Beckmann Transport Models: From Autonomous Flows to One-Step Maps

Lee Cheuk-Kit[1] **Florentin Coeurdoux**[2] **Peter Potaptchik**[3]
Yilun Du[1] **Michael S. Albergo**[1] **Eric Vanden-Eijnden**[2,4]
[1]Harvard University [2]Capital Fund Management
[3]University of Oxford [4]New York University

Abstract

We propose an instantiation of flow matching that relies on a time-independent velocity field (an *autonomous flow*) to exactly map between two distributions, so long as the target is singular, i.e. supported on a lower-dimensional data manifold. We also show that the one-step generative map associated with this flow is the unique solution of a simple conservation equation, which can be used to learn the map directly from samples. These autonomous flows and maps give a dynamical meaning to the flux constraint of Beckmann's transportation problem. Their construction provides a unifying framework that recovers, for instance, the closed-form Poisson-flow generative model and equilibrium matching with a quadratic flow-matching regression loss. We illustrate how this theory corrects inconsistencies in existing methods and demonstrate the effectiveness of the autonomous flow and the one-step map on ImageNet 256x256.

1 Introduction

Standard flow matching [Lipman et al., 2023, Albergo and Vanden-Eijnden, 2023, Liu et al., 2023] and diffusion models [Ho et al., 2020, Song et al., 2021] drive generation by transporting samples from a base distribution μ_0 to a target distribution μ_1 using a time-dependent velocity field (or *drift*) learned by quadratic regression.

An intriguing alternative is to constrain the drift to be *time-independent*: minimize the same regression loss over fields $b : \mathbb{R}^d \to \mathbb{R}^d$ that do not depend on t and perform generation using the *autonomous* equation $\dot{X}_t = b(X_t)$ solved with initial $X_0 \sim \mu_0$. This is essentially the construction proposed by Wang and Du [2025] under the name *Equilibrium Matching* (EM) and it is appealing because it simplifies both the neural network architectures needed for generative models based on dynamical transport and the dynamical system itself. However, it raises an immediate question: *does the autonomous flow with this time-independent drift actually transport μ_0 to μ_1?* The original EM paper does not prove that it does, and uses a loss whose minimizer turns out not to enforce correct transport.

This paper provides the missing justification, and uses it to derive new results. We show that when μ_1 is supported on a lower-dimensional manifold $M_1 \subset \mathbb{R}^d$, the time-independent flow matching drift defines a valid transport: trajectories of the autonomous flow converge to M_1, and the resulting map pushes μ_0 forward to μ_1. The proof rests on a divergence equation satisfied by the drift, and uncovers a general framework: for example, we recover the closed-form Poisson Flow generative model [Xu et al., 2022] as a different instance of the same principle. This framework also identifies the correct loss for EM (which differs from the loss originally proposed) and yields a simple conservation equation for the one-step transport map T from μ_0 to μ_1. This equation can be used to train T directly from samples and allows us to replace ODE integration with a single forward pass at inference. At a structural level, these autonomous flows and maps give a *dynamic* interpretation of the flux constraint

Preprint.

