Regularized Distribution Matching Distillation for One-step Unpaired Image-to-Image Translation

Denis Rakitin [*1] **Ivan Shchekotov** [*1 2] **Dmitry Vetrov** [3]

Abstract

Diffusion distillation methods aim to compress the diffusion models into efficient one-step generators while trying to preserve quality. Among them, Distribution Matching Distillation (DMD) offers a suitable framework for training general-form one-step generators, applicable beyond unconditional generation. In this work, we introduce its modification, called Regularized Distribution Matching Distillation, applicable to unpaired image-to-image problems. We demonstrate its empirical performance in application to several translation tasks, including 2D examples and I2I between different image datasets, where it performs on par or better than multi-step diffusion baselines.

1. Introduction

One of the global problems of contemporary generative modeling consists of solving the so-called generative learning trilemma (Xiao et al., 2021). It states that a perfect generative model should possess three desirable properties: high generation quality, mode coverage/diversity of samples and efficient inference. Today, most model families tend to have only 2 of the 3. Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) have fast inference and produce high-quality samples but tend to underrepresent some modes of the data set (Metz et al., 2016; Arjovsky et al., 2017). Variational Autoencoders (VAEs) (Kingma & Welling, 2013; Rezende et al., 2014) efficiently produce diverse samples while suffering from insufficient generation quality. Finally, diffusion-based generative models (Ho et al., 2020; Song et al., 2020; Dhariwal & Nichol, 2021; Karras et al., 2022) achieve SOTA generative metrics and visual quality yet require running a high-cost multi-step inference procedure.

Satisfying these three properties is essential in numerous generative computer vision tasks beyond unconditional generation. One is image-to-image (I2I) translation (Isola et al., 2017; Zhu et al., 2017), which consists of learning a mapping between two distributions that preserves the cross-domain properties of an input object while appropriately changing its source-domain features to match the target. Most examples, like transforming cats into dogs (Choi et al., 2020) or human faces into anime (Korotin et al., 2022) belong to the *unpaired* I2I because they do not assume ground truth pairs of objects in the data set. As in unconditional generation, unpaired I2I methods were previously centered around GANs (Huang et al., 2018; Park et al., 2020; Choi et al., 2020; Zheng et al., 2022), but now tend to be competed and surpassed by diffusion-based counterparts (Choi et al., 2021; Meng et al., 2021; Zhao et al., 2022; Wu & De la Torre, 2023). Most of these methods build on top of the original diffusion sampling procedure and tend to have high generation time as a consequence.

Since diffusion models succeed in both desirable qualitative properties of the trilemma, one could theoretically obtain samples of the desired quality level given sufficient computational resources. It makes the acceleration of diffusion models an appealing approach to satisfy all of the aforementioned requirements, including efficient inference.

Recently introduced diffusion distillation techniques (Song et al., 2023; Kim et al., 2023b; Sauer et al., 2023) address this challenge by compressing diffusion models into one-step students with (hopefully) similar qualitative and quantitative properties. Among them, Distribution Matching Distillation (DMD) (Yin et al., 2023; Nguyen & Tran, 2023) offers an expressive and general framework for training free-form generators based on techniques initially introduced for text-to-3D (Poole et al., 2022; Wang et al., 2024). *Free-form* here means that the method does not make any assumptions about the generator's structure and distribution at the input. This crucial observation opens a large space for its applications beyond the $noise \rightarrow data$ problems.

In this work, we introduce the modification of DMD, called

[*]Equal contribution [1]HSE University, Moscow, Russia [2]Skolkovo Institute of Science and Technology, Moscow, Russia [3]Constructor University, Bremen, Germany. Correspondence to: Denis Rakitin <rakitindenis32@gmail.com>, Ivan Shchekotov <ivanxshchekotov@gmail.com>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria. Copyright 2024 by the author(s).

Figure 1. Illustration of performance of the proposed RDMD model on $cat \to wild$ translation problem/ from the AFHQv2 (Choi et al., 2020) data set.

Regularized Distribution Matching Distillation (RDMD), that applies to the unpaired I2I problems. To achieve this, we replace the generator's input noise with the source data samples to further translate them into the target. We maintain correspondence between the generator's input and output by regularizing the objective with the transport cost between them. As our main contributions, we

1. Propose a one-step diffusion-based method for unpaired I2I;
2. Theoretically verify it by establishing its connection with optimal transport (Villani et al., 2009; Peyré et al., 2019);
3. Ablate its qualitative properties and demonstrate its generation quality on 2D and image-to-image examples, where it obtains comparable or better results than the multi-step counterparts.

2. Background

2.1. Diffusion Models

Diffusion models (Song & Ermon, 2019; Ho et al., 2020) are a class of models that sequentially perturb data distribution p^{data} with Gaussian noise, transforming it into some tractable unstructured distribution, which contains no information about initial domain.

Using this distribution as a prior and reversing the process by progressively removing the noise yields a sampling procedure from p^{data}. A convenient way to formalize diffusion models is through stochastic differential equations (SDEs) (Song et al., 2020), which describe continuous-time stochastic dynamics of particles. The forward process is commonly defined as the Variance Exploding (VE) SDE[1]

$$\mathrm{d}\boldsymbol{x}_t = g(t)\mathrm{d}\boldsymbol{w}_t, \tag{1}$$

[1]The other popular forward processes (e.g. VP-SDE) can be obtained by scaling the VE-SDE.

where $t \in [0, T]$, $\boldsymbol{x}_0 \sim p^{\text{data}}$, $g(\cdot)$ is the scalar diffusion coefficient and $\mathrm{d}\boldsymbol{w}_t$ is the differential of a standard Wiener process. We denote by $p_t(\boldsymbol{x}_t)$ marginal distribution of $\boldsymbol{x}_t$, so that $p^{\text{data}}(\boldsymbol{x}_0) = p_0(\boldsymbol{x}_0)$. p_T acts as an unstructured prior distribution that we can sample from.

Conveniently, SDE dynamics can be represented via a deterministic counterpart given by an ordinary differential equation (ODE), which yields the same marginal distributions $p_t(\boldsymbol{x}_t)$ as in Equation 1, given the same initial distribution $p_0(\boldsymbol{x}_0) = p^{\text{data}}(\boldsymbol{x}_0)$:

$$\mathrm{d}\boldsymbol{x}_t = -\frac{1}{2}g^2(t)\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)\mathrm{d}t, \tag{2}$$

where $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ is called the *score function* of $p_t(\boldsymbol{x}_t)$. Equation 2 is also called Probability Flow ODE (PF-ODE). The ODE formulation allows us to obtain a backward process by simply reversing velocity of the particle. In particular, we can obtain samples from p^{data} by taking $\boldsymbol{x}_T \sim p_T$ and running the PF-ODE backwards in time, given access to the score function.

However, in the case of generative modeling $\nabla_{\boldsymbol{x}} \log p_t(\boldsymbol{x}_t)$ is intractable due to p^{data} being intractable, and thus cannot be used directly in Equation 2. Under mild regularity conditions, the unconditional score can be expressed by:

$$\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t) = \mathbb{E}_{p_{0|t}(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\boldsymbol{s}_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)\right], \tag{3}$$

where $\boldsymbol{s}_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \nabla_{\boldsymbol{x}_t} \log p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)$ is the conditional distribution (also called perturbation kernel) and $p_{0|t}(\boldsymbol{x}_0|\boldsymbol{x}_t)$ is the corresponding posterior distribution. The perturbation kernel in the case of VE-SDE corresponds to simply adding an independent Gaussian noise:

$$p_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{x}_0, \sigma_t^2 I),\ \sigma_t^2 = \int_0^t g^2(s)\mathrm{d}s. \tag{4}$$

Denoising Score Matching (DSM) (Vincent, 2011) utilizes Equation 3 and approximates $\nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$ with the score

model $\boldsymbol{s}_t^\theta(\boldsymbol{x}_t)$ via L2-regression minimization:

$$\int_0^T \beta_t \, \mathbb{E}_{p_{0,t}(\boldsymbol{x}_0,\boldsymbol{x}_t)} \|\boldsymbol{s}_t^\theta(\boldsymbol{x}_t) - \boldsymbol{s}_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)\|^2 \mathrm{d}t \to \min_\theta, \quad (5)$$

where β_t is some positive weighting function. The minimum in the Equation 5 is obtained at $\boldsymbol{s}_t^\theta(\boldsymbol{x}_t) = \nabla_{\boldsymbol{x}_t} \log p_t(\boldsymbol{x}_t)$. Given a suitable parameterization of the score network, DSM objective is equivalent to

$$\int_0^T \beta_t \, \mathbb{E}_{p_{0,t}(\boldsymbol{x}_0,\boldsymbol{x}_t)} \|D_t^\theta(\boldsymbol{x}_t) - \boldsymbol{x}_0\|^2 \mathrm{d}t \to \min_\theta, \quad (6)$$

where D_t^θ is called the denoising network (or simply denoiser) and is related to the score network via $\boldsymbol{s}_t^\theta(\boldsymbol{x}_t) = \left(\boldsymbol{x}_t - D_t^\theta(\boldsymbol{x}_t)\right)/\sigma_t^2$. Therefore, Denoising Score Matching procedure consists of learning to denoise images at various noise levels.

Having obtained $\boldsymbol{s}_t^\theta(\boldsymbol{x}_t)$, we solve Equation 2 backward in time, starting from $x_T \sim \mathcal{N}(0, \sigma_T^2 I)$ to obtain approximate samples from p^{data}.

2.2. Distribution Matching Distillation

Distribution Matching Distillation (Yin et al., 2023) is the core technique of this paper. Essentially, it aims to train a generator $G_\theta(\boldsymbol{z})$ on matching the given distribution p^{real}. Its input $\boldsymbol{z}$ is assumed to come from a tractable input distribution p^{noise}. Formally, matching two distributions can be achieved by optimizing the KL divergence between the distribution[2] p^θ of $G_\theta(\boldsymbol{z})$ and the data distribution p^{real}:

$$\mathrm{KL}(p^\theta \,\|\, p^{\text{real}}) = \mathbb{E}_{p^{\text{noise}}(\boldsymbol{z})} \log \frac{p^\theta(G_\theta(\boldsymbol{z}))}{p^{\text{real}}(G_\theta(\boldsymbol{z}))} \to \min_\theta \quad (7)$$

Differentiating it by the parameters θ, using the chain rule, one encounters a summand, containing the difference $\boldsymbol{s}^\theta(G_\theta(\boldsymbol{z})) - \boldsymbol{s}^{\text{real}}(G_\theta(\boldsymbol{z}))$ between the score functions of the corresponding distributions [3]. The pure data score function can be very non-smooth due to the Manifold Hypothesis (Tenenbaum et al., 2000) and is generally hard to train (Song & Ermon, 2019), so the authors make the problem accessible through the diffusion framework. To this end, they replace the original loss with an ensemble of KL divergences between distributions, perturbed by the forward diffusion process:

$$\int_0^T \omega_t \, \mathrm{KL}\Big(p_t^\theta \,\|\, p_t^{\text{real}}\Big) \mathrm{d}t, \quad (8)$$

Here, ω_t is a weighting function, p_t^θ and p_t^{real} are the perturbed versions of the generator distribution and p^{real} up to the time step t. In theory, the minima of Equation 8 objective coincides (Wang et al., 2024, Thm. 1) with the original minima from Equation 7. Meanwhile in practice taking the gradient of the new loss, which can be equivalently written as

$$\int_0^T \omega_t \, \mathbb{E}_{\mathcal{N}(\varepsilon|0,I)p^{\text{noise}}(\boldsymbol{z})} \log \frac{p_t^\theta(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon)}{p_t^{\text{real}}(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon)} \mathrm{d}t, \quad (9)$$

results in obtaining difference $\boldsymbol{s}_t^\theta(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon) - \boldsymbol{s}_t^{\text{real}}(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon)$, which can be approximated by the diffusion models.

