# Switched Flow Matching: Eliminating Singularities via Switching ODEs

**Qunxi Zhu** [1]   **Wei Lin** [1 2 3 4]

## Abstract

Continuous-time generative models, such as Flow Matching (FM), construct probability paths to transport between one distribution and another through the simulation-free learning of the neural ordinary differential equations (ODEs). During inference, however, the learned model often requires multiple neural network evaluations to accurately integrate the flow, resulting in a slow sampling speed. We attribute the reason to the inherent (joint) heterogeneity of source and/or target distributions, namely the singularity problem, which poses challenges for training the neural ODEs effectively. To address this issue, we propose a more general framework, termed Switched FM (SFM), that eliminates singularities via switching ODEs, as opposed to using a uniform ODE in FM. Importantly, we theoretically show that FM cannot transport between two simple distributions due to the existence and uniqueness of initial value problems of ODEs, while these limitations can be well tackled by SFM. From an orthogonal perspective, our framework can seamlessly integrate with the existing advanced techniques, such as minibatch optimal transport, to further enhance the straightness of the flow, yielding a more efficient sampling process with reduced costs. We demonstrate the effectiveness of the newly proposed SFM through several numerical examples.

[1]Research Institute of Intelligent Complex Systems, Fudan University, China. [2]School of Mathematical Sciences, LMNS, and SCMS, Fudan University, China. [3]State Key Laboratory of Medical Neurobiology and MOE Frontiers Center for Brain Science, Institutes of Brain Science, Fudan University, China. [4]Shanghai Artificial Intelligence Laboratory, China. Correspondence to: Qunxi Zhu <qxzhu16@fudan.edu.cn>.

## 1. Introduction

Generative modeling is a fundamental task in the machine learning and data science communities, whose primary objective is to transform samples from one (empirical) probability distribution to another through a learnable transformation. Over the years, several methods have been extensively proposed for generative modeling, including generative adversarial networks (GAN) (Goodfellow et al., 2014), variational autoencoders (VAE) (Kingma & Welling, 2013; Rezende et al., 2014), energy-based models (Teh et al., 2003; LeCun et al., 2006; Du & Mordatch, 2019; Song & Kingma, 2021), normalizing flow models (Dinh et al., 2014; 2016; Rezende & Mohamed, 2015), and autoregressive models (Germain et al., 2015; Van Den Oord et al., 2016; Van den Oord et al., 2016; Oord et al., 2016).

Despite their successes across various domains, these models have some limitations. For instance, training GANs can be challenging because of several major issues, including mode collapse (Goodfellow et al., 2014; Metz et al., 2016), vanishing gradient (Arjovsky et al., 2017; Weng, 2019), and unstable convergence (Arjovsky & Bottou, 2017; Farnia & Ozdaglar, 2020). VAE and energy-based models employ surrogate losses to aid in successful training via utilizing the evidence lower bound (with the parameterization trick) (Kingma & Welling, 2013) and contrastive divergence (Hinton, 2002), respectively. Normalizing flow (Dinh et al., 2014; 2016; Rezende & Mohamed, 2015) and autoregressive models (Germain et al., 2015; Van Den Oord et al., 2016; Van den Oord et al., 2016; Oord et al., 2016) often impose architectural constraints to build a normalized probability model.

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020; Song et al., 2020a), the current state-of-the-art generative models, have delivered outstanding results in a myriad of tasks (Chen et al., 2020; Nichol et al., 2021; Rombach et al., 2022; Saharia et al., 2022), primarily due to the scalable and stable training methodologies (Dhariwal & Nichol, 2021). In a significant leap forward, Song et al. (2020b) introduced a general framework that encapsulates the essence of previous diffusion models through the stochastic differential equations (SDEs), which, equivalently, correspond to the neural ordinary differential equations (ODEs) (Chen et al., 2018) in the sense of proba-

bility flow. Recently, Lipman et al. (2022), developed the Flow matching (FM), a scalable, simulation-free approach to train the probability flow, also known as continuous normalizing flow (Chen et al., 2018; Grathwohl et al., 2018), by directly regressing vector fields along specific conditional probability paths. We note that two concurrent studies, the stochastic interpolant by Albergo & Vanden-Eijnden (2022) and the rectified flow by Liu et al. (2022), propose similar methodologies for matching distributions using flows, albeit from distinct viewpoints.

However, during inference, generating a high-quality sample via simulating the learned ODEs often requires multiple function evaluations, leading to a long inference time. This inefficiency arises from the utilization of independent couplings that overlook the intrinsic structures connecting source and target distributions (Lipman et al., 2022; Liu et al., 2022). To mitigate this issue, there has been a shift towards designing non-trivial couplings inspired by optimal transport theory (Pooladian et al., 2023; Tong et al., 2023a;b) or learning a coupling based on an auxiliary VAE-style objective function to minimize the trajectory curvature (Lee et al., 2023). Notably, these existing continuous-time generative models, have predominantly adopted a uniform/single ODE to model the transportation process between two (empirical) distributions.

**Contributions**. We introduce Switched FM (SFM), a generalized framework that eliminates singularities via switching ODEs as opposed to employing a single ODE in FM. The core principle of SFM is that according to the inherent (joint) heterogeneity of the underlying distributions, i.e., (jointly) dependent on the source or/and target data samples, a specific ODE should be selected from the pool of the candidate ODEs to facilitate the transportation process while preserving the marginal vector fields or probability paths.

To summarize, the major contributions of this study are multi-folded, including:

1. **Development of SFM**: We establish SFM, a versatile continuous-time generative model that eliminates singularities encountered in the FM via switching the candidate ODEs, and allows the intersection of probability paths from different ODEs.

2. **Theoretical insights**: Through rigorous analysis, we demonstrate that FM struggles with transporting between simple distributions due to the existence and uniqueness of initial value problems of ODEs while such limitation can be effectively addressed by SFM, offering a more efficient solution.

3. **Integration with advanced techniques**: SFM can seamlessly integrate with the existing advanced techniques, for example, minibatch optimal transport, to further enhance the straightness of the flow, facilitating a more efficient sampling process.

4. **Empirical validation**: We validate the effectiveness of the newly proposed SFM through extensive experiments on both synthetic and real-world datasets, achieving competitive or even better performance compared to existing methods, such as FM.

**Organization**. The rest of this article is organized as follows. Section 2 introduces some preliminaries on (neural) ODEs, continuous normalizing flows, flow matching, and optimal transport. In Sec. 3, we theoretically show the limitations of FM. Then, we present the SFM in Sec. 4. Related works are discussed in Sec. 5. In Sec. 6, we provide numerical verifications on synthetic and real-world datasets. Finally, we conclude the article in Sec. 7, and all the details of this work are found in the appendices.

**Notations.** Before ending this section, we provide the following notations that will be used throughout the article: $\mathbb{R}$ (resp. $\mathbb{R}^+$) – the set of (resp. positive) real numbers; $\mathbb{R}^d$ – the Euclidean space; $\|\cdot\|$ – the $d$-dimensional ($d$-d) Euclidean norm; $\nabla$ and $\nabla\cdot$ – the gradient and divergence operator, respectively; $\mathbf{1}_d$ – the $d$-d vector with all elements being 1; $\boldsymbol{I}_d$ – the $d$-d identity matrix; $\mathrm{Tr}(A)$ – the trace of the square matrix $A \in \mathbb{R}^{d\times d}$; $\delta_{\boldsymbol{x}}$ – the Dirac mass at the point $\boldsymbol{x} \in \mathbb{R}^d$; $\mathcal{P}(\mathbb{R}^d)$ – the space of Borel probability measures on $\mathbb{R}^d$; For given $q_0 \in \mathcal{P}(\mathbb{R}^d)$ and $q_1 \in \mathcal{P}(\mathbb{R}^d)$, then $\Pi(q_0, q_1)$ is defined as the set of all joint probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ whose marginals are $q_0$ and $q_1$, and $q \in \Pi(q_0, q_1)$ is called a coupling between $q_0$ and $q_1$; $\mathcal{U}(a, b)$ – the uniform distribution over the interval $[a, b]$; $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ – the multivariate Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$; $\mathcal{H}^d$ – the $d$-d Hausdorff measure (with suitable normalization); $|S|$ – the cardinality of the set $S$.

## 2. Preliminaries

### 2.1. ODE and Probability Flows

**Definition 2.1** (O'Searcoid (2006); Villani (2009))**.** A map $\boldsymbol{f} : \mathcal{X} \to \mathcal{Y}$ between metric spaces $(\mathcal{X}, d_\mathcal{X})$ and $(\mathcal{Y}, d_\mathcal{Y})$ is said to be Lipschitz continuous (or $L$-Lipschitz) if $d_\mathcal{Y}[\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}')] \le L d_\mathcal{X}(\boldsymbol{x}, \boldsymbol{x}')$ for all $\boldsymbol{x}$, $\boldsymbol{x}'$ in $\mathcal{X}$. The best admissible constant $L$ is called the Lipschitz constant of $\boldsymbol{f}$, denoted by $\|\boldsymbol{f}\|_{\mathrm{Lip}}$.

The Cauchy problem or the initial value problem (IVP) is defined as the time-dependent Ordinary Differential Equation (ODE) of the following general form:

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{u}_t(\boldsymbol{x}), \quad t \in [0, 1], \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \quad (1)$$

where $\boldsymbol{u}_t(\boldsymbol{x}) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is a smooth[1] vector field. The solution $\boldsymbol{x}(t)$ of this ODE (1) induces a map, called the time-dependent flow: $\boldsymbol{\phi}_t(\boldsymbol{x}_0) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$, defined as $\boldsymbol{\phi}_t(\boldsymbol{x}_0) := \boldsymbol{x}(t)$. For a given initial distribution $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$, the above ODE (1) induces the associated probability flows $p_t(\boldsymbol{x}) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^+$, satisfying the continuity equation (Pedlosky, 2013):

$$\frac{\partial p_t(\boldsymbol{x})}{\partial t} = -\nabla \cdot [p_t(\boldsymbol{x})\boldsymbol{u}_t(\boldsymbol{x})], \qquad (2)$$

with the initial condition $p_0(\boldsymbol{x}_0) = q_0(\boldsymbol{x}_0)$. Typically, $(\boldsymbol{\phi}_t)_{\#}p_0$ stands for the image measure or push-forward of $p_0$ by $\boldsymbol{\phi}_t$. In addition, if, for a given target distribution $\boldsymbol{x}_1 \sim q_1(\boldsymbol{x}_1)$, it holds $p_1(\boldsymbol{x}_1) = q_1(\boldsymbol{x}_1)$, then the set of all these vector fields satisfying the boundary conditions is defined as $U(q_0, q_1)$.

## 2.2. Continuous Normalizing Flow

Chen et al. (2018) proposed a continuous-time generative model, called the Continuous Normalizing Flow (CNF), that can be trained via performing maximum likelihood estimation. Specifically, the generative process works by first sampling data points from the source distribution $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$. Then, these data points are transformed into different ones by solving the initial value problem of the neural ODE (NODE) (Chen et al., 2018):

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta}), \quad t \in [0,1], \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \qquad (3)$$

where $\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})$ is a parameterized neural network with the trainable weights $\boldsymbol{\theta}$ and the flow map is defined as $\boldsymbol{\varphi}_t(x_0; \boldsymbol{\theta})$. The object is that the final states $\boldsymbol{x}(1)$ from the above ODE (3) should constitute the target data instances. In addition, based on the instantaneous change of variables formula (Chen et al., 2018), the change in log probability follows a second ODE:

$$\frac{\mathrm{d}\log p_t(\boldsymbol{x})}{\mathrm{d}t} = -\mathrm{Tr}\left[\frac{\partial \boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})}{\partial \boldsymbol{x}}\right], \qquad (4)$$

resulting in the total change in log density as follows:

$$\log p_1(\boldsymbol{x}) = \log q_0(\boldsymbol{x}_0) - \int_0^1 \mathrm{Tr}\left[\frac{\partial \boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})}{\partial \boldsymbol{x}}\right] \mathrm{d}t. \qquad (5)$$

Finally, the CNF can be trained by maximizing (5). We note that the CNF requires simulating the ODEs (3) and (4) during training, yielding high computational costs.

---

[1] In this work, we assume that the vector field is (locally) Lipschitz continuous in both arguments $t$ and $\boldsymbol{x}$ and thereby the Picard's existence theorem (Arnold, 1992) guarantees the existence and uniqueness of the solution locally defined on a maximal time interval.

## 2.3. (Conditional) Flow Matching

Different from the training of the CNF as well as its objective, Lipman et al. (2022) proposed Flow Matching (FM), a simple simulation-free training method that employs a stable objective by regressing a target vector field $\boldsymbol{u}_t(\boldsymbol{x})$ that generates the desired probability paths $p_t(\boldsymbol{x})$, satisfying $p_0[\boldsymbol{x}(0)] = q_0(\boldsymbol{x}_0)$ and $p_1[\boldsymbol{x}(1)] = q_1(\boldsymbol{x}_1)$. Then, the regression objective is

$$\mathcal{L}_{\mathrm{FM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,p_t(\boldsymbol{x})} \|\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta}) - \boldsymbol{u}_t(\boldsymbol{x})\|^2, \qquad (6)$$

where $t \sim \mathcal{U}(0,1)$ and $\boldsymbol{x}(t) \sim p_t(\boldsymbol{x})$. Ideally, when the above objective (6) approaches zero, the learned vector field $\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})$ will generate $p_t(\boldsymbol{x})$. However, this objective (6) is, in general, computationally intractable without knowing the explicit forms of $\boldsymbol{u}_t(\boldsymbol{x})$ and $p_t(\boldsymbol{x})$.

Regarding this intractable issue, Conditional FM (CFM) (Lipman et al., 2022; Pooladian et al., 2023; Tong et al., 2023a;b) employs a simpler and tractable regression objective to effectively learn the vector field $\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})$ by incorporating a latent condition $\boldsymbol{z}$:

$$\mathcal{L}_{\mathrm{CFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,q(\boldsymbol{z}),p_t(\boldsymbol{x}|\boldsymbol{z})}\|\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})\|^2, \quad (7)$$

which has the same gradient, w.r.t. $\boldsymbol{\theta}$ as the FM objective (6) (Lipman et al., 2022; Pooladian et al., 2023; Tong et al., 2023a;b). Usually, $q(\boldsymbol{z})$ is chosen as an independent coupling between two distributions, i.e.,

$$q(\boldsymbol{z}) := q(\boldsymbol{x}_0, \boldsymbol{x}_1) = q_0(\boldsymbol{x}_0)q_1(\boldsymbol{x}_1), \qquad (8)$$

with $\boldsymbol{x}(t)$ being the linear interpolation of $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$:

$$\boldsymbol{x}(t) = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \qquad (9)$$

resulting in a constant speed vector field given $\boldsymbol{z}$:

$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}) = \boldsymbol{x}_1 - \boldsymbol{x}_0. \qquad (10)$$

This specific CFM model was also extensively investigated in prior research, notably in studies such as Liu et al. (2022); Albergo & Vanden-Eijnden (2022), where it is referred to as the rectified flow or the stochastic interpolant. In addition, $q(\boldsymbol{z})$ can be also selected as the (minibatch) optimal transport coupling (Fatras et al., 2019; Pooladian et al., 2023; Tong et al., 2023a;b). Here, we call these two methods independent CFM (I-CFM) and optimal transport CFM (OT-CFM).

## 2.4. Static and Dynamic Optimal Transport

The (static) optimal transport theory (Villani, 2009; Santambrogio, 2015; Peyré & Cuturi, 2019), a field in mathematics, focuses on efficiently transferring one distribution to another. Usually, the optimal transport cost between two measures
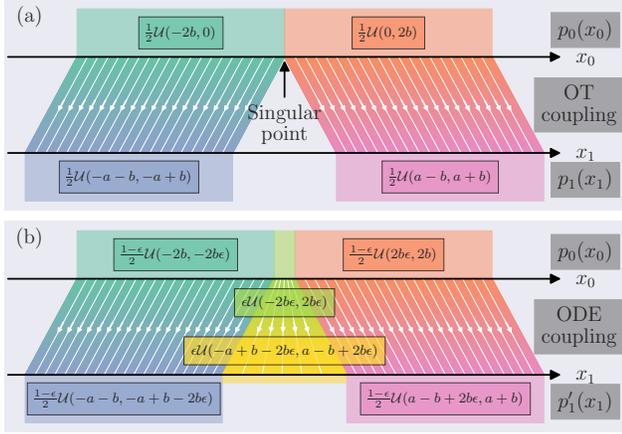
*Figure 1.* Illustration of the optimal transport (OT) coupling (a) and the ODE coupling (b) on the example in Proposition 3.1.



