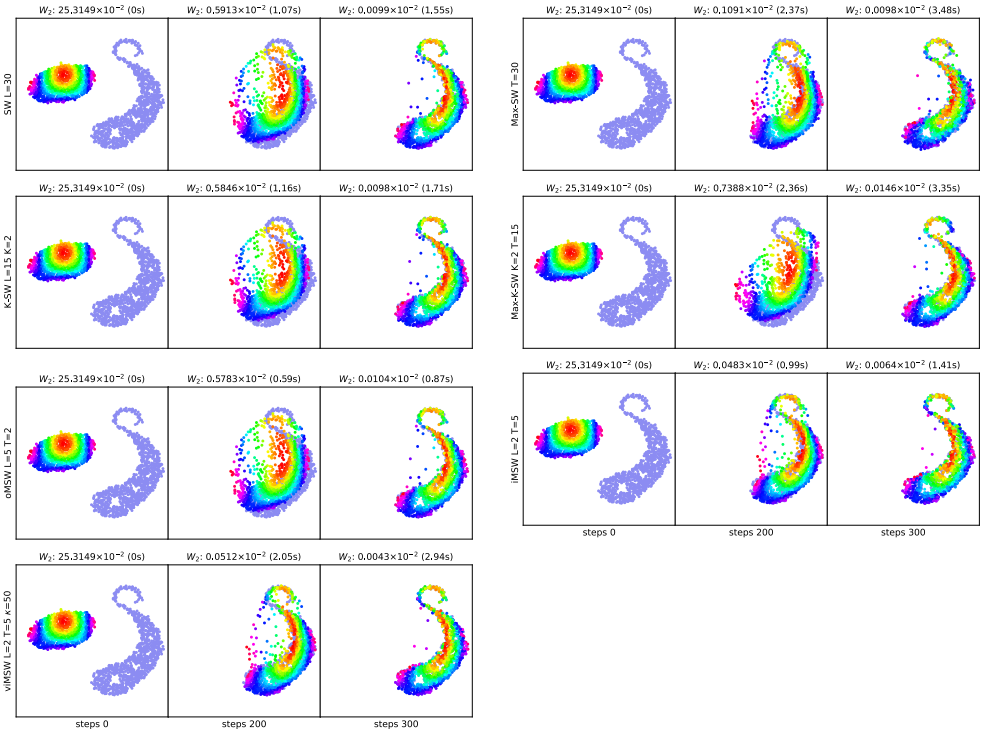


Figure 4: The figures show the gradient flows that are from the empirical distribution over the color points to the empirical distribution over S-shape points produced by different distances. The corresponding Wasserstein-2 distance between the empirical distribution at the current step and the S-shape distribution and the computational time (in second) to reach the step is reported at the top of the figure.

Table 3: Summary of Wasserstein-2 scores, computational time in second (s) of different distances in gradient flow application.

Distances	Wasserstein-2 (↓)	Time (↓)	Distances	Wasserstein-2 (↓)	Time (↓)
SW (L=10)	0.0113×10^{-2}	0.85	SW (L=100)	0.0096×10^{-2}	4.32
Max-SW (T=5)	0.0231×10^{-2}	1.02	Max-SW (T=100)	0.0083×10^{-2}	10.46
K-SW (L=5,K=2)	0.0104×10^{-2}	0.92	K-SW (L=20,K=2)	0.0096×10^{-2}	1.97
Max-K-SW (K=2,T=5)	0.0152×10^{-2}	1.41	Max-K-SW (K=2,T=100)	0.0083×10^{-2}	10.46
iMSW (L=1,T=5)	0.0109×10^{-2}	1.07	iMSW (L=5,T=5)	0.0055×10^{-2}	2.44
iMSW (L=2,T=10)	0.0052×10^{-2}	2.79	iMSW (L=5,T=2)	0.0071×10^{-2}	1.14
iMSW (L=2,T=5,M=4)	0.0101×10^{-2}	1.2	iMSW (L=2,T=5,M=2)	0.0055×10^{-2}	1.25
iMSW (L=2,T=5,M=0,N=2)	0.0066×10^{-2}	1.28	iMSW (L=2,T=5,M=2,N=2)	0.0072×10^{-2}	1.19
viMSW (L=2,T=5,κ=10)	0.0052×10^{-2}	3.12	viMSW (L=2,T=5,κ=100)	0.0053×10^{-2}	2.76