Given this, authors approximate $\boldsymbol{s}_t^{\text{real}}$ with the pre-trained diffusion model, which we will denote $\boldsymbol{s}_t^{\text{real}}$ as well with a slight abuse of notation. The whole procedure now can be considered as distillation of $\boldsymbol{s}_t^{\text{real}}$ into G_θ. At the same time, $\boldsymbol{s}_t^\theta$ is the score of the noised distribution of the generator, which is intractable and therefore approximated by an additional "fake" diffusion model $\boldsymbol{s}_t^\phi$ and the corresponding denoiser D_t^ϕ. It is trained on the standard denoising score matching objective with the generator's samples at the input. The joint training procedure is essentially the coordinate descent

$$\begin{cases} \int_0^T \omega_t \, \mathbb{E}_{\varepsilon,\boldsymbol{z}} \log \dfrac{p_t^\phi(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon)}{p_t^{\text{real}}(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon)} \mathrm{d}t \to \min_\theta; \\ \int_0^T \beta_t \, \mathbb{E}_{\varepsilon,\boldsymbol{z}} \|D_t^\phi(G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon) - G_\theta(\boldsymbol{z})\|^2 \, \mathrm{d}t \to \min_\phi, \end{cases} \quad (10)$$

where the stochastic gradient with respect to the fake network is calculated by backpropagation and the generator's stochastic gradient is calculated directly as

$$\omega_t \left(\boldsymbol{s}_t^\phi - \boldsymbol{s}_t^{\text{real}}\right) \nabla_\theta G_\theta(\boldsymbol{z}), \quad (11)$$

where the scores are evaluated in the point $G_\theta(\boldsymbol{z}) + \sigma_t\varepsilon$. Minimization of the fake network's objective ensures $\boldsymbol{s}_t^\phi = \boldsymbol{s}_t^\theta \Leftrightarrow p_t^\phi = p_t^\theta$. At this condition, the generator's objective is equal to the original ensemble of KL divergences from Equation 8, minimizing which solves the initial problem and implies $p^\theta = p^{\text{real}}$.

2.3. Unpaired I2I and optimal transport

The problem of unpaired I2I consists of learning a mapping G between the *source* distribution $p^{\mathcal{S}}$ and the *target* distribution $p^{\mathcal{T}}$ given the corresponding independent data sets of samples. When optimized, the mapping should appropriately adapt $G(\boldsymbol{x})$ to the target distribution $p^{\mathcal{T}}$, while preserving the input's cross-domain features. However, from the first glance it is unclear what the preservation of cross-domain properties should be like.

[2]The superscript θ in p^θ does not mean introducing the additional neural model of density but is rather used to emphasize its dependence on the generator.

[3]Note that there is one more summand, which contains the gradient $\nabla_\theta \log p^\theta$ with respect to the log-density parameters. We do not discuss how to approximate it, because it will be further omitted.

One way to look at that formally is by introducing the notion of "transportation cost" $c(\cdot,\cdot)$ between the generator's input and output and saying that it should not be too large on average. In a practical I2I setting, we can choose $c(\cdot,\cdot)$ as any reasonable distance between images or their features that we aim to preserve, e.g. pixel-wise distance or difference between LPIPS (Zhang et al., 2018) embeddings.

Monge optimal transport (OT) problem (Villani et al., 2009; Santambrogio, 2015) follows this reasoning and aims at finding the mapping with the least average transport cost among all the mappings that fit the target $p^{\mathcal{T}}$:

$$\inf_{G} \left\{ \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} c(\boldsymbol{x}, G(\boldsymbol{x})) \mid G(\boldsymbol{x}) \sim p^{\mathcal{T}} \right\}, \tag{12}$$

which can be seen as a mathematical formalization of the I2I task.

Under mild constraints, in the case when $p^{\mathcal{S}}$ and $p^{\mathcal{T}}$ have densities, the optimal transport map G^* is bijective, differentiable, has differentiable inverse and thus satisfies the change of variables formula $p^{\mathcal{S}}(\boldsymbol{x}) = p^{\mathcal{T}}(G^*(\boldsymbol{x}))|\det(\nabla G^*(\boldsymbol{x}))|$. This highly non-linear change of variables condition gives insight into why it is notoriously challenging to optimize Equation 12 directly.

3. Methodology

Our main goal is to adapt the DMD method for the unpaired I2I between an arbitrary source distribution $p^{\mathcal{S}}$ and target distribution $p^{\mathcal{T}}$.

3.1. Regularized Distribution Matching Distillation

First, we note that the construction of DMD requires only having samples from the input distribution. Given this, we replace the Gaussian input p^{noise} by $p^{\mathcal{S}}$, the data distribution p^{data} by $p^{\mathcal{T}}$ and aim at optimizing

$$\mathcal{L}(\theta) = \int_0^T \omega_t \, \mathrm{KL}\Big(p_t^{\theta} \,\|\, p_t^{\mathcal{T}}\Big) \mathrm{d}t = \\ = \int_0^T \omega_t \, \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})\mathcal{N}(\varepsilon|0,I)} \log \frac{p_t^{\theta}(G_\theta(\boldsymbol{x}) + \sigma_t \varepsilon)}{p_t^{\mathcal{T}}(G_\theta(\boldsymbol{x}) + \sigma_t \varepsilon)} \mathrm{d}t, \tag{13}$$

where p_t^{θ} and $p_t^{\mathcal{T}}$ are now respectively the distribution of the generator output $G_\theta(\boldsymbol{x})$ and the target distribution $p^{\mathcal{T}}$, perturbed by the forward process up to the timestep t.

Optimizing the objective in Equation 13, one obtains a generator, which takes $\boldsymbol{x} \sim p^{\mathcal{S}}$ and outputs $G_\theta(\boldsymbol{x}) \sim p^{\mathcal{T}}$, so it performs the desired transfer between the two distributions. However, there are no guarantees that the input and the output will be related. Similarly to the OT problem (Equation 12), we fix the issue by penalizing the transport cost between them. We obtain the following objective

$$\mathcal{L}^{\lambda}(\theta) = \mathcal{L}(\theta) + \lambda \, \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} c\left(\boldsymbol{x}, G_\theta(\boldsymbol{x})\right) \to \min_{\theta}, \tag{14}$$

where $c(\cdot,\cdot)$ is the cost function, which describes the object properties that we aim to preserve after transfer, and λ is the regularization coefficient. Choosing the appropriate λ will result in finding a balance between fitting the target distribution and preserving properties of the input.

As in DMD, we assume that the perturbed target distributions are represented by a pre-trained diffusion model $\boldsymbol{s}_t^{\mathcal{T}}$ and approximate the generator distribution score $\boldsymbol{s}_t^{\theta}$ by the additional fake diffusion model $\boldsymbol{s}_t^{\phi}$. Analogous to the DMD procedure (Equation 10), we perform the coordinate descent in which, however, the generator objective is now regularized. We call the procedure *Regularized Distribution Matching Distillation* (RDMD). Formally, we optimize

$$\begin{cases} \int_0^T \omega_t \, \mathbb{E}_{\varepsilon,\boldsymbol{x}} \log \dfrac{p_t^{\phi}(G_\theta(\boldsymbol{x}) + \sigma_t \varepsilon)}{p_t^{\mathcal{T}}(G_\theta(\boldsymbol{x}) + \sigma_t \varepsilon)} \mathrm{d}t \\ \qquad\qquad + \lambda \, \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} c\left(\boldsymbol{x}, G_\theta(\boldsymbol{x})\right) \to \min_{\theta}; \\ \int_0^T \beta_t \, \mathbb{E}_{\varepsilon,\boldsymbol{x}} \|D_t^{\phi}(G_\theta(\boldsymbol{x}) + \sigma_t \varepsilon) - G_\theta(\boldsymbol{x})\|^2 \, \mathrm{d}t \to \min_{\phi}. \end{cases} \tag{15}$$

Given the optimal fake score $\boldsymbol{s}_t^{\phi}$, the generator's objective becomes equal to the desired loss in Equation 14, which validates the procedure.

3.2. Analysis of the method

The optimization problem in Equation 14 can be seen as the soft-constrained optimal transport, which balances between satisfying the output distribution constraint and preserving the original image properties. Moreover, if one takes $\lambda \to 0$, the objective essentially becomes equivalent to the Monge problem (Equation 12). It can be seen by replacing the λ coefficient before the transport cost with the $1/\lambda$ coefficient before the KL divergence. In the limit, it equals $+\infty$ whenever the generator output and the target distributions are different, which makes the corresponding problem hard-constrained and, therefore, equivalent to the original optimal transport problem. Based on this observation, we prove the following

Theorem 3.1. *Let $c(\boldsymbol{x},\boldsymbol{y})$ be the quadratic cost $\|\boldsymbol{x}-\boldsymbol{y}\|^2$ and G^{λ} be the theoretical optimum in the problem 14. Then, under mild regularity conditions, it converges in probability (with respect to $p^{\mathcal{S}}$) to the optimal transport map G^*, i.e.*

$$G^{\lambda} \xrightarrow[\lambda \to 0]{p^{\mathcal{S}}} G^*. \tag{16}$$

The detailed proof can be found in Appendix A. Informally, it means that the optimal transport map can be approximated

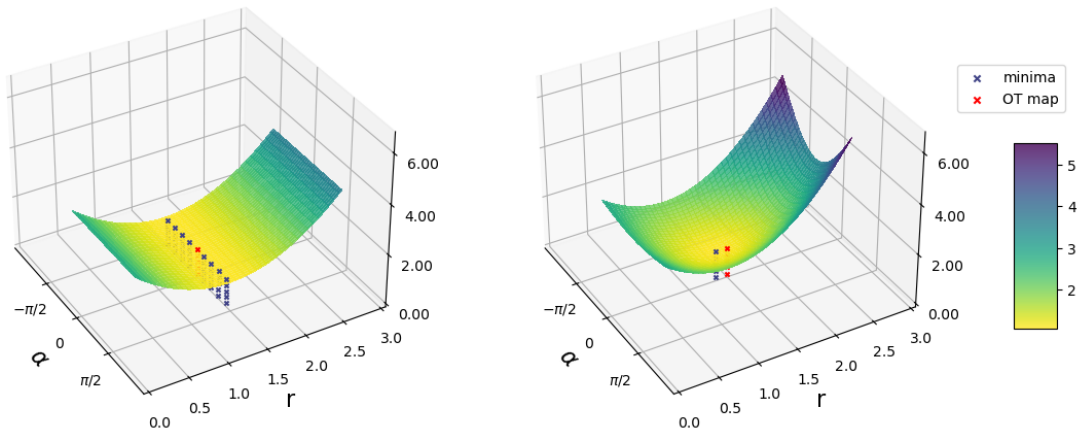

Figure 2. Comparison of the DMD loss surfaces without (left) and with (right) transport cost regularization on a toy problem of translating $\mathcal{N}(0, I)$ to $\mathcal{N}(0, 1.5^2 I)$. We set the regularization coefficient $\lambda = 0.2$. The generator is parameterized as $r \cdot C(\alpha)$, where $C(\alpha)$ is the rotation matrix, corresponding to the angle α. Minima at the left contains all orthogonal matrices, multiplied by $\sigma = 1.5$, while the minimum at the right is attained in the only point, which is close, but not equal, to the OT map. The surfaces are moved up for the sake of visualization.

by the RDMD generator, trained on Equation 15, given a sufficiently small regularization coefficient, enough capacity of the architecture, and convergence of the optimization algorithm.