*Figure 2.* Illustration of the OT coupling (a) and the crossing coupling (b) on the example in Proposition 3.5.

is defined as the Kantorovich problem (Kantorovich, 1942), which can be described as follows:

$$C(q_0, q_1) = \inf_{\pi \in \Pi(q_0, q_1)} \int c(\boldsymbol{x}_0, \boldsymbol{x}_1) \mathrm{d}\pi(\boldsymbol{x}_0, \boldsymbol{x}_1), \quad (11)$$

where $c(\boldsymbol{x}_0, \boldsymbol{x}_1)$ is the cost for transporting one unit of mass from $\boldsymbol{x}_0$ to $\boldsymbol{x}_1$. In this paper, we consider the cost defined in terms of Euclidean distance, resulting in the following squared 2-Wasserstein distance:

$$W(q_0, q_1)^2 = \inf_{\pi \in \Pi(q_0, q_1)} \int \|\boldsymbol{x}_0 - \boldsymbol{x}_1\|^2 \mathrm{d}\pi(\boldsymbol{x}_0, \boldsymbol{x}_1). \quad (12)$$

Notably, the squared 2-Wasserstein distance has the equivalent dynamic form, known as the Benamou-Brenier formula (Benamou & Brenier, 1999; Brenier, 2003; Villani, 2009):

$$W(q_0, q_1)^2 = \inf_{\boldsymbol{u}_t \in U(q_0, q_1)} \int_0^1 \int p_t(\boldsymbol{x}) \|\boldsymbol{u}_t(\boldsymbol{x})\|^2 \mathrm{d}\boldsymbol{x} \mathrm{d}t. \quad (13)$$

## 3. Limitations of Folw Matching

In reality, the inherent (joint) heterogeneity of the source or/and target distributions may lead to a scenario where even an optimally trained FM model exhibits pronounced singularity. Consequently, this section aims to theoretically elucidate the limitations inherent to FM models through a series of propositions. All the details of the proofs are relegated to the appendices.

**Proposition 3.1** (Heterogeneity in $q_0$ or $q_1$). *Suppose the source distribution $q_0$ is an 1-d uniform distribution $q_0 = \mathcal{U}(-2b, 2b)$ and the target distribution $q_1$ is an 1-d uniform mixture (2-modes) $q_1 = \frac{1}{2}\mathcal{U}(-a - b, -a + b) + \frac{1}{2}\mathcal{U}(a -*

$b, a+b$), *where $a \gg b \geq 0$. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)).*

1. *If the NODE (3) exactly[2] solves the problem, then $x(0) = 0$ is a singular point, i.e., where the flow map $\varphi_1(0; \boldsymbol{\theta}) : x(0) = 0 \to x(1)$ is not well-defined or discontinuous (with two directions to $q_1$), as shown in Fig. 1(a).*

2. *If the NODE (3) approximately[3] solves the problem, resulting in an approximated target distribution $q_1'$, then there is a neighborhood $O$ of $x(0) = x_0$ which is homeomorphically mapped to the open subset in target space connecting the two modes, as shown in Fig. 1(b).*

3. *If the two modes of $q_1$ are far away from each other, i.e., $a \gg 1$, then the flow map $\varphi_1[x_0; \boldsymbol{\theta}]$ within a neighborhood $O$ as defined in the above-approximated NODE (the second bulletin) has a large Lipchitz constant.*

**Remark 3.2.** In Proposition 3.1, without loss of generality, we only consider the target distribution $q_1$ with heterogeneity (two modes). The intuition behind the theoretical results is simple. Within the context of bulletin 1 from Proposition 3.1, the mechanism of optimal transport coupling necessitates the division of $p_0$ into two symmetrical segments at the juncture $x(0) = 0$, directing these segments towards the dual modes of $q_1$. This process engenders a singularity at $x(0) = 0$, a direct consequence of $q_1$'s heterogeneity. For the bulletins 2 & 3 of Proposition 3.1, these statements are

---

[2]To be precise, $q_0$ can be completely transported to $q_1$ with the minimum of the squared 2-Wasserstein distance (13).

[3]A small fraction ($\epsilon \ll 1$) of the mass cannot be transferred from the source $q_0$ to the target $q_1$.

the consequences of the flow map $\varphi_t(x(0); \boldsymbol{\theta})$ being a homomorphism (more precisely, diffeomorphism), i.e. a bijective and continuous function whose inverse is also continuous.

A straightforward corollary emerging from Proposition 3.1 is articulated as follows.

**Corollary 3.3.** *Given the discrete distributions $q_0 = \delta_0$ and $q_1 = \frac{1}{2}\delta_{-a} + \frac{1}{2}\delta_a$, consider the optimal coupling $q(x_0 = 0, x_1 = \pm a) = \frac{1}{2}$, then it cannot be solved by an ODE. Furthermore, the learned flow map $\varphi_1(0; \boldsymbol{\theta})$ transfers the initial Dirac mass to some point $a'$ in the open set $(-a, a)$, i.e., $q_1' = \delta_{a'}$.*

**Remark 3.4.** Intuitively, to resolve the issue identified in Corollary 3.3, the flow map should assign the initial state to two disparate target states, thereby challenging the existence and uniqueness theorem of the IVP for a smooth ODE.

**Proposition 3.5** (Heterogeneity in both $q_0$ and $q_1$). *Suppose the source and target distributions $q_0$ and $q_1$ are two different $1$-d uniform mixtures (2-modes), respectively, i.e., $q_0 = \frac{2}{3}\mathcal{U}(-a - b, -a + b) + \frac{1}{3}\mathcal{U}(3a - b, 3a)$ and $q_1 = \frac{1}{3}\mathcal{U}(-3a, -3a + b) + \frac{2}{3}\mathcal{U}(a - b, a + b)$, where $a \gg b \geq 0$. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)). If the NODE (3) exactly solves the problem, then $x(0) = -a$ (reps., $x(1) = a$) is a singular point as shown in Fig. 2(a).*

**Remark 3.6.** The conceptual underpinnings of Proposition 3.5 closely mirror that of Proposition 3.1. However, the identified singularity originates from the heterogeneity present in both $q_0$ and $q_1$ under the optimal coupling induced by the squared 2-Wasserstein distance (12). Instead, a crossing coupling, illustrated in Fig. 2(b), enables an exact transportation between two large (resp., small) modes of $q_0$ and $q_1$, adeptly sidestepping any potential singularities. This coupling is locally optimal given the source and target modes, although it does not constitute a global optimum. Regrettably, achieving such coupling via a single ODE is impossible, as ODE trajectories cannot intersect (Arnold, 1992; Dupont et al., 2019; Zhang et al., 2020; Massaroli et al., 2020; Zhu et al., 2021; Liu et al., 2022).

**Proposition 3.7** (Infinite number of singular points). *Suppose the source and target distributions $q_0$ and $q_1$ are defined on $\mathbb{R}^2$ with $q_0$ being $\mathcal{H}^1$ restricted to $\{0\} \times [-1, 1]$, and $q_1$ being $(1/2)\mathcal{H}^1$ restricted to $\{-1, 1\} \times [-1, 1]$, respectively. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)). If the NODE (3) exactly solves the problem, then all the points $\boldsymbol{x}(0) = (0, a), a \in [-1, 1]$ are singular points as shown in Fig. 6(a).*

**Remark 3.8.** We note that Proposition 3.7 presents a quintessential example often employed to demonstrate the existence of a Monge minimizer, as detailed in(Villani, 2009). To achieve an optimal cost, one must split the mass at $(0, a)$ into two equal parts, and subsequently advance one

towards $(-1, a)$ and the other towards $(1, a)$. Although this procedure does not yield a conventional map (or Monge transport), one can approximate it via a discontinuous map with finite singular points as shown in Fig. 6(b). In addition, it is always possible to construct a better map (see Fig. 6(c)) by similarly incorporating additional singular points.

**Remark 3.9.** It is worth noting that the dimensionality of the manifold corresponding to the singular points in Proposition 3.7 is 1. Conversely, Propositions 3.1 and 3.5 are characterized by a manifold dimensionality of 0. In higher-dimensional cases, the singular points are encompassed within a stratified union of manifolds with distinct dimensions (Caffarelli, 1977; 1998; Figalli & Serra, 2019). To eliminate these singularities, it is essential to ensure that the cost functions, the spaces, and the probability measures meet adequate regularity assumptions, but this is often not the case when dealing with real-world data.

## 4. Switched Flow Matching

Inspired by the limitations of FM, we construct a new class of continuous-time generative models, referred as to Switched FM (SFM) which solves the transport problem between source and target distributions via switching multiple ODEs, particularly eliminating the singularities encountered in FM using a single ODE. The comparison of the FM and SFM are summarized in Table 1.

*Table 1.* Properties for the ODE-based generative models, including the FM, CFM, and our proposed SFM. Particularly, the SFM can not only handle general source distributions, and optimal transport flows (OT-SFM), but also employ multiple ODEs to eliminate the singularity, allowing the intersection of trajectories from different ODEs, and owning the relatively good regularity.

| ODE model | General source | OT | Mult. ODEs | Intersection | Regularity |
|---|---|---|---|---|---|
| FM | ✗ | ✗ | ✗ | ✗ | ✗ |
| I-CFM | ✓ | ✗ | ✗ | ✗ | ✗ |
| OT-CFM | ✓ | ✓ | ✗ | ✗ | ✗ |
| I-SFM | ✓ | ✗ | ✓ | ✓ | ✓ |
| OT-SFM | ✓ | ✓ | ✓ | ✓ | ✓ |

### 4.1. Formulation

Consider the source (resp., target) distribution, denoted as $q_0(\boldsymbol{x})$ (resp., $q_1(\boldsymbol{x})$), which is modeled as a mixture of conditional distributions $q_0(\boldsymbol{x}|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x}|\boldsymbol{s})$) that vary in response to a latent conditioning variable $\boldsymbol{s}$, termed the switching signal. Mathematically, this is expressed as:

$$q_i(\boldsymbol{x}) = \int q_i(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})\mathrm{d}\boldsymbol{s}, \quad i \in \{0, 1\}, \quad (14)$$

where $q^\circ(\boldsymbol{s})$ represents the distribution over the switching signal. Correspondingly, the marginal probability path $p_t(\boldsymbol{x})$
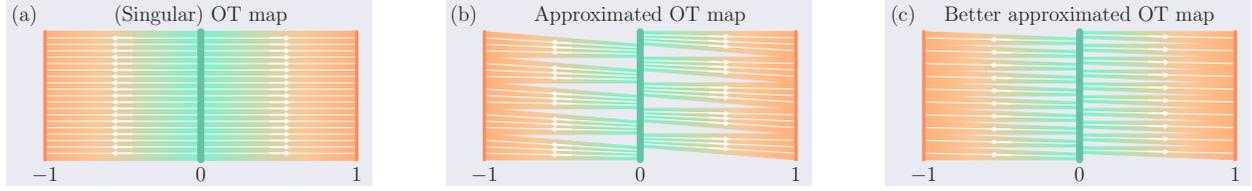
*Figure 3.* Illustration of the (singular) OT map (a) and the (better) approximated OT maps (b) & (c) on the example in Proposition 3.7.

is modeled as a mixture of probability paths $p_t(\boldsymbol{x}|\boldsymbol{s})$ of the following form:

$$p_t(\boldsymbol{x}) = \int p_t(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})\mathrm{d}\boldsymbol{s}, \qquad (15)$$

where $p_t(\boldsymbol{x}|\boldsymbol{s})$ should satisfy the boundary conditions, i.e., $p_0(\boldsymbol{x}|\boldsymbol{s}) = q_0(\boldsymbol{x}|\boldsymbol{s})$ and $p_1(\boldsymbol{x}|\boldsymbol{s}) = q_1(\boldsymbol{x}|\boldsymbol{s})$, implying $p_0(\boldsymbol{x}) = q_0(\boldsymbol{x})$ and $p_1(\boldsymbol{x}) = q_1(\boldsymbol{x})$. We assume that each conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{s})$ arises from a corresponding conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$. Significantly, our proposed SFM involves switching these ODEs rather than relying on a single ODE in FM (6). The corresponding sampling process is formalized as follows.

**Proposition 4.1** (Switching ODEs). *The marginal probability path $p_t(\boldsymbol{x})$ can be effectively sampled by switching ODEs in the following three steps:*

1. ***Sampling an ODE.*** *Sampling a switching signal $\boldsymbol{s}$ from the distribution $q^\circ(\boldsymbol{s})$, resulting in the specified ODE $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$;*

2. ***Sampling an initial state.*** *Sampling an initial state $\boldsymbol{x}_0$ (resp., backward one $\boldsymbol{x}_1$) from the conditional distribution $q_0(\boldsymbol{x}_0|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x}_1|\boldsymbol{s})$);*

3. ***Solving the IVP.*** *Generating the corresponding conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{s})$ by the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ from the initial state $\boldsymbol{x}_0$ (resp., $\boldsymbol{x}_1$).*

**Remark 4.2.** In this work, we are interested in the simple switching mechanism where the $q^\circ(\boldsymbol{s})$ and $q_0(\boldsymbol{x}_0|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x}_1|\boldsymbol{s})$) are both easily sampled, which will be presented in the Subsection 4.4. Additionally, these conditional vector fields $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$, in turn, collectively generate a marginal vector field, obtained by "marginalizing" over them as follows:

$$\boldsymbol{u}_t(\boldsymbol{x}) := \int \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})\frac{p_t(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})}{p_t(\boldsymbol{x})}\mathrm{d}\boldsymbol{s}, \qquad (16)$$

where $p_t(\boldsymbol{x}) > 0$ for all $t$ and $\boldsymbol{x}$. Crucially, as pointed out in the existing studies (Lipman et al., 2022; Pooladian et al., 2023; Tong et al., 2023a;b), the marginal vector field (16) actually generates the marginal probability path (15). However, using a single ODE to solve the transportation problem may inevitably encounter the singularity problem due to the inherent (joint) heterogeneity of the source and/or target distributions as discussed in the Section 3.

## 4.2. Training Objective

To mitigate the issue of singularity, our study aims to directly approximate the conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ by the learnable one $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$ using the following SFM objective:

$$\mathcal{L}_{\mathrm{SFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,q^\circ(\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{s})}\|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})\|^2. \quad (17)$$

Simply put, the SFM loss (17) regresses the conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ with a neural network $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$ via consistently sharing the parameter vector $\boldsymbol{\theta}$ across all switching signals $\boldsymbol{s}$. Upon minimizing the SFM loss to zero, an efficient sampling mechanism is enabled by the replacement of $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ with $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$ as proposed in Proposition 4.1.

However, akin to the FM (6), the SFM objective (17) becomes intractable in the absence of prior knowledge regarding the appropriate forms of $p_t(\boldsymbol{x}|\boldsymbol{s})$ and $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$. To address this issue, similar to the CFM (7), we further introduce a latent variable $\boldsymbol{z}$, and by marginalizing the conditional probability paths over $q(\boldsymbol{z}|\boldsymbol{s})$, we have the marginal probability path condition on $\boldsymbol{s}$,

$$p_t(\boldsymbol{x}|\boldsymbol{s}) = \int p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})q(\boldsymbol{z}|\boldsymbol{s})\mathrm{d}\boldsymbol{z}. \qquad (18)$$

Akin to the marginal vector field (16), we can also obtain the marginal vector field given $\boldsymbol{s}$, i.e., $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$, by marginalizing over the conditional vector fields $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})$ in the following sense,

$$\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s}) := \int \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})\frac{p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})q(\boldsymbol{z}|\boldsymbol{s})}{p_t(\boldsymbol{x}|\boldsymbol{s})}\mathrm{d}\boldsymbol{z}, \quad (19)$$

where $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})$ is the conditional vector field that generates $p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})$, yielding the following result.

**Proposition 4.3.** *Given the switching signal $s$, the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ in Eq. (19) generates the probability path $p_t(\boldsymbol{x}|\boldsymbol{s})$ in Eq. (18).*

Similar to the CFM (7), we then consider the Switching Conditional FM (SCFM) objective:

$$\mathcal{L}_{\mathrm{SCFM}}(\boldsymbol{\theta}) = \mathbb{E}_{t,q^\circ(\boldsymbol{s}),q(\boldsymbol{z}|\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})}\|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$$
$$- \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})\|^2. \quad (20)$$

Then, we have the following result.