Studies on hyper-parameters. We run gradient flows with different values of hyper-parameters and report the Wasserstein-2 scores and computational time in Table 3. From the table and Figure 4, we see that SW with $L = 10$ is worse than oMSW, iMSW, and viMSW with $L = 2, T = 5$ (10 total projections). Increasing the number of projections to 100, SW gets better, however, its Wasserstein-2 score is still higher than the scores of iMSW and viMSW while its computational time is bigger. Similarly, Max-(K)-SW with $T = 100$ is better than Max-(K)-SW with $T = 5$ and $T = 10$, however, it is still worse than iMSW and viMSW in terms of computation and performance. For burning and thinning, we see that the technique can help improve the computation considerably. More importantly, the burning and thinning techniques do not reduce the performance too much. For iMSW, increasing L and T leads to a better flow. For the same number of total projections e.g., 10, $L = 2, T = 5$ is better than $L = 5, T = 2$. For viMSW, it usually performs better than iMSW, however, its computation is worse due to the sampling complexity of the vMF distribution. We vary the concentration parameter $\kappa \in \{10, 50, 100\}$ and find that $\kappa = 50$ is the best. Hence, it might

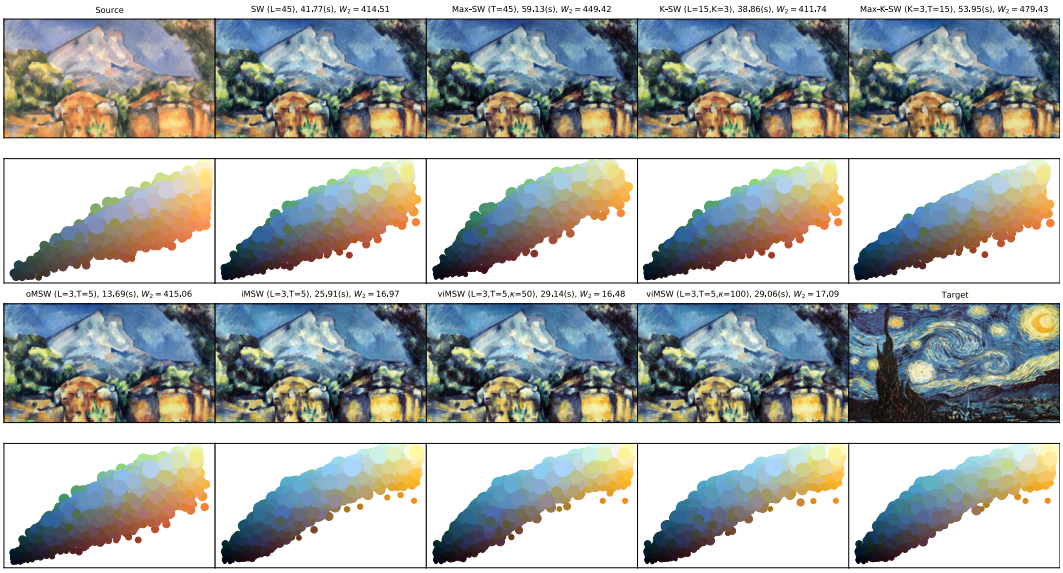


Figure 5: The figures show the source image, the target image, and transferred images from different distances. The corresponding Wasserstein-2 distance between the empirical distribution over transferred color palates and the empirical distribution over the target color palette and the computational time (in second) is reported at the top of the figure. The color palates are given below the corresponding images.