This result is important to examine from another angle. It is ideologically similar to the L_2 regularization for over-parameterized least squares regression. The original least squares, in this case, have a manifold of solutions. At the same time, by adding L_2 weight penalty and taking the limit as the regularization coefficient goes to zero, one obtains a solution with the least norm based on the Moore-Penrose pseudo-inverse (Moore, 1920; Penrose, 1955). In our case, numerous maps may be optimal in the original DMD procedure, since it only requires matching the distribution at output. However, taking the limit when $\lambda \to 0$, one obtains a feasible solution with the least transport cost.

We demonstrate this effect on a toy problem of translating $\mathcal{N}(0, I)$ to $\mathcal{N}(0, \sigma^2 I)$ and consider linear generator $G(\boldsymbol{x}) = A\boldsymbol{x}$. The solution to the optimal transport problem with the quadratic cost $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$ is $A = \sigma I$. For the DMD optimization problem without regularization, minima are obtained at the manifold of orthogonal matrices multiplied by σ: $AA^\top = \sigma^2 I$. However, if one adds the regularization, the minimum compresses into one point at the cost of introducing the bias relatively to the true OT map. We illustrate this by comparing the loss surface with and without regularization in Figure 2.

4. Related work

In this section, we give an overview of the existing methods for solving unpaired I2I including GANs, diffusion-based methods, and methods based on optimal transport.

GANs were the prevalent paradigm in the unpaired I2I for a long time. Among other methods, CycleGAN (Zhu et al., 2017) and the concurrent DualGAN (Yi et al., 2017), DiscoGAN (Kim et al., 2017) utilized the cycle-consistency paradigm, consisting in training the transfer network along with its inverse and optimizing the consistency term along with the adversarial loss. It gave rise to the whole family of two-sided methods, including UNIT (Liu et al., 2017) and MUNIT (Huang et al., 2018) that divide the encoding into style-space and content-space and SCAN (Li et al., 2018) that splits the procedure into coarse and fine stages.

The one-side GAN-based methods aim to train I2I without learning the inverse for better computational efficiency. DistanceGAN (Benaim & Wolf, 2017) achieves it by learning to preserve the distance between pairs of samples, GCGAN (Fu et al., 2019) imposes geometrical consistency constraints, and CUT (Park et al., 2020) uses the contrastive loss to maximize the patch-wise mutual information between input and output.

Diffusion-based I2I models mostly build on modifying the diffusion process using the source image. SDEdit (Meng et al., 2021) initializes the reverse diffusion process for target distribution with the noisy source picture instead of the pure noise to maintain similarity. Many methods guide (Ho & Salimans, 2022; Epstein et al., 2023) the target diffusion process. ILVR (Choi et al., 2021) adds the correc-

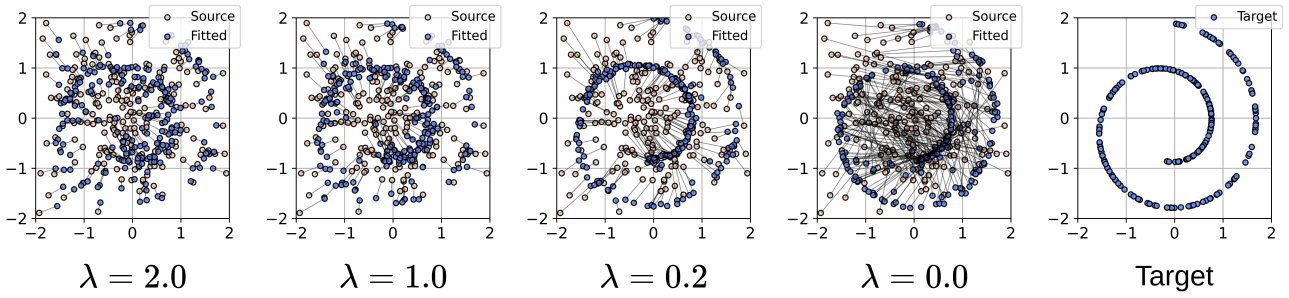

Figure 3. Visualization of RDMD mappings on $Gaussian \to Swissroll$ with different choices of the regularization coefficient λ.

tion that enforces the current noisy sample to resemble the source. EGSDE (Zhao et al., 2022) trains a classifier between domains and encourages dissimilarity between the embeddings, corresponding to the source image and the current diffusion process state. At the same time, it enforces a small distance between their downsampled versions, which allows for a balance between faithfulness and realism. The other diffusion-based approaches include two-sided methods based on the concatenation of two diffusion models with deterministic sampling (Su et al., 2022; Wu & De la Torre, 2023).

Optimal transport (Villani et al., 2009; Peyré et al., 2019) is another useful framework for the unpaired I2I. Methods based on it usually reformulate the OT problem (Equation 12) and its modifications as Entropic OT (EOT) (Cuturi, 2013) or Schrödinger Bridge (SB) (Föllmer, 1988) to be accessible in practice. In particular, NOT (Korotin et al., 2022), ENOT (Gushchin et al., 2024), and NSB (Kim et al., 2023a) use the Lagrangian multipliers formulation of the distribution matching constraint, which results in simulation-based adversarial training. The other methods obtain (partially) simulation-free techniques by iteratively refining the stochastic process between two distributions. In the works (De Bortoli et al., 2021; Vargas et al., 2021) refinement consists of learning the time-reversal with the corresponding initial distribution (source or target). The newer methods are based on Flow Matching (Lipman et al., 2022; Tong et al., 2023; Albergo & Vanden-Eijnden, 2022) and the corresponding Rectification (Liu et al., 2022; Shi et al., 2024; Liu et al., 2023) procedure. While being theoretically sound, most of these methods work well for smaller dimensions (Korotin et al., 2023) but suffer from computationally hard training in large-scale scenarios.

5. Experiments

This section presents the experimental results on 2 unpaired translation tasks. Section 5.1 is devoted to the toy 2D experiment. In Section 5.2 we compare our method with the diffusion-based baselines on the translation problem between cats and wild animals from the AFHQv2 data set (Choi et al., 2020).

In all the experiments, we use the forward diffusion process with variance $\sigma_t = t$ and $T = 80.0$ as in the paper (Karras et al., 2022). We parameterize all the diffusion models with the denoiser networks $D_\sigma(\boldsymbol{x})$, conditioned on the noise level σ, and optimize Equation 6 to train the target diffusion model. As for the RDMD procedure, we optimize Equation15, where the gradient with respect to the generator parameters is calculated analogously to Equation 11. The transport cost $c(\boldsymbol{x}, \boldsymbol{y})$ is chosen as the squared difference norm $\|\boldsymbol{x} - \boldsymbol{y}\|^2$. The average transport cost, reported in the figures, is calculated as the square root of the MSE between all input and output images for the sake of interpretability.

We use the same architecture for all networks: target score, fake score, and generator. We utilize the pre-trained target score in two ways. First, we initialize the fake model with its copy. Second, we initialize the generator $G_\theta(\boldsymbol{x})$ with the same copy $D^{\text{real}}_\sigma(\boldsymbol{x})$, but with a fixed $\sigma \in [0, T]$ (since the generator is one-step). The denoiser parameterization is trained to predict the target domain's clean images, therefore, such initialization should significantly speed up convergence and nudge the model to utilize the information about the target domain more efficiently (Nguyen & Tran, 2023; Yin et al., 2023). We explore the initialization of σ for I2I in Appendix B. The additional training details can be found in Appendix C.

5.1. Toy Experiment

We validate the qualitative properties of the RDMD method on 2-dimensional *Gaussian* $\to$ *Swissroll*. In this setting, we explore the effect of varying the regularization coefficient λ on the trained transport map G_θ. In particular, we study its impact on the transport cost and fitness to the target distribution $p^{\mathcal{T}}$.

In the experiment, both source and target distributions are represented with 5000 independent samples. We use the same small MLP-based architecture (Shi et al., 2024) for all

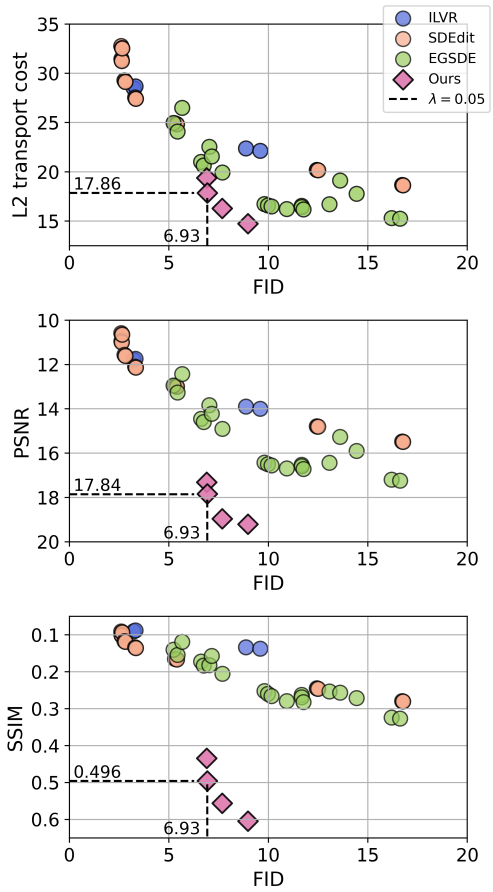

Figure 4. Comparison of RDMD with diffusion-based baselines. The figure demonstrates the tradeoff between generation quality (FID↓) and the difference between the input and output (L2↓, PSNR↑, SSIM↑). RDMD gives an overall better tradeoff given fairly strict requirements on the transport cost. In the cases of PSNR and SSIM, the y-axis is swapped for the sake of identical readability with the first plot (left is better, low is better).

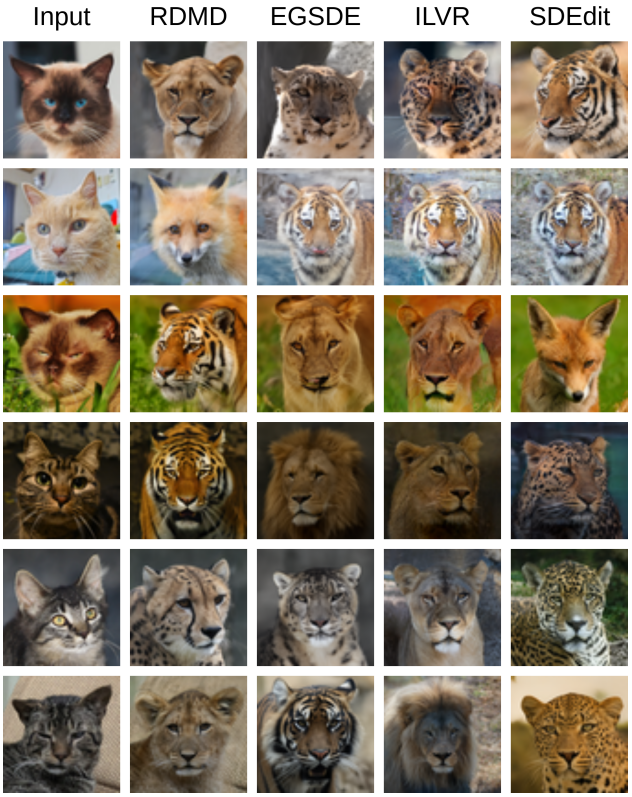

Figure 5. Visual comparison of RDMD with diffusion-based baselines.

the networks.

The main results are presented in Figure 3. The standard DMD ($\lambda = 0.0$) learns a transport map with several intersections when demonstrated as the set of lines between the inputs and the outputs. This observation means that the learned map is not OT, because it is not cycle-monotone (McCann, 1995). Increasing λ yields fewer intersections, which can be used as a proxy evidence of optimality. At the same time, the generator output distribution becomes farther and farther from the desired target. The results show the importance of choosing the appropriate λ to obtain a better trade-off between the two properties. Here, the regularization coefficient $\lambda = 0.2$ offers a good trade-off by having small intersections and producing output distribution close to the target.