**Proposition 4.4.** *Assuming that $p_t(\boldsymbol{x}|\boldsymbol{s}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $t \in [0, 1]$, then, up to a constant independent of $\boldsymbol{\theta}$, $\mathcal{L}_{SCFM}(\boldsymbol{\theta})$ and $\mathcal{L}_{SFM}(\boldsymbol{\theta})$ are equal. Hence, $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{SCFM}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}_{SFM}(\boldsymbol{\theta})$.*

**Remark 4.5.** The above result is actually the same as studied in Lipman et al. (2022); Pooladian et al. (2023); Tong et al. (2023a;b) if we consider the switching signal $s$ is a dumb variable, i.e., $q_i(\boldsymbol{x}|\boldsymbol{s}) = q_i(\boldsymbol{x}), i \in \{0, 1\}, q(\boldsymbol{z}|\boldsymbol{s}) = q(\boldsymbol{z}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) = \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z})$, and $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}) = \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta})$.

The SCFM objective (20) is useful when the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ is intractable but the conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})$ is simple even in a closed form.

### 4.3. Coupling

As delineated in Eqs. (8)-(10), one can also choose $q(\boldsymbol{z}|\boldsymbol{s})$ as an independent coupling condition on $\boldsymbol{s}$, i.e.,

$$q(\boldsymbol{z}|\boldsymbol{s}) := q(\boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{s}) = q_0(\boldsymbol{x}_0|\boldsymbol{s})q_1(\boldsymbol{x}_1|\boldsymbol{s}), \quad (21)$$

resulting in the linear interpolation $\boldsymbol{x}(t)$ and the constant speed vector field condition on both $\boldsymbol{z}$ and $\boldsymbol{s}$:

$$\boldsymbol{x}(t) = (1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \quad \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) = \boldsymbol{x}_1 - \boldsymbol{x}_0. \quad (22)$$

In addition, another choice of $q(\boldsymbol{z}|\boldsymbol{s})$ is the optimal coupling (Pooladian et al., 2023; Tong et al., 2023a;b) in terms of the squared 2-Wasserstein distance condition on $\boldsymbol{s}$, namely,

$$q(\boldsymbol{z}|\boldsymbol{s}) := q^*(\boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{s}), \quad (23)$$

where $\boldsymbol{z}$ represents a pair of points $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$. Contrary to independently sampling them from their conditional distributions (21), these points are jointly sampled in accordance with the optimal coupling $q^*(\boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{s})$ condition on $\boldsymbol{s}$. Here, we also use the simple vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})$ as defined in Eq. (22) in the SCFM objective (20). We then propose the following result.

**Proposition 4.6.** *Consider the optimal coupling $q^*(\boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{s})$ and the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})$ as defined in Eq. (22), then the optimal vector field $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$ in Eq. (20) solves the dynamic optimal transport problem (13) (condition on $\boldsymbol{s}$) between $q_0(\boldsymbol{x_0}|\boldsymbol{s})$ and $q_1(\boldsymbol{x_1}|\boldsymbol{s})$.*

**Remark 4.7.** If we consider the switching signal $s$ as a dumb variable, then the above result is actually the same as studied in Tong et al. (2023a;b). However, using a single ODE to solve the dynamic optimal transport problem (13) may not satisfy certain regularity assumptions. For example, the support of a distribution needs to be connected, which is often not the case in reality as discussed in the Section 3. On the contrary, to eliminate the singularities, the SFM uses multiple ODEs to solve it, which is conditionally or locally optimal (see the next subsection).

In practice, this optimal coupling can be approximated by addressing optimal transport problems within a given data batch (Pooladian et al., 2023; Tong et al., 2023a;b). Specifically, for each data batch $\left\{\boldsymbol{x}_0^{(k)}\right\}_{k=1}^m \sim q_0(\boldsymbol{x}_0|\boldsymbol{s})$ and $\left\{\boldsymbol{x}_1^{(k)}\right\}_{k=1}^m \sim q_1(\boldsymbol{x}_1|\boldsymbol{s})$, the optimal transport problem (12) condition on $\boldsymbol{s}$ for the discrete case can be exactly and efficiently resolved using standard solvers, such as the `POT` (Flamary et al., 2021, Python Optimal Transport).

Here, we call these two methods independent SFM (I-SFM) and optimal transport SFM (OT-SFM).

### 4.4. Switching Mechanism

Motivated by our observations and theories, we focus on constructing a simple and efficient switching mechanism such that the $q^\circ(\boldsymbol{s})$ and $q_0(\boldsymbol{x}_0|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x}_1|\boldsymbol{s})$) are both easily sampled for the general source and target distributions. One possible way is to employ the classic clustering methods to partition the empirical source (resp., target) dataset $\boldsymbol{X}_0 \sim q_0(\boldsymbol{x}_0)$ (resp., $\boldsymbol{X}_1 \sim q_1(\boldsymbol{x}_1)$) into $K_0$ (resp., $K_1$) sets, i.e., $\boldsymbol{X}_0^{(1)}, ..., \boldsymbol{X}_0^{(K_0)}$ (resp., $\boldsymbol{X}_1^{(1)}, ..., \boldsymbol{X}_1^{(K_1)}$). In addition, we assign each set $\boldsymbol{X}_0^{(i)}$ (resp., $\boldsymbol{X}_1^{(j)}$) a label $y_0^{(i)}$ (resp., $y_1^{(j)}$) and its weight or mass $\rho_0^{(i)} = |\boldsymbol{X}_0^{(i)}|/|\boldsymbol{X}_0|$ (resp., $\rho_1^{(j)} = |\boldsymbol{X}_1^{(j)}|/|\boldsymbol{X}_1|$).

**General setup.** We then construct the switching mechanism in the following manner:

1. $\boldsymbol{s}$ is a discrete variable, defined as $\boldsymbol{s} := (y_0, y_1) \in \left\{(y_0^{(i)}, y_1^{(j)})|i = 1, ..., K_0, j = 1, ..., K_1\right\}$;

2. $q^\circ(\boldsymbol{s}) := q^\circ(y_0, y_1)$ is a discrete (joint) distribution, defined as a coupling matrix $P$, satisfying the conservation of mass ($K_0 + K_1$ equality constraints),

$$\sum_{j=1}^{K_1} P(i, j) = \rho_0^{(i)}, \quad \sum_{i=1}^{K_0} P(i, j) = \rho_1^{(j)}, \quad (24)$$

where the element $P(i, j) \geq 0$ describes the amount of mass flowing from the bin $i$ (or the set $\boldsymbol{X}_0^{(i)}$) towards the bin $j$ (or the set $\boldsymbol{X}_1^{(j)}$);

3. $q_0(\boldsymbol{x_0}|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x_1}|\boldsymbol{s})$) is an empirical data distribution available as finite samples, i.e., $\boldsymbol{X}_0^{(i)}$ (resp., $\boldsymbol{X}_1^{(j)}$).

By choosing the different coupling matrix $P$, we induce the different switching signal distributions $q^\circ(\boldsymbol{s}) = q^\circ(y_0, y_1)$.

**Optimal transport setup.** If $P^*$ is the solution of the discrete Kantorovich's optimal transport problem, i.e.,

$$P^* = \arg\min_P \langle C, P \rangle := \sum_{i,j} C(i,j)P(i,j), \quad (25)$$

where $C(i, j)$ is the cost of moving a single unit from bin $i$ to bin $j$, then $P^*$ has the following property.

**Proposition 4.8** (Extremal solutions (Peyré & Cuturi, 2019))**.** $P^*$ *cannot have more than* $K_0 + K_1 - 1$ *nonzero entries, i.e.,* $|\{(y_0^{(i)}, y_1^{(j)})|P^*(i, j) > 0\}| \leq K_0 + K_1 - 1$.

**Remark 4.9.** In practice, $C(i, j)$ is simply assigned as either a uniform constant or as a function representing the appropriate distance between the sets $\boldsymbol{X}_0^{(i)}$ and $\boldsymbol{X}_1^{(j)}$. Furthermore, according to Proposition 4.8, it is possible to reduce the number of states $\boldsymbol{s} = (y_0, y_1)$ from a higher-order complexity of $K_0 K_1$ to a linear complexity of $K_0 + K_1 - 1$.

## 5. Related Works

**Switched systems.** Mathematically, switched systems are hybrid dynamical systems that consist of a family of subsystems and a rule that determines the switching between them (Liberzon & Morse, 1999; Liberzon et al., 1999; Daafouz et al., 2002; Liberzon, 2003). Typically, the rules can be largely divided into state-dependent and time-dependent switching. It should be pointed out that these switchings occur during the evolution process of a system. In contrast, as shown in Proposition 4.1, our switching mechanism involves randomly sampling a system, and then keeping it unchanged over time.

**Conditional generation.** Class-conditional generation is a common and important task, whose goal is to generate a sample that belongs to a specified class of the target distributions via incorporating the class label into their models (Van den Oord et al., 2016; Nguyen et al., 2017; Odena et al., 2017; Ho et al., 2022). This can be regarded as a special case of our framework by setting the switching signal as the target label. Since there is a lot of literature on this topic and our goal is to theoretically elucidate the limitations of FM and to eliminate singularities raised by using a single ODE, it is beyond the scope of this paper to have a complete review of the existing literature.

## 6. Experiments

**Synthetic datasets.** Figure 4 shows the proposed I-SFM and OT-SFM on transporting an 1-d Gaussian mixture (2-modes) to another. It is observed that an appropriate switching rule can eliminate the singularity raised from the heterogeneities of source and target distributions, leading to better regularity. In other words, when the data is sampled near the singularity region, it is inevitable that both the I-SFM and OT-SFM tend to perform poorly, but our framework is capable of achieving relatively good results. In addition, OT-SFM leads to a straighter flow than I-SFM.

Figure 5 shows the learned flows of the I-SFM and the OT-SFM on the example of the infinite number of singular

points under the optimal coupling in Proposition 3.7.

**CIFAR-10 dataset.** Table 2 shows the image generation results of our SFM variants on the CIFAR-10 dataset. In contrast with the existing generative models, we, here, consider a general source distribution, a Gaussian mixture with two modes, instead of a standard Gaussian distribution. Therefore, we display the results of the I-CFM and OT-CFM as the baselines. Crucially, it is observed that the I-CFM performs poorly on this task due to the mode separation of the source distribution, while it worked well for the standard Gaussian distribution (Lipman et al., 2022; Liu et al., 2022; Tong et al., 2023a;b). In addition, the I-SFM (one2ten) and OT-SFM (one2ten) perform poorly as well, as they all treat the support of the source distribution as one mode. Other SFM variants that explicitly separate the two modes of the source distribution, all perform well even better than the OT-CFM. We note that the OT-SFM did not perform as well as expected in comparison to the I-SFM. We attribute the reason to the switching mechanism that has already alleviated singularities induced by mode separation.

*Table 2.* FID results of CFM and SFM on the CIFAR-10 dataset.

| NFE | 6 | 8 | 10 | 20 | 40 | Adap. |
|---|---|---|---|---|---|---|
| I-CFM (I-SFM, one2one) | 144.52 | 130.49 | 122.44 | 106.11 | 99.19 | 94.55 |
| OT-CFM (OT-SFM, one2one) | 176.80 | 111.09 | 76.41 | 26.15 | 10.90 | 4.91 |
| I-SFM (one2ten) | **109.24** | 98.47 | 93.48 | 83.41 | 78.33 | 75.06 |
| OT-SFM (one2ten) | 122.74 | 104.19 | 93.04 | 73.47 | 63.94 | 59.72 |
| I-SFM (two2one) | 177.99 | 115.05 | 78.46 | 23.91 | 9.18 | 5.21 |
| OT-SFM (two2one) | 185.44 | 121.21 | 84.32 | 28.18 | 11.11 | 5.64 |
| I-SFM (two2ten, mixed) | 132.41 | 75.83 | **49.53** | **15.60** | **6.98** | 4.27 |
| OT-SFM (two2ten, mixed) | 133.27 | 76.31 | 49.69 | 15.50 | 7.24 | 4.39 |
| I-SFM (two2ten, extremal) | 128.55 | **75.11** | 50.12 | 17.14 | 8.39 | **4.22** |
| OT-SFM (two2ten, extremal) | 149.50 | 88.33 | 58.25 | 18.59 | 8.86 | 4.40 |

## 7. Conclusion

In this article, we highlighted and analyzed the limitations of FM, where using a single ODE for generative modeling may inevitably encounter the singularity problem due to the inherent (joint) heterogeneity of the source and/or target distributions. To eliminate singularities, we proposed SFM via switching multiple ODEs, even allowing the intersection of trajectories from distinct ODEs while it is impossible for a single ODE. In addition, a simple and efficient switching mechanism was constructed for effective training and inference. From an orthogonal perspective, our framework can seamlessly integrate with the existing advanced techniques, such as minibatch optimal transport, to further enhance the straightness of each flow. We also demonstrated the exceptional efficacy of the proposed framework by using synthetic and real-world datasets. We hope that our findings and proposed framework can contribute to the advancement of the field of generative modeling.
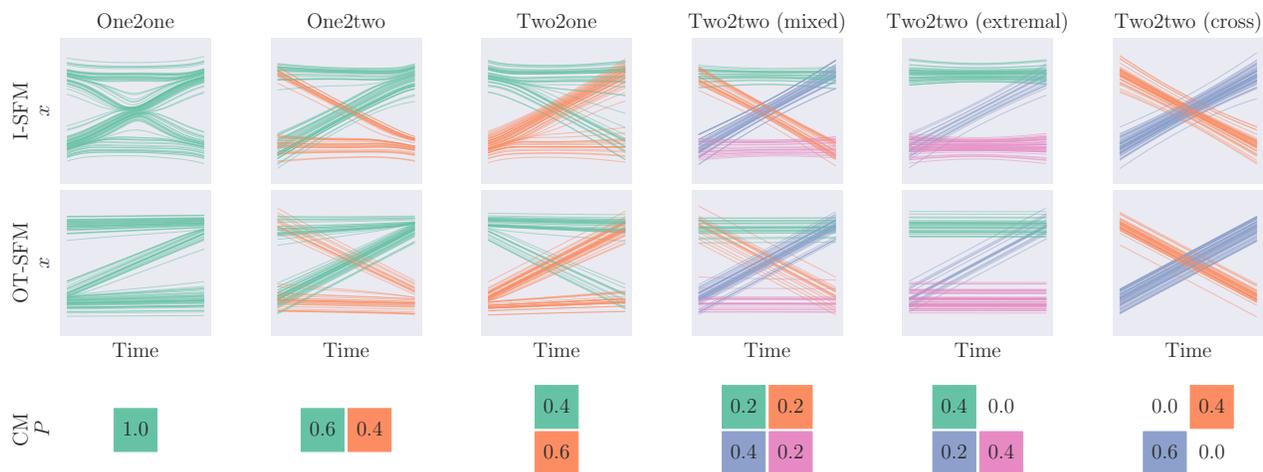
*Figure 4.* Trajectories of the I-SFM and the OT-SFM on 2-d Gaussian mixtures under different coupling matrices $P$ (from left to right). Particularly, in the first column ("one2one" coupling), the I-SFM and the OT-SFM are the I-CFM and OT-CFM, respectively.
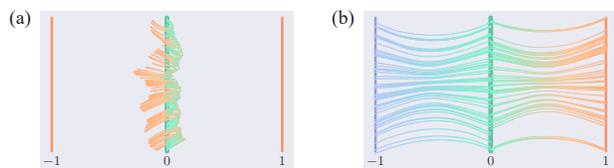


*Figure 5.* The learned flows of the I-CFM (a) and the I-SFM (one2two) (b) on the example in Proposition 3.7.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

## Acknowledgements

## References

Ahmad, S. and Ambrosetti, A. *A Textbook on Ordinary Differential Equations*, volume 88. Springer, 2015.

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

Arjovsky, M. and Bottou, L. Towards principled methods for training generative adversarial networks. *Arxiv Preprint Arxiv:1701.04862*, 2017.

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223. PMLR, 2017.

Arnold, V. I. *Ordinary Differential Equations*. Springer Science & Business Media, 1992.

Benamou, J.-D. and Brenier, Y. A numerical method for the optimal time-continuous mass transport problem and related problems. *Contemporary Mathematics*, 226:1–12, 1999.

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

Brenier, Y. Extended monge-kantorovich theory. *Optimal Transportation and Applications: Lectures Given at the Cime Summer School Held in Martina Franca, Italy, September 2–8, 2001*, pp. 91, 2003.

Caffarelli, L. A. The regularity of free boundaries in higher dimensions. *Acta Mathematica*, 139:155–184, 1977.

Caffarelli, L. A. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

Caffarelli, L. A. The obstacle problem revisited. *Journal of Fourier Analysis and Applications*, 4(4):383–402, 1998.

Chen, N., Zhang, Y., Zen, H., Weiss, R. J., Norouzi, M., and Chan, W. Wavegrad: Estimating gradients for waveform generation. *Arxiv Preprint Arxiv:2009.00713*, 2020.