Algorithm 8 Color Transfer

Input. source color palette $X \in \{0, \dots, 255\}^{n \times 3}$, target color palette $Y \in \{0, \dots, 255\}^{n \times 3}$, number of Euler iterations T (abuse of notation), Euler step size η (abuse of notation), a metric $\mathcal{D}$.
for $t = 1$ to T **do**
  $X = X - n \cdot \eta \nabla_X \mathcal{D}(P_X, P_Y)$
end for
$X = \text{round}(X, \{0, \dots, 255\})$
Output. X

suggest that a good balance between heading to the “max" projecting direction and exploring the space of projecting directions is the best strategy.

C.2 Color Transfer

Framework. In our experiments, we first compress the color palette of the source image and the target image to 3000 colors by using K-Mean clustering. After that, the color transfer application is conducted by using Algorithm 8 which is a modified version of the gradient flow algorithm since the color palette contains only positive integer in $\{0, \dots, 255\}$. The flow can be seen as an incomplete transportation map that maps from the source color palette to a color palette that is close to the target color palette. This is quite similar to the iterative distribution transfer algorithm [8], however, the construction of the iterative map is different.

Visuallization of transferred images. We show the source image, the target image, and the corresponding transferred images from distances in Figure 5 and Figure 6. The color palates are given below the corresponding images. The corresponding Wasserstein-2 distance between the empirical distribution over transferred color palates and the empirical distribution over the target color palette and the computational time (in second) is reported at the top of the figure. First, we observe that the qualitative comparison (transferred images and color palette) is consistent with the Wasserstein scores. We observe that iMSW and viMSW have their transferred images closer to the target image in terms of color than other distances. More importantly, iMSW and viMSW are faster than other distances. Max-SW and Max-K-SW do not perform well in this application, namely, they are slow