5.2. Cat to Wild

Finally, we compare the proposed RDMD method with the diffusion-based baselines ILVR (Choi et al., 2021), SDEdit (Meng et al., 2021), and EGSDE (Zhao et al., 2022) on the 64×64 $Cat \rightarrow Wild$ translation problem, based on the AFHQv2 (Choi et al., 2020) data set. Comparison with the diffusion-based models makes the setting fair since it allows to utilize the same pre-trained target diffusion model for all of the methods. We stress, however, that the GAN-based methods mostly demonstrate results inferior to EGSDE (Zhao et al., 2022) in terms of FID and PSNR at the same data set with resolution 256×256.

We pre-train the target diffusion model using the EDM (Karras et al., 2022) architecture with the hyperparameters used by the authors on the AFHQv2 data set. Our pre-trained model achieves FID equal to 2.0. We initialize the model with $\sigma = 1.0$ based on the observations from Section B and train 5 RDMD generators, corresponding to the regularization coefficients $\{0.001, 0.02, 0.05, 0.1, 0.2\}$. We slightly adapt the official baseline implementations for compatibility with the EDM setting. The EGSDE classifier is trained analogous to the paper: it is initialized from the Dhariwal UNet (Dhariwal & Nichol, 2021), pre-trained on the Ima-

geNet (Deng et al., 2009) 64×64. For each of the baselines, we run a grid of hyperparameters. The detailed hyperparameter values can be found in Appendix C.3.

We report our main quantitative results in Figure 5 and compare the achieved faithfulness-quality trade-off with the baselines. The quality metric is FID, the faithfulness metrics are L2/PSNR/SSIM. Among these metrics, L2 is the least convenient for our method. Nevertheless, RDMD achieves a better trade-off given at least moderately strict requirements on the transport cost: all of our models beat all the baselines in the L2 range between 12.5 and 20.0. In all cases, our model achieves strictly higher SSIM and almost strictly higher PSNR. We note, however, that if the lower FID is preferable over the transport cost (L2 values around $22.5 - 27.5$), then it might be better to use one of the baselines. An example of a map with a high OT cost (25.0) and low FID (5.4) is SDEdit on Figure 5.

Finally, we present a visual comparison between the methods. To this end, we randomly choose 6 pictures from the test data set and report the corresponding outputs in Figure 5. Here, we take RDMD with $\lambda = 0.05$ that achieves (FID, L2) equal to $(6.93, 17.86)$. As for the baselines, we choose the hyperparameters (Appendix C.3) with the closest FID to the RDMD: $(8.87, 22.0)$ for ILVR, $(5.4, 25.0)$ for SDEdit, and $(7.02, 22.35)$ for EGSDE.

6. Discussion and limitations

In this paper, we propose RDMD, the novel *one-step* diffusion-based algorithm for the unpaired I2I task. This algorithm is a modification of the DMD method for diffusion distillation. The main novelty is the introduction of the transport cost regularization between the input and the output of the model, which allows to control the trade-off between faithfulness and visual quality.

From the theoretical standpoint, we prove that at low regularization coefficients, the theoretical optimum of the introduced objective is close to the optimal transport map (Thm. 3.1). Our experiments in Sec. 5.1 demonstrate how the choice of regularization coefficient affects the trained mapping and allows us to build the general intuition. In Sec. 5.2 we compare our method with the diffusion-based baselines (ILVR, SDEdit, EGSDE) and obtain better results given fair restrictions on the transport cost. The results are strictly better than all of the baselines in terms of SSIM and almost strictly superior to all of the baselines in terms of PSNR.

In terms of limitations, we admit that our theory works in the asymptotic regime, while one could derive more precise non-limit bounds. Our experimental results on $Cat \to Wild$ demonstrate the lowest FID around 6.9, while the pre-trained diffusion model has 2.01. Improving the visual quality and testing our method on high dimensions is important for future work. Furthermore, the desired feature of the method would be switching among different reg. coefficients without re-training.

Acknowledgements

The work of Denis Rakitin was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139. This research was supported in part through computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021).

References

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Benaim, S. and Wolf, L. One-sided unsupervised domain mapping. *Advances in neural information processing systems*, 30, 2017.

Bogachev, V. I. and Ruas, M. A. S. *Measure theory*, volume 1. Springer, 2007.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models. *arXiv preprint arXiv:2108.02938*, 2021.

Choi, Y., Uh, Y., Yoo, J., and Ha, J.-W. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197, 2020.

Csiszár, I. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

De Bortoli, V., Thornton, J., Heng, J., and Doucet, A. Diffusion schrödinger bridge with applications to score-based

generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

De Philippis, G. and Figalli, A. The monge–ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Donsker, M. D. and Varadhan, S. S. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36 (2):183–212, 1983.

Dudley, R. M. *Real analysis and probability*. Chapman and Hall/CRC, 2018.

Epstein, D., Jabri, A., Poole, B., Efros, A., and Holynski, A. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36: 16222–16239, 2023.

Föllmer, H. Random fields and diffusion processes. *Lect. Notes Math*, 1362:101–204, 1988.

Fu, H., Gong, M., Wang, C., Batmanghelich, K., Zhang, K., and Tao, D. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2427–2436, 2019.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Gushchin, N., Kolesov, A., Korotin, A., Vetrov, D. P., and Burnaev, E. Entropic neural optimal transport via diffusion processes. *Advances in Neural Information Processing Systems*, 36, 2024.

Ho, J. and Salimans, T. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Huang, X., Liu, M.-Y., Belongie, S., and Kautz, J. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 172–189, 2018.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017.

Kantorovitch, L. On the translocation of masses. *Management science*, 5(1):1–4, 1958.

Karras, T., Aittala, M., Aila, T., and Laine, S. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35: 26565–26577, 2022.

Kim, B., Kwon, G., Kim, K., and Ye, J. C. Unpaired image-to-image translation via neural schr\" odinger bridge. *arXiv preprint arXiv:2305.15086*, 2023a.

Kim, D., Lai, C.-H., Liao, W.-H., Murata, N., Takida, Y., Uesaka, T., He, Y., Mitsufuji, Y., and Ermon, S. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. *arXiv preprint arXiv:2310.02279*, 2023b.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pp. 1857–1865. PMLR, 2017.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Korotin, A., Selikhanovych, D., and Burnaev, E. Neural optimal transport. *arXiv preprint arXiv:2201.12220*, 2022.

Korotin, A., Gushchin, N., and Burnaev, E. Light schr\" odinger bridge. *arXiv preprint arXiv:2310.01174*, 2023.

Kostenetskiy, P., Chulkevich, R., and Kozyrev, V. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, pp. 012050. IOP Publishing, 2021.

LeCun, Y. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Li, M., Huang, H., Ma, L., Liu, W., Zhang, T., and Jiang, Y. Unsupervised image-to-image translation with stacked cycle-consistent adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 184–199, 2018.

Liero, M., Mielke, A., and Savaré, G. Optimal entropy-transport problems and a new hellinger–kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Liu, M.-Y., Breuel, T., and Kautz, J. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Liu, X., Zhang, X., Ma, J., Peng, J., et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023.

McCann, R. J. Existence and uniqueness of monotone measure-preserving maps. 1995.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

Metz, L., Poole, B., Pfau, D., and Sohl-Dickstein, J. Unrolled generative adversarial networks. *arXiv preprint arXiv:1611.02163*, 2016.

Moore, E. H. On the reciprocal of the general algebraic matrix. *Bulletin of the american mathematical society*, 26:294–295, 1920.

Nguyen, T. H. and Tran, A. Swiftbrush: One-step text-to-image diffusion model with variational score distillation. *arXiv preprint arXiv:2312.05239*, 2023.

Park, T., Efros, A. A., Zhang, R., and Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pp. 319–345. Springer, 2020.

Penrose, R. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pp. 406–413. Cambridge University Press, 1955.

Peyré, G., Cuturi, M., et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Poole, B., Jain, A., Barron, J. T., and Mildenhall, B. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Posner, E. Random coding strategies for minimum entropy. *IEEE Transactions on Information Theory*, 21(4):388–391, 1975.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pp. 1278–1286. PMLR, 2014.

Santambrogio, F. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

Sauer, A., Lorenz, D., Blattmann, A., and Rombach, R. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042*, 2023.

Shi, Y., De Bortoli, V., Campbell, A., and Doucet, A. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36, 2024.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Song, Y., Dhariwal, P., Chen, M., and Sutskever, I. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023.

Su, X., Song, J., Meng, C., and Ermon, S. Dual diffusion implicit bridges for image-to-image translation. *arXiv preprint arXiv:2203.08382*, 2022.

Tenenbaum, J. B., Silva, V. d., and Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Vargas, F., Thodoroff, P., Lamacraft, A., and Lawrence, N. Solving schrödinger bridges via maximum likelihood. *Entropy*, 23(9):1134, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Villani, C. et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Vincent, P. A connection between score matching and denoising autoencoders. *Neural Computation*, 23:1661–1674, 2011. URL https://api.semanticscholar.org/CorpusID:5560643.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H., and Zhu, J. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Wu, C. H. and De la Torre, F. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7378–7387, 2023.

Xiao, Z., Kreis, K., and Vahdat, A. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

Yi, Z., Zhang, H., Tan, P., and Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pp. 2849–2857, 2017.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T., and Park, T. One-step diffusion with distribution matching distillation. *arXiv preprint arXiv:2311.18828*, 2023.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

Zhao, M., Bao, F., Li, C., and Zhu, J. Egsde: Unpaired image-to-image translation via energy-guided stochastic differential equations. *Advances in Neural Information Processing Systems*, 35:3609–3623, 2022.

Zheng, W., Li, Q., Zhang, G., Wan, P., and Wang, Z. Ittr: Unpaired image-to-image translation with transformers. *arXiv preprint arXiv:2203.16015*, 2022.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017.

A. Theory

In this section, we aim at proving the main theoretical result of the work: solution of the soft-constrained RDMD objective converges to the solution of the hard-constrained Monge problem. Our proof is largely based on the work (Liero et al., 2018). It introduces the family of entropy-transport problems, consisting in optimizing the transport cost with soft constraints based on the divergence between the map's output distribution and the target. There are, however, differences between the problems, that prevent us from reducing the functional in Equation 14 to the entropy-transport problems. First, authors consider the case of finite non-negative measures, while we stick to the probability distributions. Second, the family of Csiszár f-divergences (Csiszár, 1967), used in (Liero et al., 2018), seemingly does not contain the integral ensemble of KL divergences, used in Equation 14. Finally, we illustrate the proof in a simpler particular setting for the narrative purposes. Nevertheless, the used ideas are very similar.