Chen, R. T., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. *Advances in Neural Information Processing Systems*, 31, 2018.

Coddington, E. A., Levinson, N., and Teichmann, T. Theory of ordinary differential equations, 1956.

Daafouz, J., Riedinger, P., and Iung, C. Stability analysis and control synthesis for switched systems: A switched Lyapunov function approach. *IEEE Transactions on Automatic Control*, 47(11):1883–1887, 2002.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

Dinh, L., Krueger, D., and Bengio, Y. Nice: Non-linear independent components estimation. *Arxiv Preprint Arxiv:1410.8516*, 2014.

Dinh, L., Sohl-Dickstein, J., and Bengio, S. Density estimation using real nvp. *Arxiv Preprint Arxiv:1605.08803*, 2016.

Du, Y. and Mordatch, I. Implicit generation and generalization in energy-based models. *Arxiv Preprint Arxiv:1903.08689*, 2019.

Dupont, E., Doucet, A., and Teh, Y. W. Augmented neural ODEs. In *Advances in Neural Information Processing Systems*, pp. 3140–3150, 2019.

Farnia, F. and Ozdaglar, A. Do gans always have nash equilibria? In *International Conference on Machine Learning*, pp. 3029–3039. PMLR, 2020.

Fatras, K., Zine, Y., Flamary, R., Gribonval, R., and Courty, N. Learning with minibatch wasserstein: asymptotic and gradient properties. *arXiv preprint arXiv:1910.04091*, 2019.

Figalli, A. and Serra, J. On the fine structure of the free boundary for the classical obstacle problem. *Inventiones Mathematicae*, 215(1):311–366, 2019.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati,

H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Germain, M., Gregor, K., Murray, I., and Larochelle, H. Made: Masked autoencoder for distribution estimation. In *International Conference on Machine Learning*, pp. 881–889. PMLR, 2015.

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.

Grathwohl, W., Chen, R. T., Bettencourt, J., Sutskever, I., and Duvenaud, D. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *Arxiv Preprint Arxiv:1810.01367*, 2018.

Hinton, G. E. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Ho, J., Saharia, C., Chan, W., Fleet, D. J., Norouzi, M., and Salimans, T. Cascaded diffusion models for high fidelity image generation. *The Journal of Machine Learning Research*, 23(1):2249–2281, 2022.

Howard, R. The gronwall inequality. *Lecture Notes*, 1998.

Kantorovich, L. V. On the translocation of masses. In *Dokl. Akad. Nauk. Ussr*, volume 37, pp. 199–201, 1942.

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

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., and Huang, F. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

Lee, S., Kim, B., and Ye, J. C. Minimizing trajectory curvature of ode-based generative models. *Arxiv Preprint Arxiv:2301.12003*, 2023.

Liberzon, D. *Switching in Systems and Control*, volume 190. Springer, 2003.

Liberzon, D. and Morse, A. S. Basic problems in stability and design of switched systems. *IEEE Control Systems Magazine*, 19(5):59–70, 1999.

Liberzon, D., Hespanha, J. P., and Morse, A. S. Stability of switched systems: A lie-algebraic condition. *Systems & Control Letters*, 37(3):117–122, 1999.

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, 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.

Massaroli, S., Poli, M., Park, J., Yamashita, A., and Asama, H. Dissecting neural ODEs. *Arxiv Preprint Arxiv:2002.08071*, 2020.

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

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., and Yosinski, J. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4467–4477, 2017.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *Arxiv Preprint Arxiv:2112.10741*, 2021.

Odena, A., Olah, C., and Shlens, J. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pp. 2642–2651. PMLR, 2017.

Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. Wavenet: A generative model for raw audio. *Arxiv Preprint Arxiv:1609.03499*, 2016.

O'Searcoid, M. *Metric Spaces*. Springer Science & Business Media, 2006.

Pedlosky, J. *Geophysical Fluid Dynamics*. Springer Science & Business Media, 2013.

Peyré, G. and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11(5-6): 355–607, 2019.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. Multisample flow matching: Straightening flows with minibatch couplings. *Arxiv Preprint Arxiv:2304.14772*, 2023.

Rezende, D. and Mohamed, S. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pp. 1530–1538. PMLR, 2015.

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.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

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

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265. PMLR, 2015.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models. *Arxiv Preprint Arxiv:2010.02502*, 2020a.

Song, Y. and Ermon, S. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Song, Y. and Kingma, D. P. How to train your energy-based models. *Arxiv Preprint Arxiv:2101.03288*, 2021.

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*, 2020b.

Teh, Y. W., Welling, M., Osindero, S., and Hinton, G. E. Energy-based models for sparse overcomplete representations. *Journal of Machine Learning Research*, 4(Dec): 1235–1260, 2003.

Tong, A., Malkin, N., Fatras, K., Atanackovic, L., Zhang, Y., Huguet, G., Wolf, G., and Bengio, Y. Simulation-free schrödinger bridges via score and flow matching. *arXiv preprint arXiv:2307.03672*, 2023a.

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. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023b.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. Conditional image generation with pixelcnn decoders. *Advances in Neural Information Processing Systems*, 29, 2016.

Van Den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel recurrent neural networks. In *International Conference on Machine Learning*, pp. 1747–1756. PMLR, 2016.

Villani, C. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Weng, L. From gan to wgan. *Arxiv Preprint Arxiv:1904.08994*, 2019.

Younes, L. *Shapes and Diffeomorphisms*, volume 171. Springer, 2010.

Zhang, H., Gao, X., Unterman, J., and Arodz, T. Approximation capabilities of neural ODEs and invertible residual networks. In *International Conference on Machine Learning*, pp. 11086–11095. PMLR, 2020.

Zhu, Q., Guo, Y., and Lin, W. Neural delay differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Q1jmmQz72M2.

**Roadmap.** The structure of the appendix is outlined as follows:

- Appendix A presents the existing theoretical results on ODEs and optimal transport.

- Appendix B presents the proofs of our theoretical results.

- Appendix C presents additional results and discussions from the perspectives of the methodology, algorithm, and experiment.

- Appendix D presents the experimental details for all experiments conducted in the work.

## A. Existing theoretical results

In this section, before presenting the proofs of our theoretical results in the main text as well as the additional results in the appendix, we first review the existing theoretical results on ODEs.

Throughout the section, we consider the IVP:

$$\frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = \boldsymbol{u}_t(\boldsymbol{x}), \quad t \in [0,1],$$
$$\boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{26}$$

where $\boldsymbol{u}_t(\boldsymbol{x}) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ is a smooth vector field with a bounded Lipschitz constant $L := \|\boldsymbol{u}_t(\boldsymbol{x})\|_{\mathrm{Lip}}$. Then, we have the following existence and uniqueness theorem.

### A.1. Properties of ODEs

**Theorem A.1** (Global existence and uniqueness, (Ahmad & Ambrosetti, 2015)). *Suppose that $\boldsymbol{x} \in \mathbb{R}^d$, $t \in [0,1]$, and $\boldsymbol{u}_t(\boldsymbol{x})$ is continuous and globally lipschitzian in $\mathbb{R}^d$ with respect to $\boldsymbol{x}$, then the solution of (26) is unique and defined on all $t \in [0,1]$.*

**Theorem A.2** (Non-intersecting trajectories, (Coddington et al., 1956; Younes, 2010; Dupont et al., 2019)). *Let $\bar{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{x}}(t)$ be two solutions of the ODE (26) with different initial conditions, i.e., $\bar{\boldsymbol{x}}(0) \neq \hat{\boldsymbol{x}}(0)$. then for all $t \in (0,1]$, $\bar{\boldsymbol{x}}(t) \neq \hat{\boldsymbol{x}}(t)$. Informally, it states that ODE trajectories cannot intersect.*
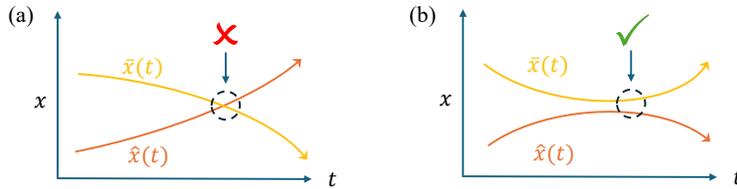


*Figure 6.* Illustration of the non-intersecting trajectories of ODEs. (a) The two trajectories intersect each other, which is not feasible for ODEs. (b) Any two trajectories cannot intersect each other at any time $t$.

**Theorem A.3** (Gronwall's inequality, (Howard, 1998; Dupont et al., 2019)). *Let $\boldsymbol{u}_t(\boldsymbol{x}) : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ be a continuous function and let $\bar{\boldsymbol{x}}(t)$ and $\hat{\boldsymbol{x}}(t)$ be two solutions of the ODE (26), satisfying the IVP:*

$$\frac{\mathrm{d}\bar{\boldsymbol{x}}(t)}{\mathrm{d}t} = \boldsymbol{u}_t[\bar{\boldsymbol{x}}(t)], \quad t \in [0,1], \quad \bar{\boldsymbol{x}}(0) = \bar{\boldsymbol{x}}_0,$$
$$\frac{\mathrm{d}\hat{\boldsymbol{x}}(t)}{\mathrm{d}t} = \boldsymbol{u}_t[\hat{\boldsymbol{x}}(t)], \quad t \in [0,1], \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0, \tag{27}$$

*Assume there is a constant $L \geq 0$ such that*

$$\|\boldsymbol{u}_t[\bar{\boldsymbol{x}}(t)] - \boldsymbol{u}_t[\hat{\boldsymbol{x}}(t)]\| \leq L\|\bar{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t)\|, \tag{28}$$

*Then for $t \in [0,1]$, we have*

$$\|\bar{\boldsymbol{x}}(t) - \hat{\boldsymbol{x}}(t)\| \leq e^{Lt}\|\bar{\boldsymbol{x}}_0 - \hat{\boldsymbol{x}}_0\|. \tag{29}$$

**Theorem A.4** (Homeomorphism, (Younes, 2010; Dupont et al., 2019))**.** *Consider the flow map $\phi_t(\boldsymbol{x}_0)$ of the ODE (26). Then, for all $t \in [0, 1]$, $\phi_t(\boldsymbol{x}_0)$ is a homomorphism, i.e.,*

1. *$\phi_t$ is continuous;*

2. *$\phi_t$ is a bijection;*

3. *$\phi_t^{-1}$ is continuous.*

**Remark A.5.** More precisely, the flow map $\phi_t(\boldsymbol{x}_0)$ of the ODE (26) is a diffeomorphism (Younes, 2010), but we will not use this stronger property in our proofs.

### A.2. Properties of optimal transport

**Theorem A.6** (Nondecreasing map, (Santambrogio, 2015))**.** *Given $q_0, q_1 \in \mathcal{P}(\mathbb{R})$, suppose that $q_0$ is atomless[4]. Then, there exists a unique nondecreasing map $T_{\mathrm{mon}} : \mathbb{R} \to \mathbb{R}$ such that $(T_{\mathrm{mon}})_{\#} q_0 = q_1$.*

**Lemma A.7** (Monotonic property, (Santambrogio, 2015))**.** *Let $\gamma \in \Pi(q_0, q_1)$ be a transport plan between two measures $q_0, q_1 \in \mathcal{P}(\mathbb{R})$. Suppose that it satisfies the property,*

$$
\begin{aligned}
(x_0, x_1), (x_0', x_1') &\in \mathrm{Spt}(\gamma), \\
x_0 < x_0' &\implies x_1 < x_1'.
\end{aligned}
\tag{30}
$$

*Then, we have $\gamma = \gamma_{\mathrm{mon}}$. In particular, there is a unique $\gamma$ satisfying (30). Moreover, if $q_0$ is atomless, then $\gamma = \gamma_{T_{\mathrm{mon}}}$.*

**Theorem A.8** (Optimality of the monotone map, (Santambrogio, 2015))**.** *Let $h : \mathbb{R} \to \mathbb{R}^+$ be a strictly convex function and $q_0, q_1 \in \mathcal{P}(\mathbb{R})$ be probability measures. Consider the cost $c(x_0, x_1) = h(x_1 - x_0)$ and suppose that the Kantorovich problem (11) has a finite value. Then, it has a unique solution, which is given by $\gamma_{\mathrm{mon}}$. In the case where $q_0$ is atomless, this optimal plan is induced by the map $T_{\mathrm{mon}}$.*

**Theorem A.9** (Cyclical monotonicity, (Villani, 2009))**.** *Let $\mathcal{X}, \mathcal{Y}$ be arbitrary sets, and $c : \mathcal{X} \times \mathcal{Y} \to (-\infty, +\infty]$ be a function. A subset $\Gamma \subseteq \mathcal{X} \times \mathcal{Y}$ is said to be c-cyclically monotone if, for any $N \in \mathbb{N}$, and any family $(\boldsymbol{x}_1, \boldsymbol{y}_1), ..., (\boldsymbol{x}_N, \boldsymbol{y}_N)$ of points in $\Gamma$, holds the inequality*

$$
\sum_{i=1}^{N} c(\boldsymbol{x}_i, \boldsymbol{y}_i) \le \sum_{i=1}^{N} c(\boldsymbol{x}_i, \boldsymbol{y}_{i+1})
\tag{31}
$$

*(with the convention $y_{N+1} = y_1$). A transport plan is said to be c-cyclically monotone if it is concentrated on a c-cyclically monotone set.*

**Remark A.10.** Informally, a $c$-cyclically monotone plan cannot be improved via perturbations. Consequently, it follows intuitively that an optimal plan should adhere the $c$-cyclical monotonicity (Villani, 2009).

**Theorem A.11** ((Villani, 2009))**.** *Let $(\mathcal{X}, \mu)$ and $(\mathcal{Y}, \nu)$ be any two Polish probability spaces, let $T$ be a continuous map $\mathcal{X} \to \mathcal{Y}$, and let $\pi = (\mathrm{Id}, T)_{\#}\mu$ be the associated transport map. Then, for each $\boldsymbol{x} \in Spt(\mu)$, the pair $(\boldsymbol{x}, T(\boldsymbol{x}))$ belongs to the support of $\pi$.*

## B. Proofs

**Proposition 3.1** (Heterogeneity in $q_0$ or $q_1$)**.** *Suppose the source distribution $q_0$ is an 1-d uniform distribution $q_0 = \mathcal{U}(-2b, 2b)$ and the target distribution $q_1$ is an 1-d uniform mixture (2-modes) $q_1 = \frac{1}{2}\mathcal{U}(-a - b, -a + b) + \frac{1}{2}\mathcal{U}(a - b, a + b)$, where $a \gg b \ge 0$. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)).*

1. *If the NODE (3) exactly[5] solves the problem, then $x(0) = 0$ is a singular point, i.e., where the flow map $\varphi_1(0; \boldsymbol{\theta})$ : $x(0) = 0 \to x(1)$ is not well-defined or discontinuous (with two directions to $q_1$), as shown in Fig. 1(a).*

2. *If the NODE (3) approximately[6] solves the problem, resulting in an approximated target distribution $q_1'$, then there is a neighborhood $O$ of $x(0) = x_0$ which is homeomorphically mapped to the open subset in target space connecting the two modes, as shown in Fig. 1(b).*

---

[4]For every $x \in \mathbb{R}$, the probabilistic measure on this single point $x$ is equal to zero.

[5]To be precise, $q_0$ can be completely transported to $q_1$ with the minimum of the squared 2-Wasserstein distance (13).

[6]A small fraction ($\epsilon \ll 1$) of the mass cannot be transferred from the source $q_0$ to the target $q_1$.

*3. If the two modes of $q_1$ are far away from each other, i.e., $a \gg 1$, then the flow map $\varphi_1[x_0; \boldsymbol{\theta}]$ within a neighborhood $O$ as defined in the above-approximated NODE (the second bulletin) has a large Lipchitz constant.*

*Proof of Proposition 3.1.* We routinely prove these three bulletins in the following.

1. If the NODE (3) exactly solves the optimal transport problem, then the flow map $\varphi_1(x_0; \boldsymbol{\theta})$ transports all mass from the source to the target and simultaneously achieves the minimum squared 2-Wasserstein distance. By construction of the source and target distributions, the support of the source distribution should be divided into two equal parts at the point 0, and each part is transported to one of the two disconnected components of the target support. In addition, from Lemma A.7 and Theorem A.8, the left (resp., right) part of the source support should be transported to the left part of the target support and this optimal transport map should satisfy the monotonic property (30), as illustrated in Fig. 1(a). Hence, the flow map $\varphi_1(0; \boldsymbol{\theta}) : x(0) = 0 \rightarrow x(1)$ is not well-defined or discontinuous at the point 0.