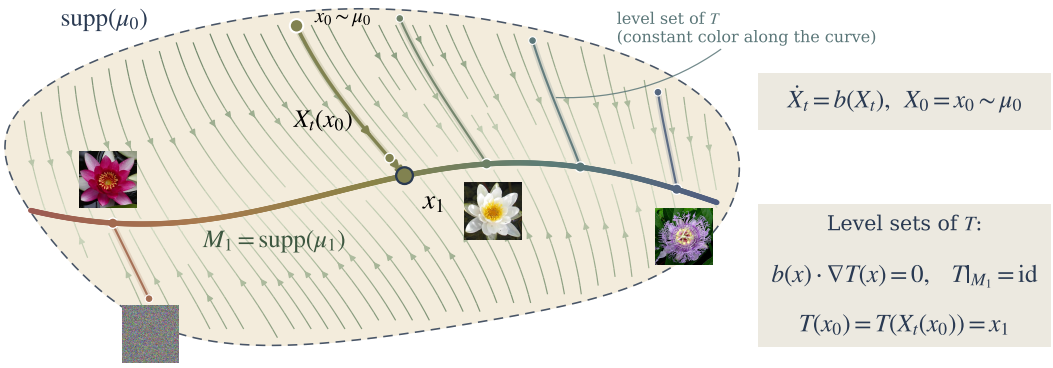


Figure 1: Overview of autonomous flow matching. When μ_1 is supported on a lower-dimensional manifold M_1, the autonomous flow $\dot{X}_t = b(X_t)$ transports every initial X_0 to M_1 at hitting time τ, defining the transport map $T = X_\tau$ that pushes μ_0 forward to μ_1. The map T is constant along the trajectories, $(d/dt)T(X_t) = 0$, implying the conservation equation $b \cdot \nabla T = 0$ that uniquely characterizes T and motivates an objective function to directly learn it.

of Beckmann's transportation problem [Beckmann, 1952, Santambrogio, 2015], a formulation of optimal transport that, until now, lacked a dynamic interpretation analogous to Benamou–Brenier's [Benamou and Brenier, 2000] of Monge–Kantorovich. As a byproduct, we show that the transport cost of the autonomous flow is always bounded by that of the original, time-dependent flow. We refer to constructions of this kind collectively as *Beckmann Transport Models* (BTM).

Summarizing, our **main contributions** are:

- We prove that, if μ_1 has singular support, the time-independent FM drift defines a valid transport: the minimizer of the FM regression loss over time-independent fields b generates a flow $\dot{X}_t = b(X_t)$ that pushes μ_0 to μ_1 (Proposition 1). This recovers Equilibrium Matching [Wang and Du, 2025], up to a correction in the originally-proposed loss.
- More generally, we show that given any positive weight ν, a drift b satisfying the **divergence equation** $\nabla \cdot (\nu b) = \mu_0 - \mu_1$ defines a valid autonomous transport (Theorem 1). The weight ν acts as a genuine design knob: the FM case corresponds to ν being the time-averaged interpolant distribution, and $\nu \equiv 1$ recovers the closed-form Poisson Flow generative model [Xu et al., 2022].
- We provide a **conservation equation** for the one-step map T that fulfills the transport, given by $b \cdot \nabla T = 0$ with boundary condition $T = \mathrm{id}$ on M_1 (Theorem 2).
- We characterize the autonomous flow X_t as the solution of an Eulerian equation $\partial_t X_t = b \cdot \nabla X_t$ (Proposition 2). Discretizing this equation in optimization time yields a residual loss for learning T directly from FM samples (Theorem 3), enabling one-step or iterative inference.

We validate these results on 2D atomic targets and on the image benchmarks of Wang and Du [2025], and show that the corrected autonomous flow associated to equilibrium matching improves performance, and that the map associated to these transports are directly learnable. A detailed discussion of related work (including flow matching and diffusion, OT formulations, and the recent literature on one-step generation) is deferred to Appendix A.

2 Autonomous transports and their associated maps

Standard flow matching [Lipman et al., 2023, Albergo and Vanden-Eijnden, 2023, Liu et al., 2023] constructs a stochastic interpolant[1]

$$I_s = \alpha_s x_0 + \beta_s x_1, \qquad x_0 \sim \mu_0,\ \ x_1 \sim \mu_1,\ \ s \in [0,1], \tag{1}$$

where α_s and β_s satisfy the boundary conditions $\alpha_0 = \beta_1 = 1$, $\alpha_1 = \beta_0 = 0$ so that the law of I_s bridges between the base μ_0 at time $s = 0$ and the target μ_1 at time $s = 1$. The main result of the

[1]We use s for the interpolant time to avoid confusion with the time t in the autonomous-flow introduced later.

framework is that, at each time $s \in [0,1]$, the probability distribution μ_s of I_s coincides with that of the solution to the probability flow ODE

$$\dot{Y}_s = b_s(Y_s), \quad Y_0 \sim \mu_0, \tag{2}$$

provided that the time-dependent drift b_s is the minimizer of the FM loss:

$$b_s(x) = \arg\min_{\hat{b}_s} \mathbb{E}_{s\sim U[0,1],x_0,x_1}\left[|\hat{b}_s(I_s) - \dot{I}_s|^2\right]. \tag{3}$$

In particular, $Y_1 \sim \mu_1$, thereby offering a way to perform inference by integrating (2) on $s \in [0,1]$ after learning b_s by minimizing (3). Importantly, the pair (μ_s, b_s) satisfy the continuity equation

$$\partial_s \mu_s + \nabla \cdot (b_s \mu_s) = 0, \quad \mu_{s=0} = \mu_0, \tag{4}$$

such that $\mu_{s=1} = \mu_1$.