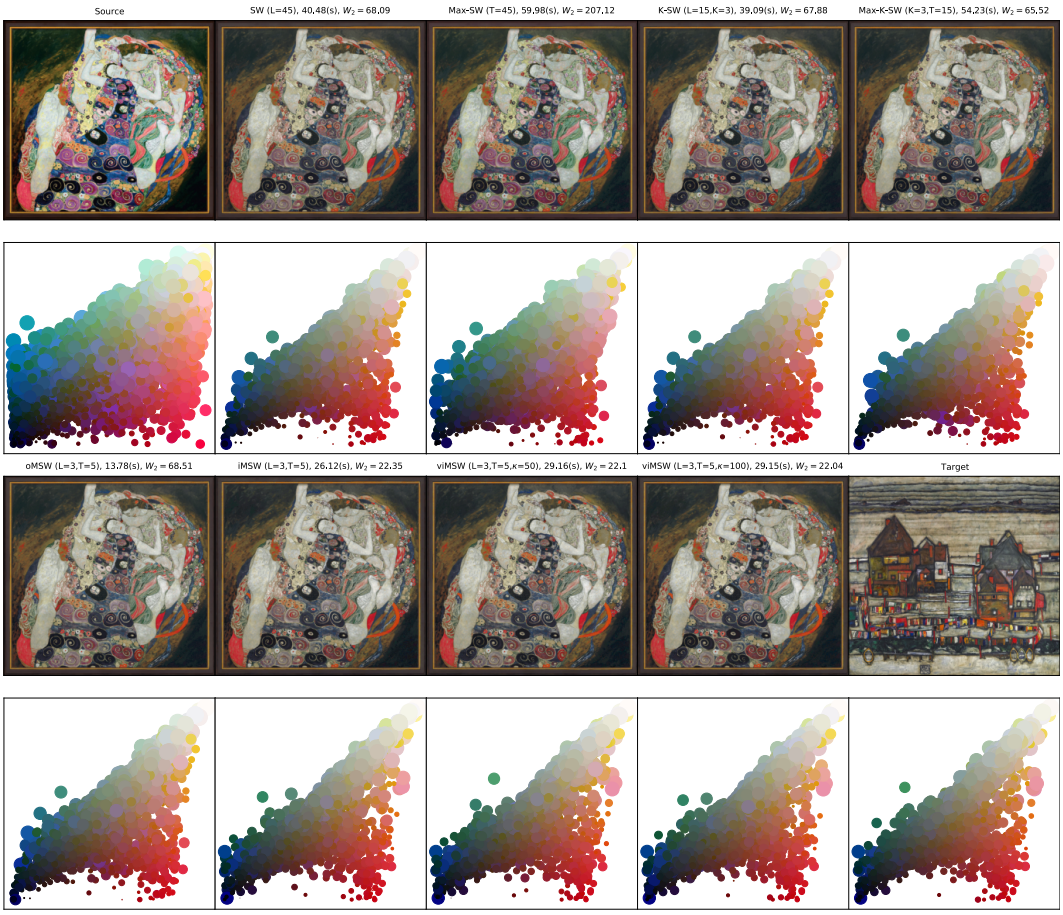


Figure 6: The figures show the source image, the target images, and transferred images from different distances. The corresponding Wasserstein-2 distance between the empirical distribution over transferred color palates and the empirical distribution over the target color palette and the computational time (in second) is reported at the top of the figure. The color palates are given below the corresponding images.

Table 4: Summary of Wasserstein-2 scores, computational time in second (s) of different distances in the color transfer application.

Distances	Wasserstein-2 (↓)	Time (↓)	Distances	Wasserstein-2 (↓)	Time (↓)
SW (L=45)	414.51	41.77	SW (L=15)	421.5	12.96
Max-SW (T=45)	449.42	59.13	Max-SW (T=15)	450.37	19.03
K-SW (L=15,K=3)	411.74	38.86	K-SW (L=5,K=3)	413.16	14.2
Max-K-SW (K=3,T=15)	479.43	53.95	Max-K-SW (K=3,T=5)	510.43	17.46
oMSW (L=3,T=5)	415.06	13.69	oMSW (L=3,T=15)	414.29	38.51
iMSW (L=3,T=5) iMSW (L=5,T=5) iMSW (L=3,T=15,M=14) iMSW (L=3,T=15,M=0,N=2)	16.97 21.63 26.23 16.6	25.91 39.82 48.08 62.66	iMSW (L=3,T=15) iMSW (L=5,T=3) iMSW (L=3,T=15,M=10) iMSW (L=3,T=15,M=10,N=2)	15.23 24.02 18.67 19.2	79.47 22.27 55.55 50.1
viMSW (L=3,T=5,κ=50)	16.48	29.14	viMSW (L=3,T=5,κ=100)	16.49	29.06

and give high Wasserstein distances. For oMSW, it is comparable to SW and K-SW while being faster.

Studies on hyper-parameters. In addition to result in Figure 5, we run color transfer with other settings of distances in Table 4. From the table, increasing the number of projections L lead to a better result for SW and K-SW. However, they are still worse than iMSW and viMSW with a

smaller number of projections. Similarly, increasing T helps Max-SW, Max-K-SW, and iMSW better. As discussed in the main paper, the burning and thinning technique improves the computation and sometimes enhances the performance.

C.3 Deep Generative Models

Framework. We follow the generative modeling framework from [19, 41]. Here, we state an adaptive formulation of the framework. We are given a data distribution $\mu \in \mathcal{P}(\mathcal{X})$ through its random samples (data). Our goal is to estimate a parametric distribution ν_ϕ that belongs to a family of distributions indexed by parameters ϕ in a parameter space Φ. Deep generative modeling is interested in constructing ν_ϕ via pushforward measure. In particular, ν_ϕ is implicitly represented by pushing forward a random noise $\nu_0 \in \mathcal{P}(\mathcal{Z})$ e.g., standard multivariable Gaussian, through a parametric function $G_\phi : \mathcal{Z} \to \mathcal{X}$ (a neural network with weights ϕ). To estimate ϕ (ν_ϕ), the expected distance estimator [53, 40] is used:

$$\text{argmin}_{\phi\in\Phi} \mathbb{E}_{(X,Z)\sim\mu^{\otimes m}\otimes\nu_0^{\otimes m}}[\mathcal{D}(P_X, P_{G_\phi(Z)})],$$