2. If the NODE (3) approximately solves the problem, resulting in an approximated target distribution $q_1'$, then there exists a small fraction of the mass that cannot be transferred from the source to the target. More precisely, based on the homeomorphism of the flow map $\varphi_1(x_0; \boldsymbol{\theta})$ (Theorem A.4), it should map the source connected support to another connected set, thereby preserving connectedness. However, the support of the target distribution has two disconnected components, implying that there is a neighborhood $O$ of $x(0) = x_0$ in the source support, which is homeomorphically mapped to the open subset connecting these two modes.

3. If the two modes of $q_1$ are far away from each other, i.e., $a \gg 1$, then it implies that the flow map $\varphi_1(x_0; \boldsymbol{\theta})$ maps the neighborhood $O$ of $x(0) = x_0$ to a large open subset connecting these two largely shifted modes, though the transported mass can be still small (say $\epsilon \ll 1$). Hence, the flow map $\varphi_1(x_0; \boldsymbol{\theta})$ has a large Lipchitz constant $L'$, satisfying

$$\|\varphi_1(\bar{x}_0; \boldsymbol{\theta}) - \varphi_1(\hat{x}_0; \boldsymbol{\theta})\| \leq L' \|\bar{x}_0 - \hat{x}_0\|. \tag{32}$$

   More precisely, the order of $L'$ is $\mathcal{O}(\frac{a}{\epsilon})$. In addition, based on the Gronwall's inequality (29), the Lipchitz constant $L$ of the vector field $\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})$ should satisfy that

$$\|\varphi_1(\bar{x}_0; \boldsymbol{\theta}) - \varphi_1(\hat{x}_0; \boldsymbol{\theta})\| \leq e^L \|\bar{x}_0 - \hat{x}_0\|, \quad \forall \bar{x}_0, \hat{x}_0 \in O, \tag{33}$$

   implying that the order of $L$ is $\mathcal{O}(\ln L') = \mathcal{O}\left(\ln \frac{a}{\epsilon}\right)$. Therefore, when $a \gg 1$, the flow map $\varphi_1[x_0; \boldsymbol{\theta}]$ within a neighborhood $O$ has a large Lipchitz constant $L$ of the order $\mathcal{O}\left(\ln \frac{a}{\epsilon}\right)$.

The proof is complete. $\qquad\square$

**Corollary 3.3.** *Given the discrete distributions $q_0 = \delta_0$ and $q_1 = \frac{1}{2}\delta_{-a} + \frac{1}{2}\delta_a$, consider the optimal coupling $q(x_0 = 0, x_1 = \pm a) = \frac{1}{2}$, then it cannot be solved by an ODE. Furthermore, the learned flow map $\varphi_1(0; \boldsymbol{\theta})$ transfers the initial Dirac mass to some point $a'$ in the open set $(-a, a)$, i.e., $q_1' = \delta_{a'}$.*

*Proof of Corollary 3.3.* Based on the optimal coupling $q(x_0 = 0, x_1 = \pm a) = \frac{1}{2}$, the mass at the point 0 should be divided into two halves, and one half is transported to $-a$, and the other half is transported to $a$. However, based on the existence and uniqueness of ODEs' solutions, the flow map induced by the ODE is a determined map that can only transport the mass at the point 0 to some point $a' := \varphi_1(0; \boldsymbol{\theta})$. Now, we claim $a' \in (-a, a)$. The optimal coupling $q(x_0 = 0, x_1 = \pm a) = \frac{1}{2}$ shows the singularity at the point $x(0) = 0$ with two directions. Hence, in an average way by eliminating the downward and upward components of these two directions, respectively, it leads to a horizon direction (between them) at the point $x(0) = 0$. However, for any point belonging to the lines $x(t) = ta$ or $x(t) = -ta$, $t \in (0, 1]$, the direction of this point is determined along the corresponding line. Moreover, by the Theorem A.2, the ideal learned solution $\varphi_t(0; \boldsymbol{\theta})$ is always sandwiched between these two lines, as illustrated in Fig. 7. Therefore, the learned flow map $\varphi_1(0; \boldsymbol{\theta})$ transfers the initial Dirac mass to some point $a'$ in the open set $(-a, a)$, i.e., $q_1' = \delta_{a'}$. $\qquad\square$

**Proposition 3.5** (Heterogeneity in both $q_0$ and $q_1$). *Suppose the source and target distributions $q_0$ and $q_1$ are two different 1-d uniform mixtures (2-modes), respectively, i.e., $q_0 = \frac{2}{3}\mathcal{U}(-a - b, -a + b) + \frac{1}{3}\mathcal{U}(3a - b, 3a)$ and $q_1 = \frac{1}{3}\mathcal{U}(-3a, -3a + b) + \frac{2}{3}\mathcal{U}(a - b, a + b)$, where $a \gg b \geq 0$. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)). If the NODE (3) exactly solves the problem, then $x(0) = -a$ (reps., $x(1) = a$) is a singular point as shown in Fig. 2(a).*
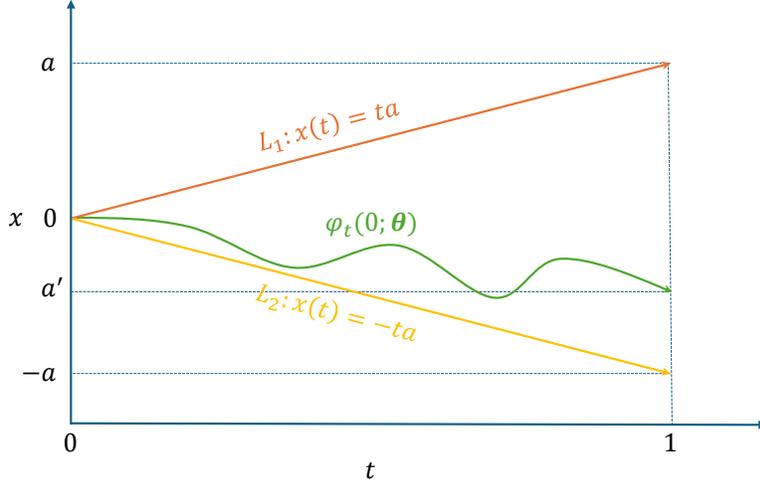
*Figure 7.* Illustration of the sandwiched property of ODEs in Corollary 3.3.

*Proof of Proposition 3.5.* The proof is similar to the one of the Proposition 3.1. Different from the case Proposition 3.1, here, the disconnectedness occurs in both the source distribution $q_0$ and the target distribution $q_1$, where each disconnected component of their supports has a distinct mass. The optimal transport coupling should split both the large components in the supports of the source and the target distributions into two equal parts, resulting in a total of three equal components, respectively. Therefore, the ODE is thereby singular at the split points $x(0) = -a$ and $x(1) = a$. $\qquad\square$

**Proposition 3.7** (Infinite number of singular points). *Suppose the source and target distributions $q_0$ and $q_1$ are defined on $\mathbb{R}^2$ with $q_0$ being $\mathcal{H}^1$ restricted to $\{0\} \times [-1, 1]$, and $q_1$ being $(1/2)\mathcal{H}^1$ restricted to $\{-1, 1\} \times [-1, 1]$, respectively. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)). If the NODE (3) exactly solves the problem, then all the points $\boldsymbol{x}(0) = (0, a), a \in [-1, 1]$ are singular points as shown in Fig. 6(a).*

*Proof of Proposition 3.7.* Indeed, to exactly solve the optimal transport problem, each source point $(0, a)$ has two expected target points $(-1, a)$ and $(1, a)$, since both of them require the same unit cost for $(0, a)$ to one of them. Therefore, the optimal flow map $\boldsymbol{\varphi}_1[(0, a); \boldsymbol{\theta}] : \boldsymbol{x}(t = 0) = (0, a) \to \boldsymbol{x}(1)$ is not well-defined (with two directions to $q_1$). $\qquad\square$

**Proposition 4.1** (Switching ODEs). *The marginal probability path $p_t(\boldsymbol{x})$ can be effectively sampled by switching ODEs in the following three steps:*

1. ***Sampling an ODE.*** *Sampling a switching signal $\boldsymbol{s}$ from the distribution $q^\circ(\boldsymbol{s})$, resulting in the specified ODE $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$;*

2. ***Sampling an initial state.*** *Sampling an initial state $\boldsymbol{x}_0$ (resp., backward one $\boldsymbol{x}_1$) from the conditional distribution $q_0(\boldsymbol{x}_0|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x}_1|\boldsymbol{s})$);*

3. ***Solving the IVP.*** *Generating the corresponding conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{s})$ by the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ from the initial state $\boldsymbol{x}_0$ (resp., $\boldsymbol{x}_1$).*

*Proof of Proposition 4.1.* Here, we aim to prove that the marginal probability path $p_t(\boldsymbol{x})$ can be equivalently sampled by switching ODEs in the above three steps.

By construction, we introduce a latent conditioning variable $\boldsymbol{s}$ to represent the source (resp., target) distribution $q_0(\boldsymbol{x})$ (resp., $q_1(\boldsymbol{x})$) as a mixture of conditional distributions $q_0(\boldsymbol{x}|\boldsymbol{s})$ (resp., $q_1(\boldsymbol{x}|\boldsymbol{s})$), satisfying (also see Eq. (14))

$$q_i(\boldsymbol{x}) = \int q_i(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})\mathrm{d}\boldsymbol{s}, \quad i \in \{0, 1\}, \tag{34}$$

where $q^\circ(\boldsymbol{s})$ is the distribution over the switching signal. In addition, we model the marginal probability path $p_t(\boldsymbol{x})$ as a mixture of probability paths $p_t(\boldsymbol{x}|\boldsymbol{s})$ (also see Eq. (15)),

$$p_t(\boldsymbol{x}) = \int p_t(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})\mathrm{d}\boldsymbol{s}. \tag{35}$$

where each conditional probability path $p_t(\boldsymbol{x}|\boldsymbol{s})$ arises from a corresponding conditional vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$, i.e., satisfying the continuity equation,

$$\frac{\partial p_t(\boldsymbol{x}|\boldsymbol{s})}{\partial t} = -\nabla \cdot [p_t(\boldsymbol{x}|\boldsymbol{s})\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})] \tag{36}$$

with the boundary conditions $p_0(\boldsymbol{x}_0|\boldsymbol{s}) = q_0(\boldsymbol{x}_0|\boldsymbol{s})$ and $p_1(\boldsymbol{x}_1|\boldsymbol{s}) = q_1(\boldsymbol{x}_0|\boldsymbol{s})$. Moreover, it holds the boundary marginal source and target distributions, i.e.,

$$\begin{aligned}
p_i(\boldsymbol{x}) &= \int p_i(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})\mathrm{d}\boldsymbol{s} \\
&= \int q_i(\boldsymbol{x}|\boldsymbol{s})q^\circ(\boldsymbol{s})\mathrm{d}\boldsymbol{s} \\
&= q_i(\boldsymbol{x}),
\end{aligned} \tag{37}$$

where $i \in \{0, 1\}$. Therefore, one can sample $p_t(\boldsymbol{x})$ through the above three steps. $\qquad\square$

**Proposition 4.3.** *Given the switching signal $s$, the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ in Eq. (19) generates the probability path $p_t(\boldsymbol{x}|\boldsymbol{s})$ in Eq. (18).*

*Proof of Proposition 4.3.* The proof is adapted from Lipman et al. (2022); Tong et al. (2023a;b).

Since $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})$ is the conditional vector field that generates $p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})$, it means that given the switching signal $s$ and the latent variable $\boldsymbol{z}$, $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})$ and $p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})$ satisfy the continuity equation:

$$\frac{\partial p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})}{\partial t} = -\nabla \cdot [p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})]. \tag{38}$$

Next, we check that given the switching signal $s$, $p_t(\boldsymbol{x}|\boldsymbol{s})$ and $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})$ satisfy the continuity equation:

$$\begin{aligned}
\frac{\partial p_t(\boldsymbol{x}|\boldsymbol{s})}{\partial t} &= \frac{\partial}{\partial t} \int p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})q(\boldsymbol{z}|\boldsymbol{s})\mathrm{d}\boldsymbol{z} \\
&= \int \left[ \frac{\partial}{\partial t} p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s}) \right] q(\boldsymbol{z}|\boldsymbol{s})\mathrm{d}\boldsymbol{z} \\
&= - \int \{ \nabla \cdot [p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})] \} q(\boldsymbol{z}|\boldsymbol{s})\mathrm{d}\boldsymbol{z} \\
&= -\nabla \cdot \int p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})q(\boldsymbol{z}|\boldsymbol{s})\mathrm{d}\boldsymbol{z} \\
&= -\nabla \cdot \left[ p_t(\boldsymbol{x}|\boldsymbol{s}) \int \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})\frac{p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})q(\boldsymbol{z}|\boldsymbol{s})}{p_t(\boldsymbol{x}|\boldsymbol{s})}\mathrm{d}\boldsymbol{z} \right] \\
&:= -\nabla \cdot [p_t(\boldsymbol{x}|\boldsymbol{s})\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})]
\end{aligned} \tag{39}$$

where we assume that the functions being integrated satisfy the regularity conditions for exchanging integration and differentiation. $\qquad\square$

**Proposition 4.4.** *Assuming that $p_t(\boldsymbol{x}|\boldsymbol{s}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^d$ and $t \in [0, 1]$, then, up to a constant independent of $\boldsymbol{\theta}$, $\mathcal{L}_{SCFM}(\boldsymbol{\theta})$ and $\mathcal{L}_{SFM}(\boldsymbol{\theta})$ are equal. Hence, $\nabla_{\boldsymbol{\theta}}\mathcal{L}_{SCFM}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\mathcal{L}_{SFM}(\boldsymbol{\theta})$.*

*Proof of Proposition 4.4.* The proof is adapted from Lipman et al. (2022); Tong et al. (2023a;b).

To ensure the existence of all integrals and to allow the changing of integration order (by Fubini's Theorem), we assume that $q(\boldsymbol{x}|\boldsymbol{s})$, $p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})$ are decreasing to zero at sufficient speed as $\|\boldsymbol{x}\| \to \infty$ and that $\boldsymbol{u}_t, \boldsymbol{v}_t, \nabla_{\boldsymbol{\theta}}\boldsymbol{v}_t$ are bounded. Since $t$ and $\boldsymbol{s}$

are sampled from $\mathcal{U}(0,1)$ and $q^\circ(s)$, respectively, where both are independent of $\boldsymbol{\theta}$, in the following $t$ and $s$ are both fixed. By the bilinearity of the Euclidean norm and since $\boldsymbol{u}_t$ is independent of $\boldsymbol{\theta}$, we have

$$
\begin{aligned}
&\nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_t(\boldsymbol{x}|\boldsymbol{s})} \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})\|^2 \\
=& \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_t(\boldsymbol{x}|\boldsymbol{s})} \left( \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 - 2 \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s}) \rangle + \|\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s})\|^2 \right) \\
=& \nabla_{\boldsymbol{\theta}} \mathbb{E}_{p_t(\boldsymbol{x}|\boldsymbol{s})} \left( \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 - 2 \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s}) \rangle \right),
\end{aligned}
\tag{40}
$$

and

$$
\begin{aligned}
&\nabla_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})} \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})\|^2 \\
=& \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})} \left( \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 - 2 \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) \rangle + \|\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})\|^2 \right) \\
=& \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})} \left( \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 - 2 \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) \rangle \right).
\end{aligned}
\tag{41}
$$

Next,

$$
\begin{aligned}
&\mathbb{E}_{p_t(\boldsymbol{x}|\boldsymbol{s})} \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 \\
&= \int \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 p_t(\boldsymbol{x}|\boldsymbol{s}) \mathrm{d}\boldsymbol{x} \\
&= \iint \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2 p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) q(\boldsymbol{z}|\boldsymbol{s}) \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})} \|\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})\|^2.
\end{aligned}
\tag{42}
$$

Finally,

$$
\begin{aligned}
\mathbb{E}_{p_t(\boldsymbol{x}|\boldsymbol{s})} \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s}) \rangle &= \int \left\langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \frac{\int \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) q(\boldsymbol{z}|\boldsymbol{s}) \mathrm{d}\boldsymbol{z}}{p_t(\boldsymbol{x}|\boldsymbol{s})} \right\rangle p_t(\boldsymbol{x}|\boldsymbol{s}) \mathrm{d}\boldsymbol{x} \\
&= \int \left\langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \int \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) q(\boldsymbol{z}|\boldsymbol{s}) \mathrm{d}\boldsymbol{z} \right\rangle \mathrm{d}\boldsymbol{x} \\
&= \iint \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) \rangle p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) q(\boldsymbol{z}|\boldsymbol{s}) \mathrm{d}\boldsymbol{z} \mathrm{d}\boldsymbol{x} \\
&= \mathbb{E}_{q(\boldsymbol{z}|\boldsymbol{s}),p_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s})} \langle \boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s}), \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z},\boldsymbol{s}) \rangle.
\end{aligned}
\tag{43}
$$

Therefore, Eq. (40) is always equal to Eq. (41) for any $\boldsymbol{s}$ and $\boldsymbol{z}$, implying that $\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{SCFM}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{SFM}}(\boldsymbol{\theta})$. The proof is complete. $\square$

**Proposition 4.6.** *Consider the optimal coupling $q^*(\boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{s})$ and the vector field $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s})$ as defined in Eq. (22), then the optimal vector field $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$ in Eq. (20) solves the dynamic optimal transport problem (13) (condition on $\boldsymbol{s}$) between $q_0(\boldsymbol{x_0}|\boldsymbol{s})$ and $q_1(\boldsymbol{x_1}|\boldsymbol{s})$.*

*Proof of Proposition 4.6.* Here, we assume that given the switching signal $\boldsymbol{s}$, the source and target distributions condition on $\boldsymbol{s}$ satisfy the regularity conditions such that by Brenier's theorem (Brenier, 1991), there is a unique optimal Monge coupling between $q_0(\boldsymbol{x_0}|\boldsymbol{s})$ and $q_1(\boldsymbol{x_1}|\boldsymbol{s})$.