A.1. Proof outline

We start by giving a simple outline of the proof. Given a pair of source and target distributions $p^{\mathcal{S}}$ and $p^{\mathcal{T}}$, RDMD optimizes the following functional with respect to the generator G:

$$\int_0^T \omega_t \, \mathrm{KL}\left(p_t^G \,\|\, p_t^{\mathcal{T}}\right) \mathrm{d}t + \lambda \, \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} c\left(\boldsymbol{x}, G(\boldsymbol{x})\right), \tag{17}$$

where p_t^G and $p_t^{\mathcal{T}}$ are the generator distribution p^G and the target distribution $p^{\mathcal{T}}$, perturbed by the forward diffusion process up to the time step t. Our goal is to prove that the optimal generator of the regularized objective converges to the optimal transport map when $\lambda \to 0$. With a slight abuse of notation, in this section we will use a different objective

$$\mathcal{L}^{\alpha}(G) = \alpha \int_0^T \omega_t \, \mathrm{KL}\left(p_t^G \,\|\, p_t^{\mathcal{T}}\right) \mathrm{d}t + \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} c\left(\boldsymbol{x}, G(\boldsymbol{x})\right) \tag{18}$$

and consider the equivalent limit $\alpha \to +\infty$. We also define

$$\mathcal{L}^{\infty}(G) = \begin{cases} \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} c\left(\boldsymbol{x}, G(\boldsymbol{x})\right), \text{ if } p^G = p^{\mathcal{T}}; \\ +\infty, \text{ else} \end{cases} \tag{19}$$

to be the objective, corresponding to the unconditional formulation of the Monge problem (Equation 12). In this section, we will denote minimum of this objective (which is, therefore, the optimal transport map) as G^{∞} [4]

We first assume that the infimum of the objective $\mathcal{L}^{\alpha}$ is reached and define G^{α} be the optimal generator. We denote by $\{\alpha_n\}_{n=1}^{+\infty}$ an arbitrary sequence with $\alpha_n \to +\infty$. We first make two informal assumptions that need to be proved (and will be in some sence further in the section):

1. The sequence G^{α_n} converges (in some sence) to some function $\hat{G}$;
2. $\mathcal{L}^{\alpha}$ is continuous with respect to this convergence, i.e. for every convergent sequence $G_n \to G$ holds $\mathcal{L}^{\alpha}(G_n) \to \mathcal{L}^{\alpha}(G)$.

Given this, we first observe that for each map G the sequence of objectives $\mathcal{L}^{\alpha_n}(G)$ monotonically converges to the objective $\mathcal{L}^{\infty}(G)$. It follows from the fact that the first summand of $\mathcal{L}^{\alpha_n}$ converges to $+\infty$ if and only if the KL divergence is non-zero, which is equivalent to saying that p^G and $p^{\mathcal{T}}$ differ (Wang et al., 2024). If instead $p^G = p^{\mathcal{T}}$, the summand zeroes out. This also means that the minimal values of the corresponding objectives form a monotonic sequence:

$$\mathcal{L}^{\alpha_n}(G^{\alpha_n}) \leq \mathcal{L}^{\alpha_{n+1}}(G^{\alpha_{n+1}}) \leq \mathcal{L}^{\infty}(G^{\infty}). \tag{20}$$

Finally, the monotonicity implies that for a fixed m

$$\lim_{n\to\infty} \mathcal{L}^{\alpha_n}(G^{\alpha_n}) \geq \lim_{n\to\infty} \mathcal{L}^{\alpha_m}(G^{\alpha_n}), \tag{21}$$

[4] Solution to the Monge problem is not always unique, but we will further impose assumptions that will guarantee the uniqueness.

since the input G^{α_n} is fixed and $\mathcal{L}^{\alpha_n}$ monotonically increases. Using the assumed continuity of the objective, we obtain

$$\lim_{n\to\infty} \mathcal{L}^{\alpha_n}(G^{\alpha_n}) \geq \mathcal{L}^{\alpha_m}(\hat{G}) \tag{22}$$

for each m. Taking the limit $m \to \infty$, we obtain

$$\lim_{n\to\infty} \mathcal{L}^{\alpha_n}(G^{\alpha_n}) \geq \mathcal{L}^{\infty}(\hat{G}). \tag{23}$$

Combining this set of equations, we obtain:

$$\mathcal{L}^{\infty}(G^{\infty}) \geq \lim_{n\to\infty} \mathcal{L}^{\alpha_n}(G^{\alpha_n}) \geq \mathcal{L}^{\infty}(\hat{G}) \geq \mathcal{L}^{\infty}(G^{\infty}), \tag{24}$$

where the first inequality comes from the monotonicity of the minimal values and the last inequality uses that G^{∞} is the minimum of the objective $\mathcal{L}^{\infty}$. Hence, that limiting map $\hat{G}$ achieves minimal value of the objective $\mathcal{L}^{\infty}$ and is, therefore, the optimal transport map.

At this point, we only need to define and prove some versions of the aforementioned facts:

1. Infimum of $\mathcal{L}^{\alpha}$ is reached;
2. The sequence of minima G^{α_n} converges;
3. $\mathcal{L}^{\alpha}$ is continuous with respect to this convergence.

From now on, we formulate the result in details and stick to the formal proof.

A.2. Assumptions and theorem statement

First, we list the assumptions.

Assumption A.1. The distributions $p^{\mathcal{S}}$ and $p^{\mathcal{T}}$ have densities with respect to the Lebesgue measure. The distributions are defined on open bounded subsets $\mathcal{X} \subset \mathbb{R}^d$ and $\mathcal{Y} \subset \mathbb{R}^d$, where $\mathcal{Y}$ is convex. The densities are bounded away from zero and infinity on $\mathcal{X}$ and $\mathcal{Y}$, respectively.

We admit that boundedness of the support is a very restrictive assumption from the theoretical standpoint, however in our applications (I2I) both source and target distributions are supported on the bounded space of images. We thus can set $\mathcal{X} = \mathcal{Y} = (0, 1)^d$.

Assumption A.2. The cost $c(\boldsymbol{x}, \boldsymbol{y})$ is quadratic $\|\boldsymbol{x} - \boldsymbol{y}\|^2$.

Here, we stick to proving the theorem only for L_2 cost due to difficulties in investigation of Monge map existence and regularity for general transport costs (De Philippis & Figalli, 2014).

Assumption A.3. The weighting function ω_t is positive and bounded.

Assumption A.4. Standard deviation σ_t of the noise, defined by the forward process, is continuous in t.

Theorem A.1. *Let $p^{\mathcal{S}}$, $p^{\mathcal{T}}$, c, ω_t, and σ_t satisfy the assumptions **1-3**. Then, there exists a minimum G^{α} of the objective $\mathcal{L}^{\alpha}$ from the Equation 18. If $\alpha_n \to \infty$, the sequence G^{α_n} converges in probability (with respect to the source distribution) to the optimal transport map G^{∞}:*

$$G^{\alpha_n} \xrightarrow[n\to\infty]{p^{\mathcal{S}}} G^{\infty}. \tag{25}$$

A.3. Theoretical background

We start by listing all the results necessary for the proof. They are mostly related to the topics of measure theory (weak convergence, in particular) and optimal transport. Most of these classic facts can be found in the books (Bogachev & Ruas, 2007; Dudley, 2018). Otherwise, we make the corresponding citations.

Definition A.2. A sequence of probability distributions $p^n(\boldsymbol{x})$ converges weakly to the distribution $p(\boldsymbol{x})$ if for all continuous bounded test functions $\varphi \in \mathcal{C}_b(\mathbb{R}^d)$ holds

$$\mathbb{E}_{p^n(\boldsymbol{x})}\varphi(\boldsymbol{x}) \xrightarrow[n\to\infty]{} \mathbb{E}_{p(\boldsymbol{x})}\varphi(\boldsymbol{x}). \tag{26}$$

Notation: $p^n \xrightarrow{w} p$.

Definition A.3. A function $f : \mathbb{R}^d \to \mathbb{R}$ is called lower semi-continuous (lsc), if for all $\boldsymbol{x}_n \to \boldsymbol{x}$ holds

$$\liminf_{n\to\infty} f(\boldsymbol{x}_n) \geq f(\boldsymbol{x}). \tag{27}$$

Theorem A.4 (Portmanteau/Alexandrov)**.** *$p^n \xrightarrow{w} p$ is equivalent to the following statement: for every lsc function f, bounded from below, holds*

$$\liminf_{n\to\infty} \mathbb{E}_{p^n(\boldsymbol{x})} f(\boldsymbol{x}) \geq \mathbb{E}_{p(\boldsymbol{x})} f(\boldsymbol{x}). \tag{28}$$

Definition A.5. A sequence of probability measures p^n is called relatively compact, if for every subsequence p^{n_k} there exists a weakly convergent subsequece $p^{n_{k_j}}$.

Definition A.6. A sequence of probability measures p^n is called tight, if for every $\varepsilon > 0$ there exists a compact set K_ε such that $p^n(K_\varepsilon) \geq 1 - \varepsilon$ for all n.

Theorem A.7 (Prokhorov)**.** *A sequence of probability measures p^n is relatively compact if and only if it is tight. In particular, every weakly convergent sequence is tight.*

Corollary A.8. *If there exists a function $\varphi(\boldsymbol{x})$ such that its sublevels $\{\boldsymbol{x} : \varphi(x) \leq r\}$ are compact and for all n*

$$\mathbb{E}_{p^n(\boldsymbol{x})} \varphi(x) \leq C$$

holds with some constant C, then p^n is tight (i.e. at least it has a weakly convergent subsequence).

Corollary A.9. *If a sequence p^n is tight and all of its weakly convergent subsequences converge to the same measure p, then $p^n \xrightarrow{w} p$.*

Definition A.10. The functional $\mathcal{L}(p)$ is called lower semi-continuous (lsc) with respect to the weak convergence if for all weakly convergent sequences $p^n \xrightarrow{w} p$ holds

$$\liminf_{n\to\infty} \mathcal{L}(p^n) \geq \mathcal{L}(p). \tag{29}$$

Theorem A.11 (Posner 1975)**.** *The* KL *divergence* $\mathrm{KL}(p \,\|\, q)$ *is lsc (in sense of weak convergence) with respect to each argument, i.e. if $p^n \xrightarrow{w} p$ and $q_n \xrightarrow{w} q$, then*

$$\liminf_{n\to\infty} \mathrm{KL}(p^n \,\|\, q) \geq \mathrm{KL}(p \,\|\, q) \tag{30}$$

$$\liminf_{n\to\infty} \mathrm{KL}(p \,\|\, q_n) \geq \mathrm{KL}(p \,\|\, q). \tag{31}$$

Theorem A.12 (Donsker & Varadhan 1983)**.** *The* KL *divergence can be expressed as*

$$\mathrm{KL}(p\|q) = \sup_g \left(\mathbb{E}_{p(\boldsymbol{x})} g(\boldsymbol{x}) - \log \mathbb{E}_{q(\boldsymbol{x})} e^{g(\boldsymbol{x})} \right). \tag{32}$$

Definition A.13. The expression

$$\mathbb{E}_{p(\boldsymbol{x})} e^{i\langle s, \boldsymbol{x}\rangle} \tag{33}$$

is called the characteristic function (Fourier transform) of the distribution $p(\boldsymbol{x})$.