where $m \geq 1$, $\mathcal{D}$ can be any distance on space of probability measures, $\mu^{\otimes}$ is the product measures, namely, $X = (x_1, \ldots, x_m) \sim \mu^{\otimes}$ is equivalent to $x_i \sim \mu$ for $i = 1, \ldots, m$, and $P_X = \frac{1}{m}\sum_{i=1}^m \delta_{x_i}$. Similarly, $Z = (z_1, \ldots, z_m)$ with $z_i \sim \nu_0$ for $i = 1, \ldots, m$, and $G_\phi(Z)$ is the output of the neural work given the input mini-batch Z.

By using Wasserstein distance, sliced Wasserstein distance, and their variants as the distance $\mathcal{D}$, we obtain the corresponding estimators. However, applying directly those estimators to natural image data cannot give perceptually good results [19, 15]. The reason is that Wasserstein distance, sliced Wasserstein distances, and their variants require a ground metric as input e.g., $\mathcal{L}_2$, however, those ground metrics are not meaningful on images. Therefore, previous works propose using a function that maps the original data space $\mathcal{X}$ to a feature space $\mathcal{F}$ where the $\mathcal{L}_2$ norm is meaningful [51]. We denote the feature function $F_\gamma : \mathcal{X} \to \mathcal{F}$. Now the estimator becomes:

$$\text{argmin}_{\phi\in\Phi} \mathbb{E}_{(X,Z)\sim\mu^{\otimes m}\otimes\nu_0^{\otimes m}}[\mathcal{D}(P_{F_\gamma(X)}, P_{F_\gamma(G_\phi(Z))})].$$

The above optimization can be solved by stochastic gradient descent algorithm with the following stochastic gradient estimator:

$$\begin{aligned}\nabla_\phi \mathbb{E}_{(X,Z)\sim\mu^{\otimes m}\otimes\nu_0^{\otimes m}}[\mathcal{D}(P_{F_\gamma(X)}, P_{F_\gamma(G_\phi(Z))})] &= \mathbb{E}_{(X,Z)\sim\mu^{\otimes m}\otimes\nu_0^{\otimes m}}[\nabla_\phi \mathcal{D}(P_{F_\gamma(X)}, P_{F_\gamma(G_\phi(Z))})] \\ &\approx \frac{1}{K}\sum_{k=1}^K \nabla_\phi \mathcal{D}(P_{F_\gamma(X_k)}, P_{F_\gamma(G_\phi(Z_k))}),\end{aligned}$$

where $X_1, \ldots, X_K$ are drawn i.i.d from $\mu^{\otimes m}$ and $Z_1, \ldots, Z_K$ are drawn i.i.d from $\nu_0^{\otimes m}$. There are several ways to estimate the feature function F_γ in practice. In our experiments, we use the following objective [15]:

$$\min_\gamma \left(\mathbb{E}_{X\sim\mu^{\otimes m}}[\min(0, -1 + H(F_\gamma(X)))] + \mathbb{E}_{Z\sim\nu_0^{\otimes m}}[\min(0, -1 - H(F_\gamma(G_\phi(Z))))]\right),$$

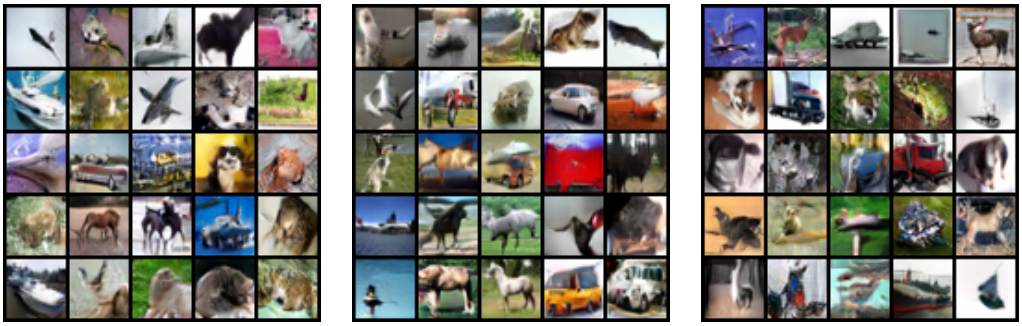
where $H : \mathcal{F} \to \mathbb{R}$. The above optimization problem is also solved by the stochastic gradient descent algorithm with the following gradient estimator:

$$\begin{aligned}&\nabla_\gamma \left(\mathbb{E}_{X\sim\mu^{\otimes m}}[\min(0, -1 + H(F_\gamma(X)))] + \mathbb{E}_{Z\sim\nu_0^{\otimes m}}[\min(0, -1 - H(F_\gamma(G_\phi(Z))))]\right) \\ &= \mathbb{E}_{X\sim\mu^{\otimes m}}[\nabla_\gamma \min(0, -1 + H(F_\gamma(X)))] + \mathbb{E}_{Z\sim\nu_0^{\otimes m}}[\nabla_\gamma \min(0, -1 - H(F_\gamma(G_\phi(Z))))] \\ &\approx \frac{1}{K}\sum_{k=1}^K [\nabla_\gamma \min(0, -1 + H(F_\gamma(X_k)))] + \frac{1}{K}\sum_{k=1}^K [\nabla_\gamma \min(0, -1 - H(F_\gamma(G_\phi(Z_k))))],\end{aligned}$$

where $X_1, \ldots, X_K$ are drawn i.i.d from $\mu^{\otimes m}$ and $Z_1, \ldots, Z_K$ are drawn i.i.d from $\nu_0^{\otimes m}$.

Settings. We use the following neural networks for G_ϕ and F_γ:

- **CIFAR10**:



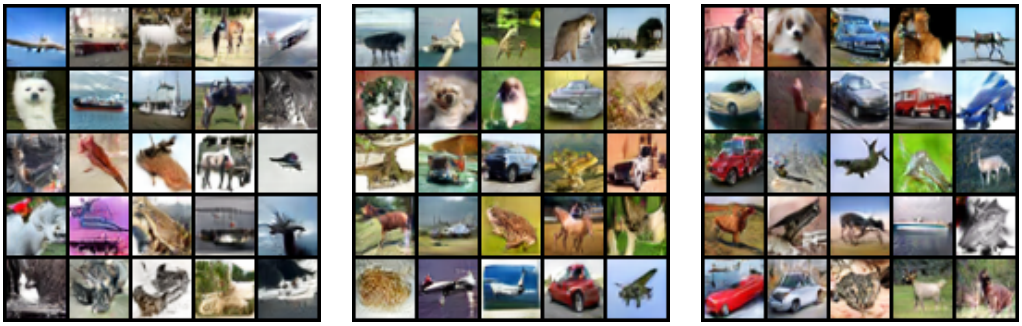


Figure 7: Random generated images of distances on CIFAR10.

- G_ϕ: $z \in \mathbb{R}^{128} (\sim \nu_0 : \mathcal{N}(0,1)) \to 4 \times 4 \times 256$(Dense, Linear) $\to$ ResBlock up 256 $\to$ ResBlock up 256 $\to$ ResBlock up 256 $\to$ BN, ReLU, $\to$ 3×3 conv, 3 Tanh .
- F_{γ_1}: $x \in [-1,1]^{32\times32\times3} \to$ ResBlock down 128 $\to$ ResBlock down 128 $\to$ ResBlock down 128 $\to$ ResBlock 128 $\to$ ResBlock 128.
- F_{γ_2}: $\boldsymbol{x} \in \mathbb{R}^{128\times8\times8} \to$ ReLU $\to$ Global sum pooling(128) $\to$ 1(Spectral normalization).
- $F_\gamma(x) = (F_{\gamma_1}(x), F_{\gamma_2}(F_{\gamma_1}(x)))$ and $H(F_\gamma(x)) = F_{\gamma_2}(F_{\gamma_1}(x))$.