Under the optimal coupling $q^*(\boldsymbol{z}|\boldsymbol{s}) := q^*(\boldsymbol{x}_0, \boldsymbol{x}_1|\boldsymbol{s})$, it induces a unique optimal Monge transport map $T(\cdot|\boldsymbol{s})$, which can be represented by the gradient of some convex function $\Phi(\cdot|\boldsymbol{s})$, i.e.,

$$
\boldsymbol{x}_1 = T(\boldsymbol{x}_0|\boldsymbol{s}) = \nabla\Phi(\boldsymbol{x}_0|\boldsymbol{s}),
\tag{44}
$$

where $\boldsymbol{x}_0 \sim q_0(\boldsymbol{x_0}|\boldsymbol{s})$ and $\boldsymbol{x}_1 \sim q_1(\boldsymbol{x_1}|\boldsymbol{s})$. In addition, we can construct the conditional probability path or equivalently the flow map as:

$$
\phi_t(\boldsymbol{x}_0|\boldsymbol{s}) = \boldsymbol{x}_0 + t[T(\boldsymbol{x}_0|\boldsymbol{s}) - \boldsymbol{x}_0],
\tag{45}
$$

with the associated vector field:

$$
\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{s}) = T(\boldsymbol{x}_0|\boldsymbol{s}) - \boldsymbol{x}_0.
\tag{46}
$$

Therefore, the optimal vector field $\boldsymbol{v}_t(\boldsymbol{x};\boldsymbol{\theta}|\boldsymbol{s})$ in Eq. (20) (i.e., equal to the above Eq. (46)), solves the dynamic optimal transport problem (13) (condition on $\boldsymbol{s}$) between $q_0(\boldsymbol{x_0}|\boldsymbol{s})$ and $q_1(\boldsymbol{x_1}|\boldsymbol{s})$. $\square$

**Proposition 4.8** (Extremal solutions (Peyré & Cuturi, 2019))**.** *$P^*$ cannot have more than $K_0 + K_1 - 1$ nonzero entries, i.e., $|\{(y_0^{(i)}, y_1^{(j)})|P^*(i,j) > 0\}| \le K_0 + K_1 - 1$.*
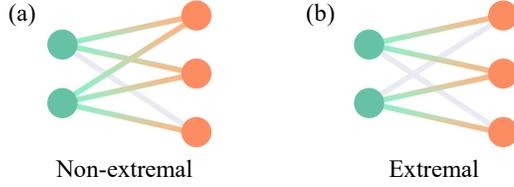
Figure 8. Illustration of the extremal property in Proposition 4.8.

*Proof of Proposition 4.8.* See Proposition 3.4 in (Peyré & Cuturi, 2019). The extremal property is illustrated in Fig. 8. □
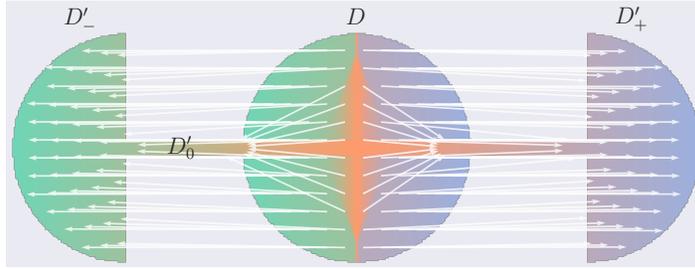


Figure 9. Illustration of the Caffarelli's counterexample.

# C. Additional results

**Regularity.** Here, we present a famous result from the study of optimal transport regularity. Specifically, Caffarelli (1992) demonstrated that the optimal transport map can be discontinuous when the target measure is supported on a non-convex domain. Therefore, the ideal flow map $\varphi_1[\boldsymbol{x}(0); \boldsymbol{\theta}]$ induced by the optimal transport map is discontinuous as well. This is formalized in the following result.

**Proposition C.1** (Caffarelli's counterexample). *Suppose the source and target distributions $q_0$ and $q_1$ are defined on 2-d disc $D$ and dumbbell $D'$ (see Fig. 9, and both normalized to be a probability measure), respectively. Consider the (dynamic) optimal transport problem as defined in Eq. (12) (or Eq. (13)). If the NODE (3) exactly solves the problem, then the flow map $\varphi_1[\boldsymbol{x}(0); \boldsymbol{\theta}]$ is discontinuous, as shown in Fig. 9.*

*Proof.* For clarity and completeness, we mainly provide the idea of its proof, and more detailed proof can be found in (Villani, 2009).

Here, the cost function $c(x_0, x_1)$ is the Euclidean distance. Consider the upper regions of the ball $D$, the left half-ball $D'_-$, and the right half-ball $D'_+$, respectively. Then, a large fraction (say 0.99) of the mass in $D$ has to go to $D'_-$ (if it lies on the left) or to $D'_+$ (if it lies on the right). Due to the homomorphism of the flow map, it should preserve the topology of the source support, i.e., the connectedness. Therefore, the map should transport the mass of a small subset in $D$ into the tube $D'_0$ connecting the left half-ball $D'_-$ and the right half-ball $D'_+$. Moreover, there is some point $\boldsymbol{x}_0 \in D$ such that the corresponding target point $\boldsymbol{x}_1$, obtained by the transport map, is close to the left end of the tube $D'_0$.

In particular, without loss of generality, we assume that $\boldsymbol{x}_1 - \boldsymbol{x}_0$ has a large downward component. From the convergence in probability, many of the neighbors $\boldsymbol{x}'_0$ of $\boldsymbol{x}_0$ have to be transported to, say, $D'_-$, with nearly horizontal displacements $\boldsymbol{x}'_1 - \boldsymbol{x}'_0$. If such an $\boldsymbol{x}'_0$ is picked below $\boldsymbol{x}_0$, we shall have,

$$\langle \boldsymbol{x}_0 - \boldsymbol{x}'_0, \boldsymbol{x}_1 - \boldsymbol{x}'_1 \rangle < 0, \tag{47}$$

or equivalently,

$$|\boldsymbol{x}_0 - \boldsymbol{x}_1|^2 + |\boldsymbol{x}'_0 - \boldsymbol{x}'_1|^2 > |\boldsymbol{x}_0 - \boldsymbol{x}'_1|^2 + |\boldsymbol{x}'_0 - \boldsymbol{x}_1|^2. \tag{48}$$

If the flow map $\boldsymbol{\varphi}_1[\boldsymbol{x}(0); \boldsymbol{\theta}]$ is continuous, in view of Theorem A.11 this contradicts the $c$-cyclical monotonicity of the optimal coupling. The conclusion is that when the tube $D'_0$ is "thin" enough, the optimal flow map $\boldsymbol{\varphi}_1[\boldsymbol{x}(0); \boldsymbol{\theta}]$ is discontinuous. $\qquad \square$

**Remark C.2.** In the above Caffarelli's counterexample, though the target space (i.e., the thin dumbbell) of $q_1$ (extremely smooth, constant) is connected but not convex, the optimal transport map can be discontinuous as well. Significantly, since the (convex) source space and the (non-convex) target space have the same topology, one can naturally employ an ODE to construct a dynamic transport map (not optimal) while preserving the topology over time.

**Joint clustering.** As shown in the Fig. 2(a), the source and target distributions both have two disconnected supports, while the corresponding optimal transport coupling clearly leads to a joint partition with three joint clusters separated by two singular points. Motivated by this observation, we employ the FM to achieve a joint clustering of the source and target pair data points. Specifically, we first use the well-trained FM model $\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})$ to construct the pair $(\boldsymbol{x}_0, \boldsymbol{x}_1)$ by solving the IVP (3), where $\boldsymbol{x}(0) = \boldsymbol{x}_0 \sim q_0(\boldsymbol{x}_0)$ is the initial state and $\boldsymbol{x}(1) = \boldsymbol{x}_1 \sim p_1(\boldsymbol{x}_1) \approx q_1(\boldsymbol{x}_1)$ is the numerically sampled final state. Then, one can use the classical clustering algorithms to partition the constructed pair datasets into $K$ sets, i.e., $\boldsymbol{X}^{(1)}, ..., \boldsymbol{X}^{(K)}$. Moreover, it partitions the both supports of the source and target distributions into $K$ sets as well, i.e., $K_0 = K_1 = K$, and we, without loss of generality, assume that the set $\boldsymbol{X}_0^{(i)}$ and the set $\boldsymbol{X}_1^{(i)}$ are paired with equal mass $\rho_0^{(i)} = \rho_1^{(i)} = \rho^{(i)}$. Therefore, the coupling matrix $P$ (as defined above) is a diagonal matrix of the explicit form $P = \mathrm{diag}\{\rho^{(1)}, ..., \rho^{(K)}\}$. Here, we consider the ODE couplings induced by I-CFM and OT-CFM, and we call the corresponding SFM the IC-SFM and the OTC-SFM, respectively. Notably, under the ODE couplings, it does not require searching the optimal transport couplings within a data bach for the IC-SFM and the OTC-SFM.

As shown in Fig. 10, different from the transportation way illustrated in Fig. 4, both the IC-SFM and OTC-SFM can address the singularity problem as well and transport each source data point to the target space in a determined way. However, the identification of the singularity in high-dimensional situations, such as image datasets, is a challenging task, which is out of the scope of this work. We therefore leave it as one of our future directions.
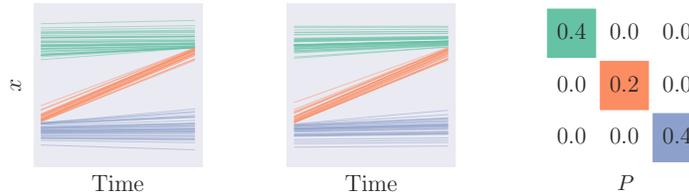


*Figure 10.* Trajectories of the IC-SFM (left) and the OTC-SFM (middle) with 3 joint clusters, and the coupling matrix $P$ (right).

**Gaussian flow.** In the original work of CFM (Tong et al., 2023a;b), they introduce a more general conditional probability path, defined by:

$$
\begin{aligned}
p_t(\boldsymbol{x}|\boldsymbol{z}) &= \mathcal{N}[\boldsymbol{x}|(1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \sigma^2], \\
\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}) &= \boldsymbol{x}_1 - \boldsymbol{x}_0,
\end{aligned} \tag{49}
$$

also referred as to the Gaussian flow, where the pair $\boldsymbol{z} := (\boldsymbol{x}_0, \boldsymbol{x}_1)$ is sampled from the Independent coupling or optimal transport coupling, and $\sigma$ is typically chosen as a small value, serving as a regularization for optimization. Notably, the CFM mentioned in the main text can be regarded as a special case of Eq. (50), i.e., $\sigma = 0$ or equivalently $p_t(\boldsymbol{x}|\boldsymbol{z}) = \delta_{(1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1}$ (the Dirac mass). Naturally, we can generalize our SCFM framework in a similar way, yielding a more general conditional probability path for SCFM:

$$
\begin{aligned}
p_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s}) &= \mathcal{N}[\boldsymbol{x}|(1-t)\boldsymbol{x}_0 + t\boldsymbol{x}_1, \sigma^2], \\
\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{z}, \boldsymbol{s}) &= \boldsymbol{x}_1 - \boldsymbol{x}_0,
\end{aligned} \tag{50}
$$

where the pair $\boldsymbol{z} := (\boldsymbol{x}_0, \boldsymbol{x}_1)$ is now sampled from the Independent coupling or optimal transport coupling condition on the switching signal $\boldsymbol{s}$, i.e., $\boldsymbol{z} \sim q(\boldsymbol{z}|\boldsymbol{s})$. Notably, we set $\sigma = 0$ for all experiments and leave the impact of different values of $\sigma$ on the experimental results for future study, which is out of scope in this work.

**Switching coupling.** For easy implementation of the switching coupling, we make use of the batch data to construct an implicit switching signal distribution. Specifically, given batch data, we employ the following construction:

1. Sample $\left\{ x_0^{(k)} \right\}_{k=1}^{m} \sim q_0(x_0)$ and $\left\{ x_1^{(k)} \right\}_{k=1}^{m} \sim q_1(x_1)$ with the corresponding label sets $\left\{ y_0^{(k)} \right\}_{k=1}^{m}$ and $\left\{ y_1^{(k)} \right\}_{k=1}^{m}$, respectively.

2. Construct the batch coupling matrix $P^m \in \mathbb{R}^{m \times m}$ using the batch label data as follows:

$$P^m(i,j) = \frac{P\left( y_0^{(i)}, y_1^{(j)} \right)}{\mathrm{Count}\left( y_0^{(i)}, y_1^{(j)} \right)}, \tag{51}$$

   where $P \in \mathbb{R}^{K_0 \times K_1}$ is a predefined coupling matrix (see Eq. (24)), and $\mathrm{Count}\left( y_0^{(i)}, y_1^{(j)} \right)$ represents the number of the pair $\left( y_0^{(i)}, y_1^{(j)} \right)$ within the batch data,

3. Sample the switching signal $s := \left( y_0^{(i)}, y_1^{(j)} \right)$ based on the coupling matrix $P^m$, and obtain the corresponding data pair $z^{(i,j)} := \left( x_0^{(i)}, x_1^{(j)} \right)$, resulting in $n$ samples, defined by:

$$\left\{ \left( \hat{x}_0^{(k)}, \hat{x}_1^{(k)}, \hat{y}_0^{(k)}, \hat{y}_1^{(k)} \right) \right\}_{k=1}^{n}, \tag{52}$$

   where $n$ need not match the batch size $m$, but for simplicity in our experiments we choose $n = m$,

4. (Optional for OT-SFM) Construct the joint distribution $\pi_{\mathrm{batch}}(z|s)$ induced by the optimal coupling in terms of the Euclidean distance between data points of the pair $\hat{z}^{(i,j)} := \left( \hat{x}_0^{(i)}, \hat{x}_1^{(j)} \right)$ within the same switching signal data pair set $\left\{ (z, s) | s = \left( \hat{y}_0^{(i)}, \hat{y}_1^{(j)} \right) \right\}$,

5. (Optional for OT-SFM) Sample from the joint distribution $\pi_{\mathrm{batch}}(z|s)$, yielding the samples

$$\left\{ \left( \bar{x}_0^{(k)}, \bar{x}_1^{(k)}, \bar{y}_0^{(k)}, \bar{y}_1^{(k)} \right) \right\}_{k=1}^{l}, \tag{53}$$

   where $l$ need not match the batch size $m$ or $n$, but for simplicity in our experiments we choose $l = n = m$.