Theorem A.14 (Lévy)**.** *Weak convergence of probability measures $p^n \xrightarrow{w} p$ is equivalent to the point-wise convergence of characteristic functions, i.e. $\mathbb{E}_{p^n(\boldsymbol{x})} e^{i\langle s, \boldsymbol{x}\rangle} \to \mathbb{E}_{p(\boldsymbol{x})} e^{i\langle s, \boldsymbol{x}\rangle}$ for all s.*

Definition A.15. A sequence of measurable functions $\varphi^n(\boldsymbol{x})$ is said to converge in measure (in probability) to the function φ with respect to the measure $p(\boldsymbol{x})$, if for all $\varepsilon > 0$ holds

$$p\left(\{\boldsymbol{x} : |\varphi^n(\boldsymbol{x}) - \varphi(\boldsymbol{x})| > \varepsilon\}\right) \to 0.$$

Theorem A.16 (Lebesgue)**.** *Let φ^n, φ be measurable functions such that $\|\varphi^n(\boldsymbol{x})\|, \|\varphi(\boldsymbol{x})\| \leq C$ and $\varphi^n(\boldsymbol{x}) \to \varphi(\boldsymbol{x})$ pointwise. Then $\mathbb{E}_{p(\boldsymbol{x})} \varphi^n(\boldsymbol{x}) \to \mathbb{E}_{p(\boldsymbol{x})} \varphi(\boldsymbol{x})$.*

Lemma A.17 (Fatou)**.** *For any sequence of measurable functions φ^n the function $\liminf_n \varphi^n$ is measurable and*

$$\int_a^b \liminf_{n\to\infty} \varphi^n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq \liminf_{n\to\infty} \int_a^b \varphi^n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{34}$$

Theorem A.18 (Brenier 1991)**.** *Given the Assumption A.1, there exists a unique optimal transport map that solves the Monge problem 12 for the quadratic cost.*

Proof. This result can be found e.g. in (De Philippis & Figalli, 2014, Theorem 3.1). □

Theorem A.19. *Given the Assumption A.1, the unique OT Monge map is continuous.*

Proof. This is a simplified version of (De Philippis & Figalli, 2014, Theorem 3.3). □

A.4. Lower semi-continuity of the loss

Having defined all the needed terms and results, we start the proof by re-defining the objective in Equation 18 with respect to the joint distribution π input and output of the generator instead of the generator G itself. Analogous to the Kantorovitch formulation of the optimal transport problem (Kantorovitch, 1958), for each measure π on $\mathbb{R}^d \times \mathbb{R}^d$ (which is also called a *transport plan* or just plan) we define the corresponding fuctional as

$$\mathcal{L}^{\alpha}(\pi) = \alpha \int_0^T \omega_t \, \mathrm{KL}\left(\pi_{\boldsymbol{y},t} \,\|\, p_t^{\mathcal{T}}\right) \mathrm{d}t + \mathbb{E}_{\pi(\boldsymbol{x},\boldsymbol{y})} c\left(\boldsymbol{x},\boldsymbol{y}\right), \tag{35}$$

where $\pi_{\boldsymbol{x}}$ and $\pi_{\boldsymbol{y}}$ are the corresponding projections (marginal distributions) of π and $\pi_{\boldsymbol{y},t}$ is the perturbed $\boldsymbol{y}$-marginal distribution of π. Note that for π, corresponding to the joint distribution of $(\boldsymbol{x}, G(\boldsymbol{x}))$, $\mathcal{L}^{\alpha}(\pi)$ coincides with $\mathcal{L}^{\alpha}(G)$, defined in Equation 18. Thus, we aim to optimize $\mathcal{L}^{\alpha}(\pi)$ with respect to such plans π, that their $\boldsymbol{x}$ marginal is equal to $p^{\mathcal{S}}$ and $\pi(\boldsymbol{y} = G(\boldsymbol{x})) = 1$ for some G.

Definition A.20. We will call a measure π generator-based if its $\boldsymbol{x}$-marginal is equal to $p^{\mathcal{S}}$ and $\pi(\boldsymbol{y} = G(\boldsymbol{x}))$ for some function G.

For the sake of clearity, we note that the distributions $\pi_t^{\boldsymbol{y}}$ and $p_t^{\mathcal{T}}$ can be represented as $\pi^{\boldsymbol{y}} * q_t$ and $p^{\mathcal{T}} * q_t$, where $*$ is the convolution operation and $q_t = \mathcal{N}(0, \sigma_t^2 I)$. We thus rewrite the functional as

$$\mathcal{L}^{\alpha}(\pi) = \alpha \int_0^T \omega_t \, \mathrm{KL}\left(\pi_{\boldsymbol{y}} * q_t \,\|\, p^{\mathcal{T}} * q_t\right) \mathrm{d}t + \mathbb{E}_{\pi(\boldsymbol{x},\boldsymbol{y})} c\left(\boldsymbol{x},\boldsymbol{y}\right), \tag{36}$$

Previously, we wanted to establish continuity of the objective. This may not be the case in general. Instead, we prove the following

Lemma A.21. *$\mathcal{L}^{\alpha}(\pi)$ is lsc with respect to the weak convergence, i.e. for all weakly convergent sequences $\pi^n \xrightarrow{w} \pi$ holds*

$$\liminf_{n\to\infty} \mathcal{L}^{\alpha}(\pi^n) \geq \mathcal{L}^{\alpha}(\pi). \tag{37}$$

This result is a direct consequence of the Theorem A.11 about lower semi-continuity of the KL divergence.

Proof. We start by proving that the projection and the convolution operation preserve weak convergence. For the first, we need to prove that for any test function $g \in \mathcal{C}_b(\mathbb{R}^d)$ holds

$$\mathbb{E}_{\pi_{\boldsymbol{y}}^n(\boldsymbol{y})} g(\boldsymbol{y}) \to \mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})} g(\boldsymbol{y}) \tag{38}$$

given $\pi^n \xrightarrow{w} \pi$. For this, we note that the function $\varphi(\boldsymbol{x},\boldsymbol{y}) = g(\boldsymbol{y})$ is also bounded and continuous and, thus

$$\mathbb{E}_{\pi_{\boldsymbol{y}}^n(\boldsymbol{y})} g(\boldsymbol{y}) = \mathbb{E}_{\pi^n(\boldsymbol{x},\boldsymbol{y})} \varphi(\boldsymbol{x},\boldsymbol{y}) \to \mathbb{E}_{\pi(\boldsymbol{x},\boldsymbol{y})} \varphi(\boldsymbol{x},\boldsymbol{y}) = \mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})} g(\boldsymbol{y}). \tag{39}$$

Regarding the convolution, recall that $\pi_{\boldsymbol{y}}^n * q_t$ is the distribution of the sum of independent variables with corresponding distributions. Its characteristic function is equal to

$$\mathbb{E}_{\pi_{\boldsymbol{y}}^n * q_t(\boldsymbol{y}_t)} e^{i\langle s, \boldsymbol{y}_t\rangle} = \mathbb{E}_{\pi_{\boldsymbol{y}}^n(\boldsymbol{y}) q_t(\varepsilon_t)} e^{i\langle s, \boldsymbol{y}+\varepsilon_t\rangle} = \mathbb{E}_{\pi_{\boldsymbol{y}}^n(\boldsymbol{y})} e^{i\langle s, \boldsymbol{y}\rangle} \mathbb{E}_{q_t(\varepsilon_t)} e^{i\langle s, \varepsilon_t\rangle}. \tag{40}$$

Applying the Lévy's continuity theorem to $\pi^n_{\boldsymbol{y}} \xrightarrow{w} \pi_{\boldsymbol{y}}$, we take the limit and obtain

$$\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})} e^{i\langle s, \boldsymbol{y}\rangle} \mathbb{E}_{q_t(\varepsilon_t)} e^{i\langle s, \varepsilon_t\rangle} = \mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y}) q_t(\varepsilon_t)} e^{i\langle s, \boldsymbol{y}+\varepsilon_t\rangle} = \mathbb{E}_{\pi_{\boldsymbol{y}} * q_t(\boldsymbol{y}_t)} e^{i\langle s, \boldsymbol{y}_t\rangle}, \tag{41}$$

which implies

$$\mathbb{E}_{\pi^n_{\boldsymbol{y}} * q_t(\boldsymbol{y}_t)} e^{i\langle s, \boldsymbol{y}_t\rangle} \to \mathbb{E}_{\pi_{\boldsymbol{y}} * q_t(\boldsymbol{y}_t)} e^{i\langle s, \boldsymbol{y}_t\rangle}. \tag{42}$$

We apply the continuity theorem for the convolutions and obtain $\pi^n_{\boldsymbol{y}} * q_t \xrightarrow{w} \pi_{\boldsymbol{y}} * q_t$.

With this observation, we prove that the first term of $\mathcal{L}^\alpha(\pi)$ is lsc. First, we apply Lemma A.17 (Fatou) and move the limit inside the integral

$$\liminf_{n\to\infty} \int_0^T \omega_t \, \mathrm{KL}\left(\pi^n_{\boldsymbol{y}} * q_t \,\|\, p^{\mathcal{T}} * q_t\right) \mathrm{d}t \geq \int_0^T \liminf_{n\to\infty} \omega_t \, \mathrm{KL}\left(\pi^n_{\boldsymbol{y}} * q_t \,\|\, p^{\mathcal{T}} * q_t\right) \mathrm{d}t. \tag{43}$$

Using the lower semi-continuity of the KL divergence (Theorem A.11), we obtain

$$\int_0^T \liminf_{n\to\infty} \omega_t \, \mathrm{KL}\left(\pi^n_{\boldsymbol{y}} * q_t \,\|\, p^{\mathcal{T}} * q_t\right) \mathrm{d}t \geq \int_0^T \omega_t \, \mathrm{KL}\left(\pi_{\boldsymbol{y}} * q_t \,\|\, p^{\mathcal{T}} * q_t\right) \mathrm{d}t. \tag{44}$$

Finally, the Assumption A.2 on the continuity of $c(\cdot,\cdot)$ implies its lower semi-coninuity. Theorem A.4 (Portmanteau) states that

$$\liminf_{n\to\infty} \mathbb{E}_{\pi^n(\boldsymbol{x},\boldsymbol{y})} c(\boldsymbol{x},\boldsymbol{y}) \geq \mathbb{E}_{\pi(\boldsymbol{x},\boldsymbol{y})} c(\boldsymbol{x},\boldsymbol{y}). \tag{45}$$

Combining inequalities from Equation 43, Equation 44 and Equation 45, we obtain

$$\liminf_{n\to\infty} \mathcal{L}^\alpha(\pi^n) \geq \mathcal{L}^\alpha(\pi). \tag{46}$$

□

A.5. Existence of the minimizer

Now we aim to prove that the objective $\mathcal{L}^\alpha(\pi)$ has a minimum over generator-based plans. First, we need the following technical lemma about sublevels of the KL part of the functional.

Lemma A.22. *Let $\{\pi^n\}_{n=1}^\infty$ be a sequence of generator-based plans that satisfy*

$$\int_0^T \omega_t \, \mathrm{KL}\left(\pi^n_{\boldsymbol{y},t} \,\|\, p^{\mathcal{T}}_t\right) \mathrm{d}t \leq C \tag{47}$$

for some constant C. Then, the sequence $\{\pi^n\}_{n=1}^\infty$ is tight.

Proof. We take arbitrary π from the sequence and apply the Donsker-Varadhan representation (Theorem A.12) of the KL divergence. We take the test function $g(\boldsymbol{x}) = \|x\|^2/(2\sigma_T^2)$ and obtain

$$\int_0^T \omega_t \, \mathrm{KL}\left(\pi_{\boldsymbol{y},t} \,\|\, p^{\mathcal{T}}_t\right) \mathrm{d}t \geq \int_0^T \omega_t \left(\mathbb{E}_{\pi_{\boldsymbol{y},t}(\boldsymbol{y}_t)} \frac{1}{2\sigma_T^2} \|\boldsymbol{y}_t\|^2 - \log \mathbb{E}_{p^{\mathcal{T}}_t(\boldsymbol{y}_t)} e^{\|\boldsymbol{y}_t\|^2/(2\sigma_T^2)}\right) \mathrm{d}t. \tag{48}$$

The choice of $g(\boldsymbol{x})$ is not very specific, i.e. every function that will produce finite expectations and integrals is suitable. In the right-hand side, we rewrite the expectations with repect to the original variable and noise:

$$\int_0^T \omega_t \left(\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})\mathcal{N}(\varepsilon|0,I)} \frac{1}{2\sigma_T^2} \|\boldsymbol{y} + \sigma_t \varepsilon\|^2 - \log \mathbb{E}_{p^{\mathcal{T}}(\boldsymbol{y})\mathcal{N}(\varepsilon|0,I)} e^{\|\boldsymbol{y}+\sigma_t\varepsilon\|^2/(2\sigma_T^2)}\right) \mathrm{d}t. \tag{49}$$

We rewrite $\|\boldsymbol{y}+\sigma_t\varepsilon\|^2$ as $\|\boldsymbol{y}\|^2+2\sigma_t\langle\boldsymbol{y},\sigma_t\varepsilon\rangle+\sigma_t^2\|\varepsilon\|^2$ and note that expectation of the second term is zero. The first term is then equal to

$$\frac{1}{2\sigma_T^2}\int\limits_0^T \omega_t\,\mathrm{d}t\cdot\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})}\|\boldsymbol{y}\|^2+\frac{1}{2\sigma_T^2}\int\limits_0^T \omega_t\,\sigma_t^2\mathrm{d}t\cdot\mathbb{E}_{\mathcal{N}(\varepsilon|0,I)}\|\varepsilon\|^2. \tag{50}$$