- **CelebA.**
  - G_ϕ: $z \in \mathbb{R}^{128} (\sim \nu_0 : \mathcal{N}(0,1)) \to 4 \times 4 \times 256$(Dense, Linear) $\to$ ResBlock up 256 $\to$ ResBlock up 256 $\to$ ResBlock up 256 $\to$ ResBlock up 256 $\to$ BN, ReLU, $\to 3 \times 3$ conv, 3 Tanh .
  - F_{γ_1}: $\boldsymbol{x} \in [-1,1]^{32\times32\times3} \to$ ResBlock down 128 $\to$ ResBlock down 128 $\to$ ResBlock down 128 $\to$ ResBlock 128 $\to$ ResBlock 128.
  - F_{γ_2}: $\boldsymbol{x} \in \mathbb{R}^{128\times8\times8} \to$ ReLU $\to$ Global sum pooling(128) $\to$ 1(Spectral normalization).
  - $F_\gamma(x) = (F_{\gamma_1}(x), F_{\gamma_2}(F_{\gamma_1}(x)))$ and $H(F_\gamma(x)) = F_{\gamma_2}(F_{\gamma_1}(x))$.

For all datasets, the number of training iterations is set to 50000. We update the generator G_ϕ each 5 iterations while we update the feature function F_γ every iteration. The mini-batch size m is set 128 in all datasets. The learning rate for G_ϕ and F_γ is 0.0002 and the optimizer is Adam [24] with parameters $(\beta_1, \beta_2) = (0, 0.9)$. We use the order $p = 2$ for all sliced Wasserstein variants. We use 50000 random samples from estimated generative models G_ϕ for computing the FID scores and the Inception scores. In evaluating FID scores, we use all training samples for computing statistics of datasets[1].

Generated images. We show generated images on CIFAR10 and CelebA from different generative models trained by different distances in Figure 7 and in Figure 8 in turn. Overall, images are visually consistent with the quantitative FID scores in Table 2.

[1]We evaluate the scores based on the code from `https://github.com/GongXinyuu/sngan.pytorch`.

Table 5: Summary of FID and IS scores of methods on CIFAR10 (32x32), and CelebA (64x64)

Method	CIFAR10 (32x32)		CelebA (64x64)
	FID (↓)	IS (↑)	FID (↓)
iMSW (L=100,T=10,M=0,N=1)	14.61±0.72	8.15±0.15	9.73±0.33
iMSW (L=100,T=10,M=9,N=1)	14.16±1.11	8.17±0.07	9.10±0.34
iMSW (L=100,T=10,M=5,N=1)	13.93±0.21	8.15±0.05	9.49±0.52
iMSW (L=100,T=10,M=0,N=2)	14.33±0.32	8.15±0.06	8.99±0.64
iMSW (L=10,T=100,M=0,N=1)	14.26±0.74	8.15±0.07	8.89±0.23
iMSW (L=10,T=100,M=99,N=1)	14.50±0.70	8.12±0.08	9.55±0.35
iMSW (L=10,T=100,M=50,N=1)	14.41±0.58	8.12±0.06	9.46±0.73
iMSW (L=10,T=100,M=0,N=2)	14.65±0.01	8.11±0.06	9.49±0.39

Figure 8: Random generated images of distances on CelebA.

Studies on hyperparameters. We run some additional settings of iMSW to investigate the performance of the burning thinning technique and to compare the role of L and T in Table 5. First, we see that burning and thinning helps to improve FID score and IS score on CIFAR10 and CelebA in the settings of $L = 100, T = 10$. It is worth noting that the original purpose of burning and thinning is to reduce computational complexity and memory complexity. The side benefit of improving performance requires more investigation that is left for future work. In addition, we find that for the same number of total projections 1000 without burning and thinning, the setting of $L = 10, T = 100$ is better than the setting of $L = 100, T = 10$ on CIFAR10. However, the reverse direction happens on CelebA. Therefore, on different datasets, it might require hyperparameter tunning for finding the best setting of the number of projections L and the number of timesteps T.