**Proposition C.3.** *The joint distribution $q^m(s) := q^m(y_0, y_1)$ induced by the batch coupling matrix $P^m$ constructed in the above Steps [1-3] has the same distribution $q^\circ(s) := q^\circ(y_0, y_1)$ induced by the coupling matrix $P$, i.e., $q^m(y_0, y_1) = q^\circ(y_0, y_1) = P(y_0, y_1)$.*

*Proof.* For an arbitrary test function $f(y_0, y_1)$, by construction of the batch coupling matrix $P^m$ (see Eq. (51)), it holds

$$
\begin{aligned}
\mathbb{E}_{q^m(y_0,y_1)} f(y_0, y_1) &= \sum_{y_0, y_1} q^m(y_0, y_1) f(y_0, y_1) \\
&= \sum_{y_0, y_1} \sum_{y_0^{(i)} = y_0, y_1^{(j)} = y_1} \frac{P\left( y_0^{(i)}, y_1^{(j)} \right)}{\mathrm{Count}\left( y_0^{(i)}, y_1^{(j)} \right)} f(y_0, y_1) \\
&= \sum_{y_0, y_1} P(y_0, y_1) f(y_0, y_1) \\
&= \mathbb{E}_{q^\circ(y_0, y_1)} f(y_0, y_1).
\end{aligned}
\tag{54}
$$

The proof is complete. $\qquad\square$

**Algorithms.** For a better understanding of the proposed SFM, we provide the main pseudocode as shown in Algorithm 1 as well as the pseudocode for constructing the switching coupling as shown in Algorithm 2 with the optional module for OT-SFM (see Algorithm 3). For inference, the pseudocode is displayed in Algorithm 4. The code is available at https://github.com/zhuqunxi/switched-flow-matching.

**Additional experimental results.** The additional experimental results are summarized as follows.

---

**Algorithm 1** `Switched Flow Matching`

---

```
# Input:   Data={𝒙₀,𝒙₁,y₀,y₁}, sampled from q₀(𝒙₀) and q₁(𝒙₁)
# Output:  Model 𝒗ₜ(𝒙;𝜽|𝒔) for the vector field given the switching signal 𝒔 = (y₀,y₁)
Initialize Model # Model can be arbitrary neural network structure
Initialize OT # OT=True/False indicates whether OT-SFM is adopted
Initialize P # P is the predefined coupling matrix
for 𝒙₀,𝒙₁,y₀,y₁ in Data:  # Generate batch data
    Optimizer.zero_grad()
    𝒙₀,𝒙₁,y₀,y₁ = Sample_plan(𝒙₀, 𝒙₁, y₀, y₁, P, OT) # Construct switching coupling
    𝒔 = (y₀,y₁) # Switching signal
    t = torch.rand(batchsize)  # Randomly sample t ∈ [0,1]
    Loss = { Model[(1-t)*𝒙₀ + t*𝒙₁, t, 𝒔] - (𝒙₁ - 𝒙₀) }.pow(2).mean()
    Loss.backward()
    Optimizer.step()
return Model
```

---

**Algorithm 2** `Switching Coupling`

---

```
def Sample_plan(𝒙₀, 𝒙₁, y₀, y₁, P, OT) # Construct switching coupling
    # 𝒙₀, 𝒙₁:  shape=(m, dim); y₀, y₁:  shape=(m, )
    m, K₀, K₁ = len(y₀), P.shape[0], P.shape[1]
    Pᵐ(i,j) = P(y₀[i],y₁[j])/Count(y₀[i],y₁[j])   # Construct the batch coupling matrix
    # Sample from batch coupling matrix (the following three lines)
    choices = np.random.choice(m*m, p=Pᵐ.flatten(), size=m)
    index0, index1 = np.divmod(choices, m)
    𝒙₀, 𝒙₁, y₀, y₁ = 𝒙₀[index0], 𝒙₁[index1], y₀[index0], y₁[index1]
    if not OT:  # Return the sampled data for I-SFM
        return 𝒙₀, 𝒙₁, y₀, y₁
    # Sample from πbatch induced by OT-SFM (the following three lines)
    Y_pair = y₀ * K₁ + y₁   # Construct the switching signal
    M_mask = Y_pair[:, None] == Y_pair[None, :]    # Construct the mask matrix for
sampling the data within the same switching signal data pair set
    index0, index1 = Sample_plan_with_mask(𝒙₀, 𝒙₁, M_mask)   # Sample from πbatch
return 𝒙₀[index0], 𝒙₁[index1], y₀[index0], y₁[index1]
```

---

- **Figure 11**: An experimental illustration of the limitations as pointed out in the bulletins 1 & 2 of Proposition 3.1 as well as in the Corollary 3.3. Specifically, when the source data is near the singularity point, the corresponding flow is "out of distribution", i.e., transporting the source data to the target space but out of the target distribution (see the second row of the Fig. 11). Particularly, when the source and target distributions degenerate into the Dirac masses (see the right column of the Fig. 11), the CFM cannot solve the transport problem, since the flow map of the CFM is determined due to the existence and uniqueness theorem. On the contrary, our proposed SFM can address these limitations.

- **Figure 12**: An experimental illustration of the limitations as pointed out in the bulletins 3 of Proposition 3.1. As the gap between the two modes of target distribution increases, data near the singularity will be mapped to far data points. In other words, the flow map's output (or value) can vary rapidly with small changes in the source singular point, signifying a large Lipchitz constant of the flow map or a worse generalization. In addition, this can present challenges in terms of the numerical stability of the ODE solver. On the contrary, our proposed SFM does not have these issues.

- **Table 3**: An experimental illustration of the example of an infinite number of singular points in the Propositioin 3.7. The learned flow maps of the CFM, including the I-CFM and the OT-CFM, cannot solve the transportation problem due to the singularity, while our proposed SFM performs very well.

- **Table 4**: Generated target samples from the trained CFM and the SFM at $t = 1$ for different NFE on transporting the 2-d Gaussian mixture (8 modes) to the checkerboard. It is observed that both the I-CFM and the OT-CFM face significant

---

**Algorithm 3** `Sample from `$\pi_{\texttt{batch}}$

```
def Sample_plan_with_mask(𝒙₀, 𝒙₁, M_mask) # Sample from π_batch
    # 𝒙₀, 𝒙₁:  shape=(m, dim); M_mask:  shape=(m, m)
    # The following lines are adapted from the source code https://github.com/
atong01/conditional-flow-matching/blob/main/torchcfm/optimal_transport.py
    m = len(y₀)
    a, b = pot.unif(m), pot.unif(m)  # Uniform weights for each sample
    D = torch.cdist(x0, x1) ** 2  # Distance matrix
    D = D + (1 - M_mask) * 1e10  # Refined distance matrix by assigning large
values for miss match pairs
    p = pot.emd(a, b, M.detach().cpu().numpy())  # Return the OT matrix
    choices = np.random.choice(m*m, p=p, size=m)  # Sample from the OT matrix
    return np.divmod(choices, m)
```

---

**Algorithm 4** `Inference`

```
# Input:  Model, source Data={𝒙₀,y₀}, and coupling matrix P
# Output:  Generated samples
Samples = []
for 𝒙₀,y₀ in Data:
    y₁ = sampler(y₀, P)   # Sample y₁ based on P and y₀
    𝒔 = (y₀, y₁)  # Sampled switching signal
    𝒙₁ = model.ODE_solver(𝒙₀, 𝒔)
    Samples.append(𝒙₁)   # Generate a target sample via the ODE solver
return Samples
```

---

challenges in effectively learning the transportation flows. The SFM works well when using the adaptive step size ODE solver, but the I-SFM performs much worse than the OT-SFM for small NFEs due to the straightness issue of the learned flows.

- **Table 5**: Generated target samples from the trained CFM and the SFM at $t = 1$ for different training iterations on transporting the 2-d Gaussian mixture (8 modes) to the checkerboard. Notably, the SFM demonstrates improved efficiency in training and faster convergence compared to the CFM, owing to its superior regularity.

- **Figure 13**: True samples from the source distribution Gaussian mixture and the target distribution, i.e., the CIFAR-10 image dataset. Different from the existing works (Lipman et al., 2022; Liu et al., 2022; Pooladian et al., 2023; Tong et al., 2023a;b), to illustrate the efficiency of our proposed SFM under the heterogeneity in both $q_0$ and $q_1$, we here consider the Gaussian mixture with 2 modes as the source distribution instead of the standard Gaussian distribution.

- **Table 6 (resp., Table 7)**: Generated target samples from the trained I-CFM (resp., OT-CFM) and the I-SFM (resp., OT-SFM) at $t = 1$ for different NFE on transporting the Gaussian mixture (2 modes) to the CIFAR-10 image dataset.

# D. Experimental details

It should be noted that our experimental implementations are heavily adapted from the open source code https://github.com/atong01/conditional-flow-matching provided in Tong et al. (2023a;b). All our experiments were conducted on a single 11GB GTX 1080 Ti GPU.

For the synthetic experiments, we provide the detailed setup for different datasets (see Table 8).

For the CIFAR-10 experiments, all methods used in our work were trained with the same setup as reported in Tong et al. (2023a;b), only differences in the source distribution, the choice of probability path, and the switching mechanism for our proposed SFM and its variants. More precisely, we apply the UNet with the following structures and training configurations:

- Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, and no weight decay,

- channels $= 128$,

- depth $= 2$ ,

- channels multiple $= [1, 2, 2, 2]$,

- heads $= 4$,

- heads channels $= 64$,

- attention resolution $= 16$,

- dropout $= 0.1$,

- batch size per gpu $= 128$, gpus $= 1$,

- learning rate $= 2 \times 10^{-4}$,

- gradient clipping with norm $= 1.0$,

- exponential moving average weights with decay $= 0.9999$.

For sampling, we use the traditional Euler integration or the adaptive step size solver `dopri5` from the torchdiffeq package. We use a batch size of 500 for 100 total batches for computing the Fréchet Inception Distance (FID) through the TensorFlow-GAN library `https://github.com/tensorflow/gan`.

Here, we provide the specific setup for the initial distribution of the CIFAR10 experiments, where $x_0$ has a 50% chance of being $x_0 = torch.randn(3, 32, 32)/4 + 0.5$ and a 50% chance of being $x_0 = torch.randn(3, 32, 32)/4 - 0.5$. The coupling matrices $P$ are set as

$$\text{two2one} : \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix},$$

$$\text{one2ten} : \begin{pmatrix} 0.1 & 0.1 & \cdots & 0.1 \end{pmatrix},$$

$$\text{two2ten, mixed} : \begin{pmatrix} 0.05 & 0.05 & \cdots & 0.05 \\ 0.05 & 0.05 & \cdots & 0.05 \end{pmatrix},$$

$$\text{two2ten, extremal} : \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.1 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \end{pmatrix}.$$
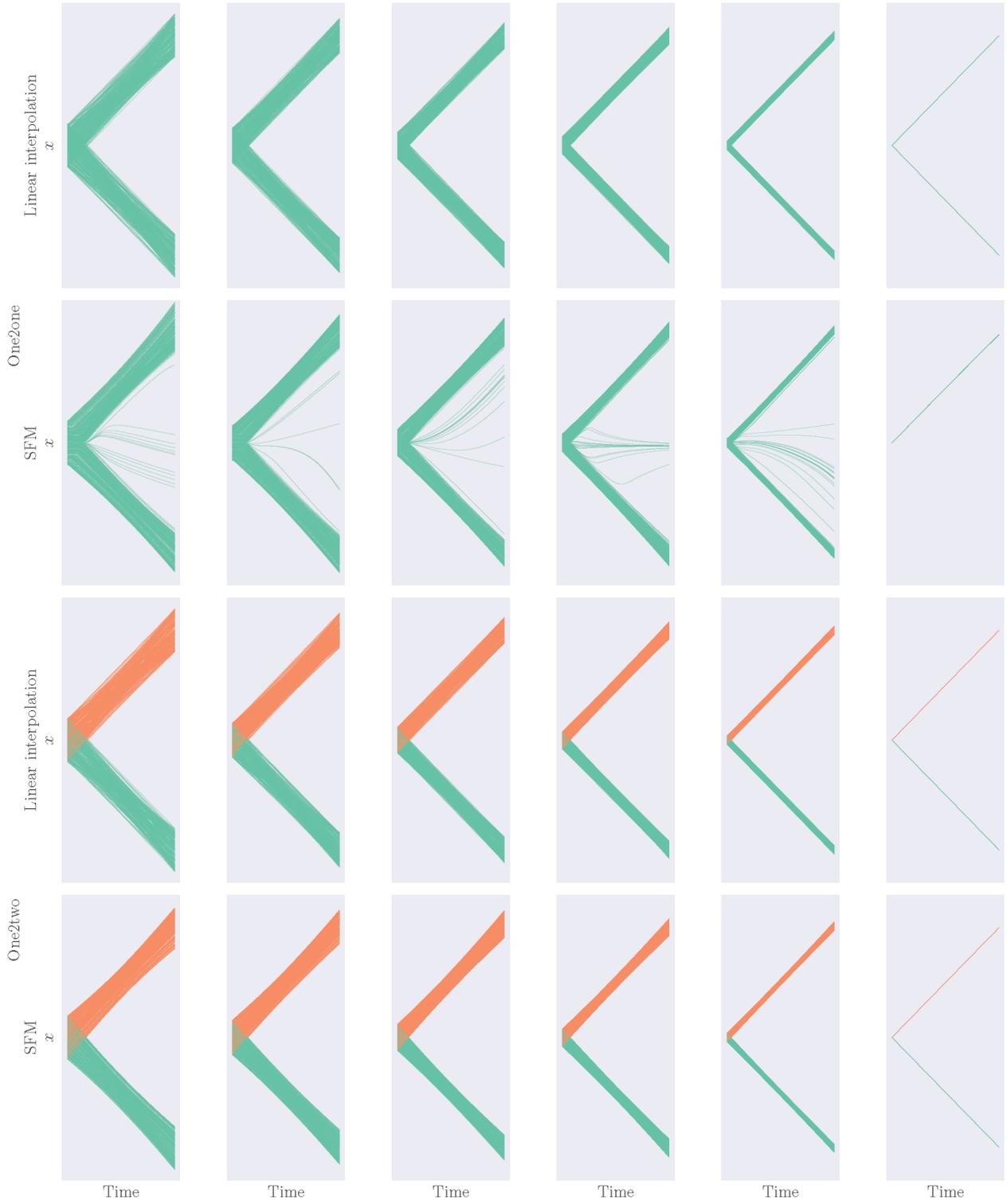
*Figure 11.* Linear interpolation of the source and target data (sampled from the unimodal and bimodal uniform distributions), and the trajectories trained from two kinds of SFM, i.e., one2one (1 switching signal, equivalently CFM) and one2two (2 switching signals), via shrinking the size of the supports (from the left column to the right column). In the right column, the supports of the source and target distributions are both shrunk to a Dirac measure and a 2-mixture Dirac measure, respectively.
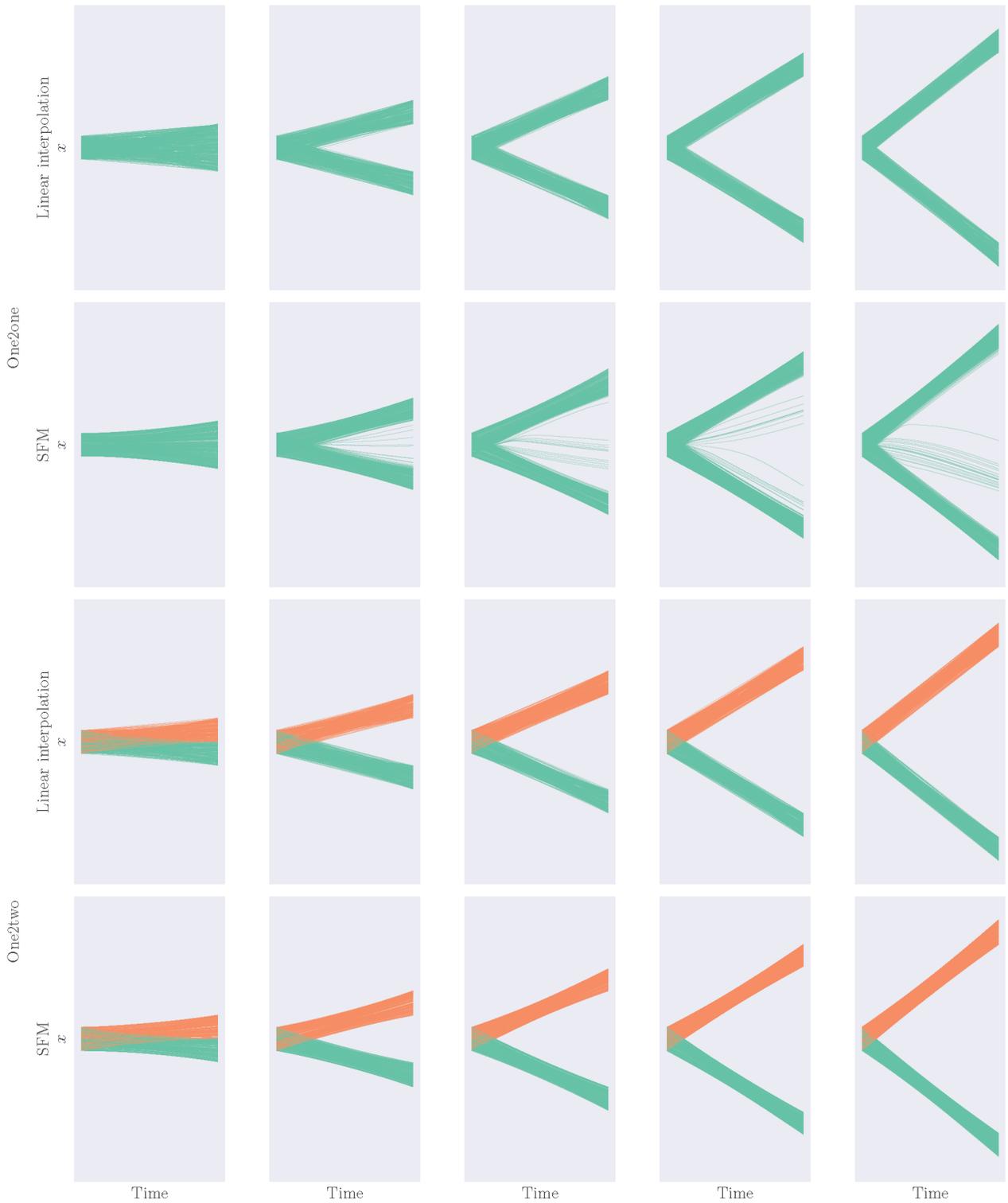
*Figure 12.* Linear interpolation of the source and target data (sampled from the unimodal and bimodal uniform distributions), and the trajectories trained from two kinds of SFM, i.e., one2one (1 switching signal, equivalently CFM) and one2two (2 switching signals), via enlarging the distance of the two modes of the target distribution (from the left column to the right column).