Boundedness of ω_t (Assumption A.3) implies that the first integral is finite and, say, equal to C_1. The second integral contains a product of bounded ω_t and continuous σ_t^2 (Assumtion A.4), which is also integrable. We then denote the second summand by C_2 and rewrite the first summand as

$$C_1\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})}\|\boldsymbol{y}\|^2+C_2. \tag{51}$$

As for the second summand, we see that the expectation

$$\mathbb{E}_{p^{\mathcal{T}}(\boldsymbol{y})\mathcal{N}(\varepsilon|0,I)}e^{\|\boldsymbol{y}+\sigma_t\varepsilon\|^2/(2\sigma_T^2)} \tag{52}$$

with respect to ε will be finite, because $\sigma_t^2/(2\sigma_T^2)$ is always less than $1/2$, which will make the exponent have negative degree. Moreover, simple calculations show that this function will be continuous with respect to σ_t and have only quadratic terms with respect to $\boldsymbol{y}$ inside the exponent, i.e. have the form

$$e^{a(\sigma_t)\|\boldsymbol{y}-b(\sigma_t)\|^2+c(\sigma_t)} \tag{53}$$

with continuous a,b,c. We now want to prove that the expectation

$$\mathbb{E}_{p^{\mathcal{T}}(\boldsymbol{y})}e^{\alpha(\sigma_t)\|\boldsymbol{y}-\beta(\sigma_t)\|^2+\gamma(\sigma_t)} \tag{54}$$

will also be continuous in t. First, due to the boundedness of $\boldsymbol{y}$, this expectation is finite. Second, for $t_n\to t$:

$$\lim_{n\to\infty}\mathbb{E}_{p^{\mathcal{T}}(\boldsymbol{y})}e^{a(\sigma_{t_n})\|\boldsymbol{y}-b(\sigma_{t_n})\|^2+c(\sigma_{t_n})}= \tag{55}$$

$$=\mathbb{E}_{p^{\mathcal{T}}(\boldsymbol{y})}\lim_{n\to\infty}e^{a(\sigma_{t_n})\|\boldsymbol{y}-b(\sigma_{t_n})\|^2+c(\sigma_{t_n})}= \tag{56}$$

$$=\mathbb{E}_{p^{\mathcal{T}}(\boldsymbol{y})}e^{a(\sigma_t)\|\boldsymbol{y}-b(\sigma_t)\|^2+c(\sigma_t)} \tag{57}$$

due to the Theorem A.16 (Lebesgue's dominated convergence). It is applicable, since $\boldsymbol{y}$ is bounded and all the functions are continuous, thus bounded in $[0,T]$.

We thus obtain that the second integral contains bounded ω_t multiplied by the logarithm of continuous function, which is always ≥ 1 (positive exponent). This means that the whole integral is finite. Denoting it by C_3, we obtain

$$C_1\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})}\|\boldsymbol{y}\|^2+C_2-C_3\leq\int\limits_0^T\omega_t\,\mathrm{KL}\left(\pi_{\boldsymbol{y},t}\,\|\,p_t^{\mathcal{T}}\right)\mathrm{d}t. \tag{58}$$

Combined with the condition of the lemma, we obtain

$$C_1\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})}\|\boldsymbol{y}\|^2+C_2-C_3\leq\int\limits_0^T\omega_t\,\mathrm{KL}\left(\pi_{\boldsymbol{y},t}\,\|\,p_t^{\mathcal{T}}\right)\mathrm{d}t\leq C, \tag{59}$$

which implies

$$\mathbb{E}_{\pi_{\boldsymbol{y}}(\boldsymbol{y})}\|\boldsymbol{y}\|^2\leq\frac{C+C_3-C_2}{C_1}:=C_4. \tag{60}$$

We thus obtained a uniform bound on some statistic with respect to all measures from $\{\pi^n\}$. The function $\|\boldsymbol{y}\|^2$ has compact sublevel sets $\{\|\boldsymbol{y}\|^2\leq r\}$. Lemma A.8 then states that the sequence $\pi^n_{\boldsymbol{y}}$ is tight, i.e. for all $\varepsilon>0$ there is a compact set K_ε with $\pi^n_{\boldsymbol{y}}(\boldsymbol{y}\in K_\varepsilon)\geq 1-\varepsilon$.

Finally, marginal $\boldsymbol{x}$ distribution of each of the π^n is $p^{\mathcal{S}}$, which is bounded (Assumption A.1), i.e. there is a compact K that $\pi^n(\boldsymbol{x}\in K)=1$. Combined with the previous observation, we obtain

$$\pi^n(\boldsymbol{x}\in K,\boldsymbol{y}\in K_\varepsilon)\geq 1-\varepsilon \tag{61}$$

for all n. The cartesian product $K\times K_\varepsilon$ is also compact. Theorem A.7 (Prokhorov) then implies that the sequence π^n is tight. □

Now we are ready to prove the following

Lemma A.23. *Infimum of the loss $\mathcal{L}^\alpha(\pi)$ over all generator-based transport plans π (with $\pi_{\boldsymbol{x}} = p^{\mathcal{S}}$ and $\pi(\boldsymbol{y} = G(\boldsymbol{x}))$ for some G) is attained on some plan $\hat{\pi}$.*

Proof. We start by observing that there is at least one feasible π with the aforementioned properties. For this purpose one can take the optimal transport map G^∞ between $p^{\mathcal{S}}$ and $p^{\mathcal{T}}$, which is unique by Theorem A.18 under Assumptions A.1, A.2.

Let π^n be a sequence of feasible generator-based measures that $\mathcal{L}^\alpha(\pi^n)$ converges to the corresponding infimum $\mathcal{L}^\alpha_{\text{inf}}$ (it exists by the definition of the infimum). Without loss of generality, we can assume that $\mathcal{L}^\alpha(\pi^n) \leq \mathcal{L}^\alpha_{\text{inf}} + 1$ for all n (if not, one can drop large enough sequence prefix). This implies that for all n holds

$$\alpha \int_0^T \omega_t \, \text{KL}\left(\pi_{\boldsymbol{y},t} \,\|\, p_t^{\mathcal{T}}\right) \text{d}t \leq \mathcal{L}^\alpha_{\text{inf}} + 1. \tag{62}$$

Lemma A.22 implies that the sequence π^n is tight. Prokhorov theorem then states that π^n has a weakly convergent subsequence $\pi^{n_k} \xrightarrow{w} \hat{\pi}$. Lower semi-continuity of the loss $\mathcal{L}^\alpha$ implies that

$$\liminf_{k\to\infty} \mathcal{L}^\alpha(\pi^{n_k}) \geq \mathcal{L}^\alpha(\hat{\pi}) \geq \mathcal{L}^\alpha_{\text{inf}}. \tag{63}$$

At the same time, $\mathcal{L}^\alpha(\pi^{n_k})$ is assumed to converge to $\mathcal{L}^\alpha_{\text{inf}}$, which means that $\hat{\pi}$ is indeed the minimum. □

A.6. Finish of the proof

Theorem A.1 proof. Finally, we combine the previous technical observations with the proof sketch from the Section A.1. Let $\alpha_n \to \infty$ be a sequence of coefficients, G^{α_n} be the optimal generators with respect to $\mathcal{L}^{\alpha_n}$ and π^{α_n} the joint distributions of $(\boldsymbol{x}, G^{\alpha_n}(\boldsymbol{x}))$. Additionally, we define π^∞ to be the optimal transport plan, corresponding to $(\boldsymbol{x}, G^\infty(\boldsymbol{x}))$, where $G^\infty(\boldsymbol{x})$ is the optimal transport map. First, due to the monotonicity of $\mathcal{L}^\alpha$ with respect to α, we have

$$\mathcal{L}^{\alpha_n}(\pi^{\alpha_n}) \leq \mathcal{L}^{\alpha_{n+1}}(\pi^{\alpha_{n+1}}) \leq \mathcal{L}^\infty(\pi^\infty). \tag{64}$$

This implies that for all n holds

$$\alpha_n \int_0^T \omega_t \, \text{KL}\left(\pi^{\alpha_n}_{\boldsymbol{y},t} \,\|\, p_t^{\mathcal{T}}\right) \text{d}t \leq \mathcal{L}^\infty(\pi^\infty) \Rightarrow \tag{65}$$

$$\Rightarrow \int_0^T \omega_t \, \text{KL}\left(\pi^{\alpha_n}_{\boldsymbol{y},t} \,\|\, p_t^{\mathcal{T}}\right) \text{d}t \leq \frac{\mathcal{L}^\infty(\pi^\infty)}{\alpha_n} \leq \frac{\mathcal{L}^\infty(\pi^\infty)}{\min\limits_n \alpha_n}, \tag{66}$$

which is finite, since $\alpha_n \to +\infty$. One more time, we apply Lemma A.22 and conclude that the sequence π^{α_n} is tight.

Let $\pi^{\alpha_{n_k}}$ be its weakly convergent subsequence: $\pi^{\alpha_{n_k}} \xrightarrow{w} \hat{\pi}$. Analogously to the Section A.1, we observe that

$$\liminf_{k\to\infty} \mathcal{L}^{\alpha_{n_k}}(\pi^{\alpha_{n_k}}) \geq \liminf_{k\to\infty} \mathcal{L}^{\alpha_{n_m}}(\pi^{\alpha_{n_k}}) \geq \mathcal{L}^{\alpha_{n_m}}(\hat{\pi}) \tag{67}$$

for any fixed m. The first inequality is due to the monotonicity of $\mathcal{L}^\alpha$ with respect to α and second is the implication of lower semi-continuity of the loss $\mathcal{L}^\alpha$ with respect to weak convergence. Taking the limit $m \to \infty$, we obtain

$$\liminf_{k\to\infty} \mathcal{L}^{\alpha_{n_k}}(\pi^{\alpha_{n_k}}) \geq \mathcal{L}^\infty(\hat{\pi}). \tag{68}$$

Combining all these observations, we obtain the following sequence of inequalities

$$\mathcal{L}^\infty(\pi^\infty) \geq \liminf_{k\to\infty} \mathcal{L}^{\alpha_{n_k}}(\pi^{\alpha_{n_k}}) \geq \mathcal{L}^\infty(\hat{\pi}) \geq \mathcal{L}^\infty(\pi^\infty), \tag{69}$$

which implies that the limiting measure $\hat{\pi}$ reaches the minimum of the objective over generator-based plans. By the uniqueness of the optimal transport map G^∞ under the Assumptions A.1, A.2, A.3, we conclude that all the convergent subsequences $\pi^{\alpha_{n_k}}$ converge to the optimal measure π^∞. Using Corollary A.9 of the Prokhorov theorem, we deduce that $\pi^{\alpha_n} \xrightarrow{w} \pi^\infty$.

Finally, we want to replace the weak convergence of π^{α_n} to π^∞ with the convergence in probability of the generators, i.e. show

$$G^{\alpha_n} \xrightarrow{p^{\mathcal{S}}} G^\infty. \tag{70}$$

To this end, we represent the corresponding probability as the expectation of the indicator and upper bound it with a continuous function:

$$p^{\mathcal{S}}\left(\|G^{\alpha_n}(\boldsymbol{x}) - G^\infty(\boldsymbol{x})\| > \varepsilon\right) = \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} I\{\|G^{\alpha_n}(\boldsymbol{x}) - G^\infty(\boldsymbol{x})\| > \varepsilon\} \tag{71}$$

$$\leq \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} d\left(G^{\alpha_n}(\boldsymbol{x}), G^\infty(\boldsymbol{x})\right), \tag{72}$$

where d is a continuous indicator approximation, defined as

$$d(\boldsymbol{u}, \boldsymbol{v}) = \begin{cases} \frac{\|\boldsymbol{u}-\boldsymbol{v}\|}{\varepsilon}, & \text{if } 0 \leq \|\boldsymbol{u}-\boldsymbol{v}\| < \varepsilon; \\ 1, & \text{if } \|\boldsymbol{u}-\boldsymbol{v}\| \geq \varepsilon. \end{cases} \tag{73}$$

We define the test function

$$\varphi(\boldsymbol{x}, \boldsymbol{y}) = d\left(\boldsymbol{y}, G^\infty(\boldsymbol{x})\right) \tag{74}$$

and rewrite the upper bound as

$$\mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} d\left(G^{\alpha_n}(\boldsymbol{x}), G^\infty(\boldsymbol{x})\right) = \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} \varphi(\boldsymbol{x}, G^{\alpha_n}(\boldsymbol{x})) = \mathbb{E}_{\pi^{\alpha_n}(\boldsymbol{x},\boldsymbol{y})} \varphi(\boldsymbol{x}, \boldsymbol{y}). \tag{75}$$

Due to Assumptions A.1, A.2 and Theorem A.14 the optimal transport map G^∞ is continuous, which implies that this test function is bounded and continuous. Given the weak convergence of π^{α_n}, we have

$$\mathbb{E}_{\pi^{\alpha_n}(\boldsymbol{x},\boldsymbol{y})} \varphi(\boldsymbol{x}, \boldsymbol{y}) \to \mathbb{E}_{\pi^{\infty}(\boldsymbol{x},\boldsymbol{y})} \varphi(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} \varphi(\boldsymbol{x}, G^\infty(\boldsymbol{x})) = \tag{76}$$

$$= \mathbb{E}_{p^{\mathcal{S}}(\boldsymbol{x})} d(G^\infty(\boldsymbol{x}), G^\infty(\boldsymbol{x})) = 0, \tag{77}$$

which implies the desired

$$p^{\mathcal{S}}\left(\|G^{\alpha_n}(\boldsymbol{x}) - G^\infty(\boldsymbol{x})\| > \varepsilon\right) \to 0. \tag{78}$$

□

B. Ablation of the initialization parameter

In this section, we further explore the design space of our method by investigating the effect of the fixed generator input noise parameter σ on the resulting quality. To this end, we take the colored version of the MNIST (LeCun, 1998) data set and perform translation between the digits "2" and "3" initializing from various σ. We use a small UNet architecture from (Gushchin et al., 2024).

The parameter σ is residual from the pre-trained diffusion architecture and therefore fixed throughout training and evaluation. However, the target denoiser network tries to convert the expected noisy input into the corresponding sample from the output distribution. Consequently, one may expect that at a suitable noise level, the generator may change the input's details to make them look appropriate for the target while preserving the original structural properties.

We demonstrate this effect on various noise levels in Figure 6. Here we observe that the small sigmas lead to the mapping close to the identity, whereas the large sigmas lead to almost constant blurry images, corresponding to the average "3" of the data set. However, there is a segment $[1.0, 10.0]$ of levels that gives a moderate-quality mapping in terms of both faithfulness and realism, which makes it a suitable initial point. Note that the FID-L2 plot is not monotone at high L2 values due to the overall poor quality of the generator, i.e. it outputs bad-quality pictures slightly related to the source. We further investigate optimal σ choice by going through a 2D grid of the hyperparameters (σ, λ) and aim to see if it is possible to choose the uniform best noise level. In Figure 6 we report the faithfulness-quality trade-off concerning various σ. We

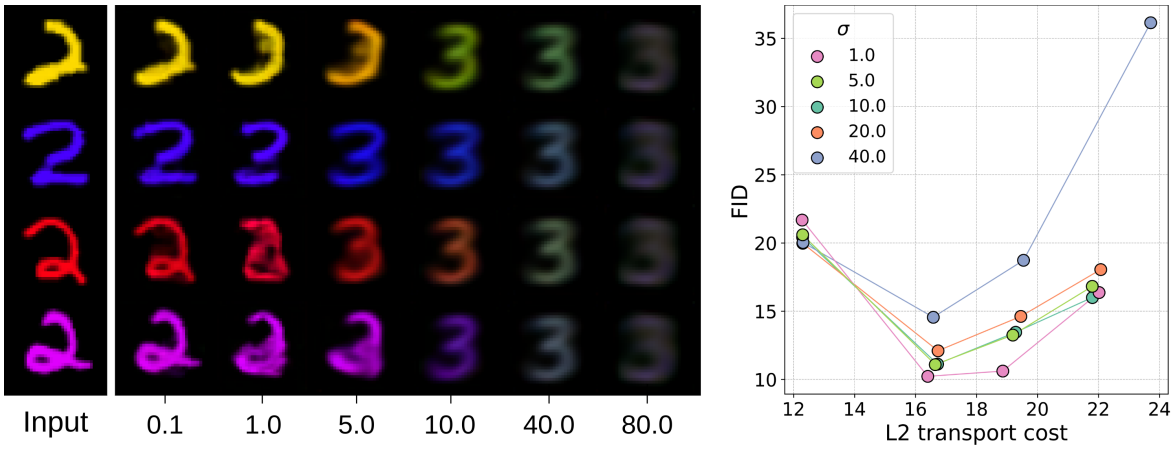

Figure 6. Left: visualization of the generator initialization at various $\sigma \in [0.1, 80.0]$, where σ is the noise level parameter residual from the pre-trained diffusion architecture. Right: comparison of different σ in terms of the quality-faithfulness trade-off. The metrics are obtained by initializing the generator at the corresponding σ level and training it with the RDMD procedure. Here, $\lambda \in \{0, 1.0, 2.0, 4.0\}$. Higher λ corresponds to the lower transport cost values.

observe that there is almost monotone dependence on σ on the segment $[1.0, 40.0]$: here the $\sigma = 1.0$ gives almost uniformly best results in terms of both metrics. Similar results are obtained by the values $5.0, 10.0$ which have fair quality visual results at initialization. Therefore, we conclude that it is best to choose the least parameter σ among the parameters with appropriate visuals at the initial point.

C. Experimental Details

C.1. 2D experiments

Architecture. The architecture used for training diffusion model and generator (De Bortoli et al., 2021) consists of input-encoding MLP block, time-encoding MLP block, and decoding MLP block. Input encoding MLP block consists of 4 hidden layers with dimensions $[16, 32, 32, 32]$ interspersed by LeakyReLU activations. Time encoding MLP consists of a positional encoding layer (Vaswani et al., 2017) and then follows the same MLP block structure as the input encoder. The decoding MLP block consists of 5 hidden layers with dimensions $[128, 256, 128, 64, 2]$ and operates on concatenated time embedding and input embedding each obtained from their respective encoder. The model contains $88k$ parameters.

Training Diffusion Model. Diffusion model was trained for 100k iterations with batch size 1024 with Adam optimizer (Kingma & Ba, 2014) with learning rate 10^{-4}.

Training RDMD. Fake denoising network was trained with Adam optimizer with learning rate 10^{-4}. The generator model was trained with a different Adam optimizer with a learning rate equal to $2 \cdot 10^{-5}$. We trained RDMD for 100k iterations with batch size 1024.

Computational resources. All the experiments were run on CPU. Running 100k iterations with the batch size 1024 took approximately 1 hour.

C.2. Colored MNIST

Architecture. We used the architecture from (Gushchin et al., 2024), which utilizes convolutional UNet with conditional instance normalization on time embeddings used after each upscaling block of the decoder[5]. Model produces time embeddings via positional encoding. The model size was approximately $9.9M$ parameters.

[5] https://github.com/ngushchin/EntropicNeuralOptimalTransport/blob/06efb6ba8b43865a30b0b626384fa64da39b
src/cunet.py

Training Diffusion Model. The diffusion model was trained for 24500 iterations with batch size 8192. We used the Adam optimizer with a learning rate equal to $4 \cdot 10^{-3}$. The model was trained in FP32. It obtained FID equal to 2.09.

Training RDMD. Fake denoising network was trained with Adam optimizer with a learning rate equal to $2 \cdot 10^{-3}$. The generator model was trained with Adam optimizer with learning rate $5 \cdot 10^{-5}$. RDMD was trained for 7300 iterations with batch size 4096.

Computational resources. All the experiments were run on 2x NVIDIA GeForce RTX 4090 GPUs. Training Diffusion model for 24500 iterations with batch size 8192 took approximately 6 hours. Training RDMD for 7300 iterations with batch size 4096 took approximately 3 hours.

C.3. Cat2Wild

Architecture. We used the SongUNet architecture from EDM repository[6], which corresponds to DDPM++ and NCSN++ networks from the work (Song et al., 2020). The model contains approximately $55M$ parameters.

Training Diffusion Model. The diffusion model was trained for 80k iterations. We set the batch size to 512 and chose the best checkpoint according to FID. We used the Adam optimizer with a learning rate equal to $2 \cdot 10^{-4}$. We used a dropout rate equal to 0.25 during training and the augmentation pipeline from (Karras et al., 2022) with probability 0.15. The model was trained in FP32. It obtained FID equal to 2.01.

Training RDMD. In all runs we initialized the generator from the target diffusion model with the fixed $\sigma = 1.0$. We've run 4 models, corresponding to the regularization coefficients $\{0.02, 0.05, 0.1, 0.2\}$. All models were trained with Adam optimizer with generator learning rate $5 \cdot 10^{-5}$ and fake diffusion learning rate $3 \cdot 10^{-4}$. We trained all models for 25000 iterations with batch size 512. Training took approximately 35 hours on $4\times$ NVidia Tesla A100 80GB.

ILVR hyperparameters. The only hyperparameter of ILVR is the downsampling factor N for the low-pass filter, which determines whether guidance would be conducted on coarser or finer information. n_{steps} denotes number of sampling steps. All metrics in Figure 4 for ILVR were obtained on the following hyperparameter grid: $N = [2, 4, 8, 16, 32]$, $n_{\text{steps}} = [18, 32, 50]$. We excluded runs that have the same statistical significance and achieve FID higher than 20.0. The images in Figure 5 were obtained with hyperparameters $N = 16$ and $n_{\text{steps}} = 18$.

SDEdit hyperparameters. The only hyperparameter of SDEdit is the noise level σ, which acts as a starting point for sampling. The higher the noise level, the closer is the sampling procedure to the unconditional generation. The smaller the noise values, the more features are carried over to the target domain at the expense of generation quality. n_{steps} denotes number of sampling steps. All metrics in Figure 4 for SDEdit were obtained on the following hyperparameter grid: $\sigma = [4, 5, 10, 15, 20, 30, 40]$, $n_{\text{steps}} = [18, 32, 50]$. We exclude runs that have the same statistical significance and achieve FID higher than 20.0. The images in Figure 5 were obtained with hyperparameters $\sigma = 10$ and $n_{\text{steps}} = 50$.

EGSDE hyperparameters. EGSDE sampling hyperparameters include the initial noise level σ at which the source image is perturbed, and the downsampling factor N for the low-pass filter. n_{steps} denotes number of sampling steps. All metrics in Figure 4 for EGSDE were obtained on the following hyperparameter grid: $\sigma = [5, 10, 15, 20, 40]$, $N = [8, 16, 32]$, $n_{\text{steps}} = [18, 32]$. We exclude runs that have the same statistical significance and achieve FID higher than 20.0. The images in Figure 5 were obtained with hyperparameters $\sigma = 10, N = 32, n_{\text{steps}} = 50$.

[6] https://github.com/NVlabs/edm/blob/008a4e5316c8e3bfe61a62f874bddba254295afb/training/networks.py