*Table 3.* Trajectories of the learned CFM and SFM via varying the number of function evaluations NFE on the example of infinite number of singular points in the Propositioin 3.7.

*Table 4.* Generated target samples from the trained CFM and the SFM at $t = 1$ for different NFE on transporting the 2-d Gaussian mixture (8 modes) to the checkerboard.

| Method | I-CFM or I-SFM (one2one) | I-SFM (eight2eight) | OT-CFM or OT-SFM (one2one) | OT-SFM (eight2eight) |
|---|---|---|---|---|
| Source data | | | | |
| NFE=1 | | | | |
| NFE=2 | | | | |
| NFE=5 | | | | |
| NFE=Adaptive | | | | |
| Target data | | | | |
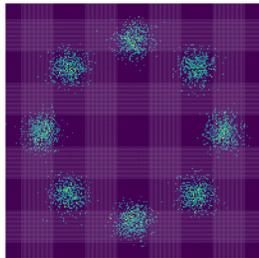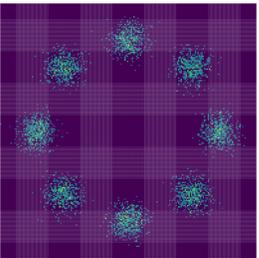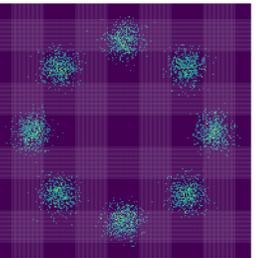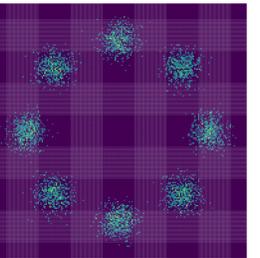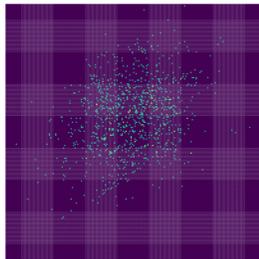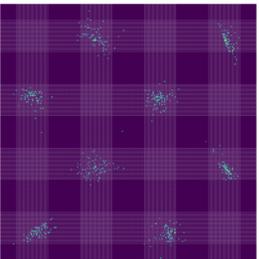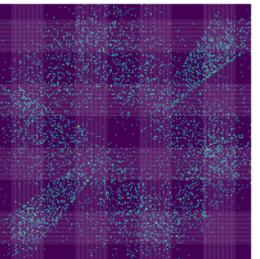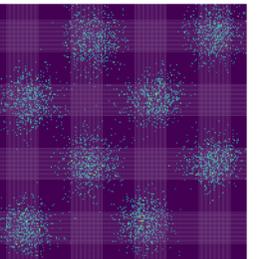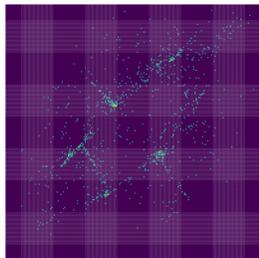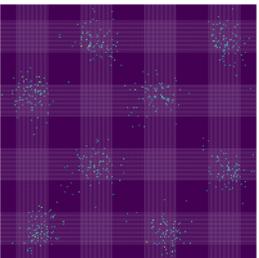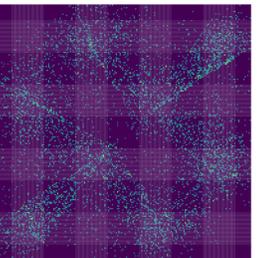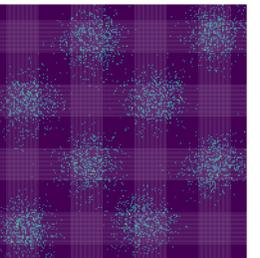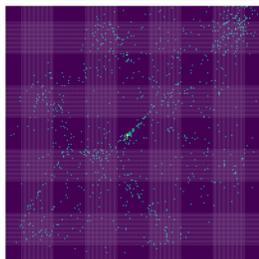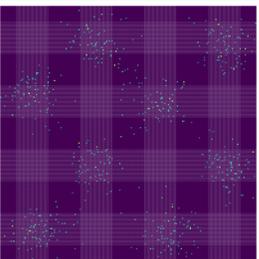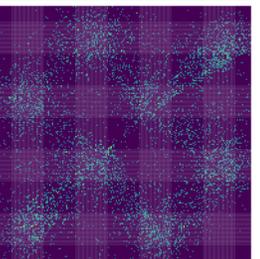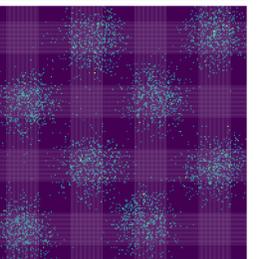
*Table 5.* Generated target samples from the trained CFM and the SFM at $t = 1$ for different training iterations on transporting the 2-d Gaussian mixture (8 modes) to the checkerboard.

| Method | I-CFM or I-SFM (one2one) | I-SFM (eight2eight) | OT-CFM or OT-SFM (one2one) | OT-SFM (eight2eight) |
|---|---|---|---|---|
| Iteration=0.5k | | | | |
| Iteration=1k | | | | |
| Iteration=2.5k | | | | |
| Iteration=5k | | | | |
| Iteration=10k | | | | |
| Iteration=20k | | | | |
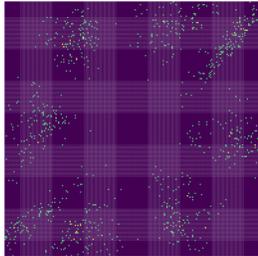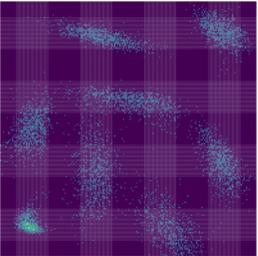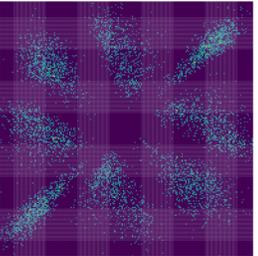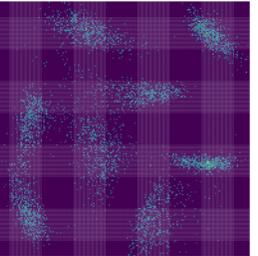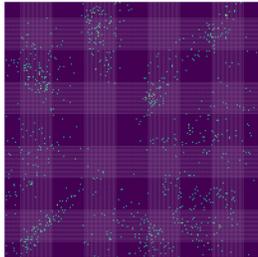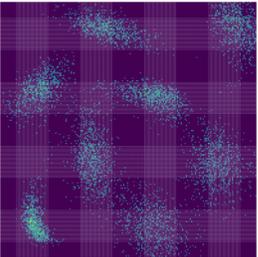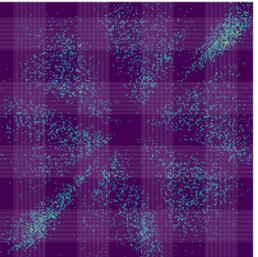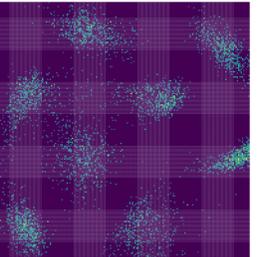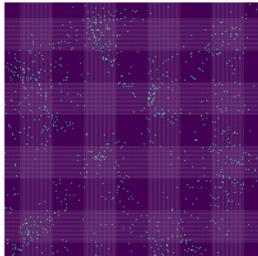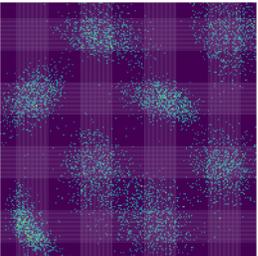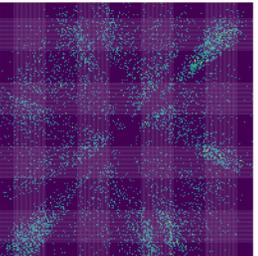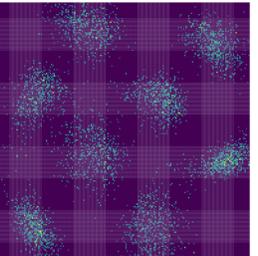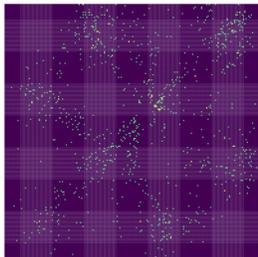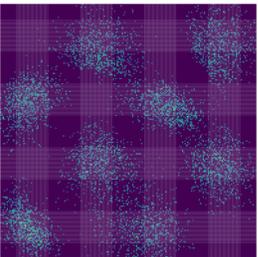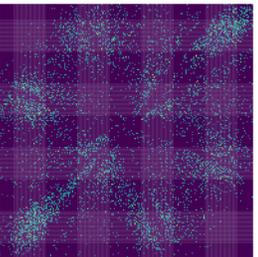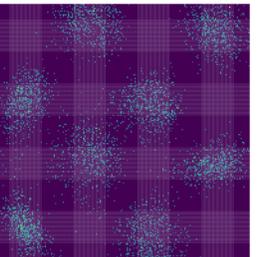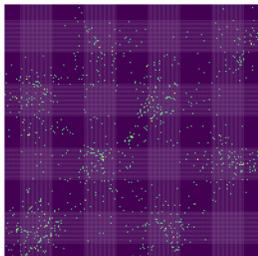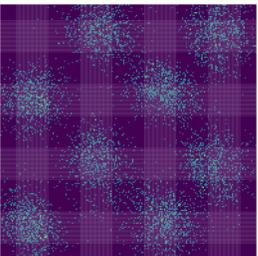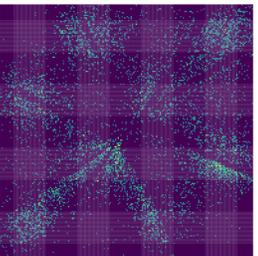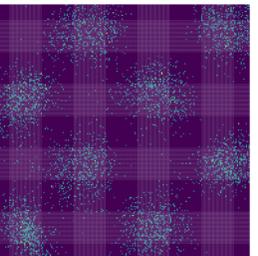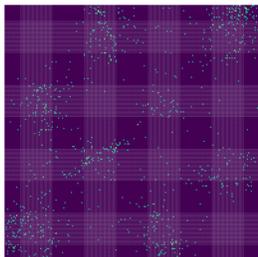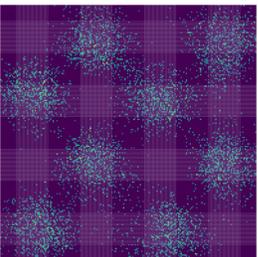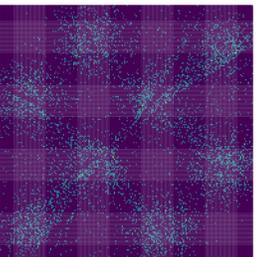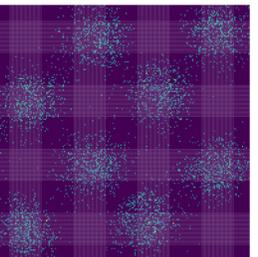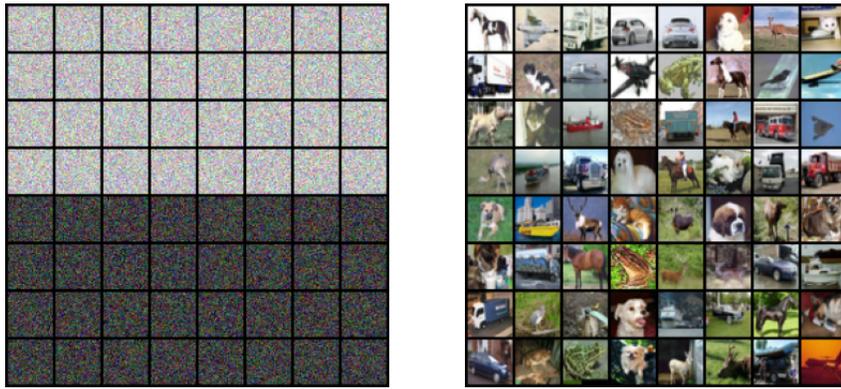
*Figure 13.* True samples from the source distribution (left, Gaussian mixture) and the target distribution (right, CIFAR-10 dataset).

*Table 6.* Generated target samples from the trained I-CFM and the I-SFM at $t = 1$ for different NFE on transporting the Gaussian mixture (2 modes) to the CIFAR-10 image dataset.

| Method | NFE=10 | NFE=20 | NFE=40 | NFE=Adaptive |
|---|---|---|---|---|
| I-SFM (one2one) | | | | |
| I-SFM (one2ten) | | | | |
| I-SFM (two2one) | | | | |
| I-SFM (two2ten, mixed) | | | | |
| I-SFM (two2ten, extremal) | | | | |

*Table 7.* Generated target samples from the trained OT-CFM and the OT-SFM at $t = 1$ for different NFE on transporting the Gaussian mixture (2 modes) to the CIFAR-10 image dataset.

| Method | NFE=10 | NFE=20 | NFE=40 | NFE=Adaptive |
|---|---|---|---|---|
| OT-SFM (one2one) | | | | |
| OT-SFM (one2ten) | | | | |
| OT-SFM (two2one) | | | | |
| OT-SFM (two2ten, mixed) | | | | |
| OT-SFM (two2ten, extremal) | | | | |

*Table 8.* The detailed setup for different synthetic datasets.

| | Setup | Dataset 1 | Dataset 2 | Dataset 3 | Dataset 4 |
|---|---|---|---|---|---|
| Data | Dimension | 1 | 1 | 2 | 2 |
| | $q_0$ | Gaussian mixture (2 modes) | Uniform distribution | $\mathcal{H}^1$ | Gaussian mixture (8 modes) |
| | $q_1$ | Gaussian mixture (2 modes) | Uniform mixture (2 modes) | $(1/2)\mathcal{H}^1$ | Checkerboard (8 squares) |
| Structure | Hidden layer | 2 | 2 | 2 | 2 |
| | Hidden neuron | 64 | 64 | 64 | 64 |
| | Activation | SELU | SELU | SELU | SELU |
| | Time input | True | True | True | True |
| | Switching signal input | True (SFM) / False (CFM) | True (SFM) / False (CFM) | True (SFM) / False (CFM) | True (SFM) / False (CFM) |
| Training | Batch size | 256 | 256 | 256 | 256 |
| | Iteration | $20k$ | $20k$ | $10k$ | $20k$ |
| | Optimizer | Adam | Adam | Adam | Adam |
| | Learning rate | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ |
| Inference | ODE solver | Euler/dopri5 | Euler/dopri5 | Euler/dopri5 | Euler/dopri5 |
| List | Figures or tables | Figs. 4 & 10 | Figs. 11 & 12 | Fig. 5 & Tab. 3 | Tabs. 4 & 5 |