2.1 Autonomous flow matching

Suppose that we constrain the minimization of the FM loss to *time-independent* fields, i.e. define

$$b(x) = \arg\min_{\hat{b}} \mathbb{E}_{s\sim U[0,1],x_0,x_1}\left[|\hat{b}(I_s) - \dot{I}_s|^2\right], \tag{5}$$

and consider the *autonomous* ODE associated with this minimizer:

$$\dot{X}_t(x_0) = b(X_t(x_0)), \quad X_0(x_0) = x_0. \tag{6}$$

This ODE defines a different flow than (2) since it involves a different drift: optimizing the FM loss over time-dependent drift gives $b_s(x) = \mathbb{E}_{x_0,x_1}[\dot{I}_s | I_s = x]$ where the expectation is taken over x_0, x_1 at *fixed* $s \in [0,1]$, conditional on $I_s = x$, whereas optimizing this loss over autonomous drifts gives $b(x) = \mathbb{E}_{x_0,x_1,s}[\dot{I}_s | I_s = x]$ where the expectation is taken over x_0, x_1 *and* $s \sim U([0,1])$, conditional on $I_s = x$. A natural question is therefore whether (6) still transports μ_0 to μ_1. Surprisingly, the answer is yes, provided that we make some structural assumption on the base μ_0 and the target μ_1:

Assumption 1 (Setup)**.** *The base distribution μ_0 is absolutely continuous on $\mathbb{R}^d$ with a positive density ρ_0. The target distribution μ_1 is supported on a closed k-dimensional smooth submanifold $M_1 := \mathrm{supp}(\mu_1) \subset \mathbb{R}^d$ with $0 \le k < d$.*

Note that the assumption on the target can always be satisfied since any μ_1 on $\mathbb{R}^{d_1}$ embeds into $\mathbb{R}^{d_1+d_2}$ by zero-padding, as in [Xu et al., 2022]; this places M_1 on a d_1-dimensional affine subspace of $\mathbb{R}^d$, satisfying the singular-target constraint. Note also that the atomic case $\mu_1 = \sum_{j\in[N]} p_j \delta_{x_j}$ is included as $k = 0$ with $M_1 = \{x_j\}_{j\in[N]}$.

Proposition 1 (Time-independent FM transport)**.** *Under Assumption 1, for μ_0-almost every $x_0 \in \mathbb{R}^d$, the trajectory $X_t(x_0)$ solving the autonomous flow* (6) *converges to a point of M_1. The hitting time*

$$\tau(x_0) := \inf\{t \ge 0 : X_t(x_0) \in M_1\} \in [0, +\infty] \tag{7}$$

is finite if $\dot{I}_1 \neq 0$ (transverse crossing) and infinite if $\dot{I}_1 = 0$ (asymptotic approach). In both cases, the limit

$$T(x_0) := \lim_{t\uparrow\tau(x_0)} X_t(x_0) \in M_1 \tag{8}$$

defines a map $T : \mathbb{R}^d \to M_1$ that pushes μ_0 forward to μ_1, i.e. $T_\sharp \mu_0 = \mu_1$.

Proof sketch. Recall that the probability distribution μ_t of the interpolant satisfies the continuity equation (4) with the time-dependent FM drift b_t. Integrating this equation from $t = 0$ to $t = 1$ gives $\nabla \cdot j = \mu_0 - \mu_1$ where we define the averaged current $j = \int_0^1 b_t \mu_t dt$. A direct calculation shows that the minimizer b of the FM loss (5) over *time-independent* fields satisfies $j = \nu b$ where $\nu = \int_0^1 \mu_t dt > 0$ is the time-averaged interpolant distribution (aka occupation measure). The pair (ν, b) therefore satisfies a divergence condition that encodes μ_0 and μ_1 as source and sink for the current $j = \nu b$ and guarantees that the flow lines of b converge to the target support M_1 of μ_1:

Divergence condition: $$\nabla \cdot (\nu b) = \mu_0 - \mu_1 \tag{9}$$

Hence, the map T defined in (8) sends each $x_0 \in \mathbb{R}^d$ to a point $T(x_0) \in M_1$. Next, we show that $T_\sharp \mu_0 = \mu_1$. To this end, pick any μ_1-measurable set $A \subseteq M_1$, consider its basin of attraction $B_A = T^{-1}(A) \equiv \{x_0 \in \mathbb{R}^d : X_{\tau(x_0)}(x_0) \in A\}$, and integrate (9) over $B_A \setminus A$. By the divergence theorem this gives $\mu_0(B_A) = \mu_1(A)$ since the outer boundary of B_A is a stable manifold of the flow ($\nu b \cdot \hat{n} = 0$) and contributes nothing, while the singular flux of νb across the inner boundary near A recovers $\mu_1(A)$. □

The full details of the proof are in Appendix B. The role of the singularity assumption is now visible from the divergence equation (9) since only a singular μ_1 creates the localized sink that the autonomous flow can converge to.

Remark 1 (The autonomous clock). We stress that the autonomous flow (6) runs on a different clock than the interpolant I_t and the solution of the original, time-dependent flow $\dot{Y}_t = b_t(Y_t)$. In particular, $X_t(x_0)$ reaches M_1 at time $t = \tau(x_0)$, with $\tau(x_0) \neq 1$, and this time is trajectory-dependent. As a result, $\text{law}(X_t) \neq \text{law}(I_t) = \mu_t$ as a function of t; the laws only match at the endpoints, with $X_0 \sim \mu_0$ and $X_\tau \sim \mu_1$.

2.2 Autonomous flow from the divergence equation

The proof of Proposition 1 rests on the fact that $\nu = \int_0^1 \mu_t dt > 0$ and $b(x) = \mathbb{E}_{x_0,x_1,t}[\dot{I}_t | I_t = x]$ satisfy divergence condition (9). Our next result shows that this argument extends to *any* positive weight ν and any drift $b = \nabla\phi$ that satisfy (9).

Theorem 1 (Autonomous Flow)**.** *Let $\nu > 0$ and $b = \nabla\phi$ satisfy the* divergence condition (9) *with μ_0 and μ_1 satisfying Assumption 1 and consider the autonomous ODE* (6) *with this b. Then the solution $X_t(x_0)$ to this ODE inherits the properties listed in Proposition 1 with $\tau(x_0)$ finite if $b(x) \neq 0$ on M_1 and $\tau(x_0) = \infty$ otherwise and, in particular, the associated map $T(x_0) = X_{\tau(x_0)}(x_0)$ pushes μ_0 forward to μ_1, i.e. $T_\sharp \mu_0 = \mu_1$.*

The proof is in Appendix B. The theorem shows that freedom in ν is a genuine design knob, and the FM-derived b is only one principled instance. A second instance, with $\nu \equiv 1$, recovers the training-free Poisson Flow generative model of Xu et al. [2022]: $b = \nabla\phi$ with $\Delta\phi = \mu_0 - \mu_1$ is the Coulomb potential of the signed charge distribution $\mu_0 - \mu_1$. Details in Appendix C.

2.3 Connection to Beckmann's transportation problem

The divergence condition (9) is the flux constraint of the transportation problem of Beckmann [1952] (see also Santambrogio, 2015): minimize the action $\int_{\mathbb{R}^d} |b|^2 \nu \, dx$ over time-independent fields b, subject to $\nabla \cdot (\nu b) = \mu_0 - \mu_1$ with ν a fixed positive weight. Until now, Beckmann's flux constraint has lacked a dynamical realization analogous to Benamou–Brenier's formulation of Monge–Kantorovich's optimal transport problem; BTM provides one. As long as the pair (ν, b) satisfies (9), the autonomous flow $\dot{X}_t = b(X_t)$ pushes μ_0 to μ_1 at the first hitting of M_1 (Proposition 1), and $\int_{\mathbb{R}^d} |b|^2 \nu \, dx$ is the action of this autonomous flow. The structural analogy is:

Static formulation		**Dynamic / flow interpretation**
Monge–Kantorovich optimal transport	$\longleftrightarrow$	Benamou–Brenier (time-dependent flow)
Beckmann transportation	$\longleftrightarrow$	Beckmann Transport Models (autonomous flow)

This positioning has a quantitative consequence. By Benamou–Brenier applied to the autonomous flow and Jensen's inequality applied to the FM time-averaging, we show in Appendix D that

$$W_2^2(\mu_0, \mu_1) \;\leq\; \int_{\mathbb{R}^d} |b|^2 \nu \;\leq\; \int_0^1 \int_{\mathbb{R}^d} |b_t|^2 \mu_t \, dt, \tag{10}$$

so the autonomous Beckmann action of b upper-bounds W_2^2 and is itself bounded by the FM Benamou–Brenier action.

2.4 The transport map

Theorem 1 defines the transport map $T(x_0) = X_{\tau(x_0)}(x_0)$ associated with the autonomous flow (6), and shows that this map can be used to generate samples from μ_1. One way to sample from T is to integrate the ODE directly: given $x_0 \sim \mu_0$, evolve the flow with drift b until the trajectory reaches M_1 at time $\tau(x_0)$, and then set $T(x_0) = X_{\tau(x_0)}(x_0)$. A more attractive option, which we develop in this section, is to learn the map T directly and thereby bypass ODE integration at inference time. The key observation is that the value of the endpoint T does not change as a point moves along the flow generated by b, which leads to the following characterization.

Theorem 2 (Conservation equation for the map)**.** *The map T of Theorem 1 is, up to its values on a μ_0-null set, the unique continuous solution of*

$$b(x) \cdot \nabla T(x) = 0 \quad \textit{for } x \notin M_1, \qquad T(x) = x \quad \textit{for } x \in M_1. \tag{11}$$

Proof. Fix $x_0 \in \mathbb{R}^d$ and write $X_t = X_t(x_0)$. For any $t \in [0, \tau(x_0))$, restarting the autonomous flow from X_t simply follows the same trajectory with time shifted by t. Hence, the endpoint is unchanged, so $T(X_t(x_0)) = T(x_0)$. Differentiating this identity at $t = 0$ gives $b(x_0) \cdot \nabla T(x_0) = 0$, which is the PDE in (11). The boundary condition follows from $\tau(x_0) = 0$ for $x_0 \in M_1$, so $T(x_0) = x_0$ on M_1. For uniqueness, let $\tilde{T}$ be any continuous solution of (11). The chain rule gives $\frac{d}{dt}\tilde{T}(X_t) = b(X_t) \cdot \nabla \tilde{T}(X_t) = 0$, so $\tilde{T}$ is constant along the trajectory X_t. Taking the limit as $t \to \tau(x_0)$ and using the boundary condition on M_1 gives

$$\tilde{T}(x_0) = \lim_{t \to \tau(x_0)} \tilde{T}(X_t) = \tilde{T}(T(x_0)) = T(x_0),$$

which holds for μ_0-a.e. x_0 since basin boundaries are μ_0-null (Proposition 1). □

The two theorems form a complementary pair. Theorem 1 characterizes the drift b via a divergence equation with source $\mu_0 - \mu_1$, while Theorem 2 characterizes the map T via a conservation equation along the flow trajectories of b. The first determines the vector field that drives transport; the second determines the map produced by that field.

2.5 The flow map

The transport map T is the long-time limit of the autonomous flow X_t. In the transverse case $b \neq 0$ on M_1, this limit is reached in finite time at each starting point: $X_t(x_0) = T(x_0)$ for all $t \geq \tau(x_0)$, by Theorem 1. We extend $X_t(x_0)$ to all $t \geq 0$ by *freezing the flow after first hitting* — setting $X_t(x_0) := X_{\tau(x_0)}(x_0)$ for $t \geq \tau(x_0)$. Under this convention, X_t inherits the semigroup property of the autonomous ODE,

$$X_{t+s} = X_t \circ X_s \qquad \text{for all } s, t \geq 0, \tag{12}$$

with $X_0 = \mathrm{id}$, since $X_t \circ X_s$ leaves M_1 fixed once any trajectory has reached it. We now show that X_t also satisfies a transport equation in t.

Proposition 2 (Eulerian form of the flow)**.** *Under Assumption 1, let $\nu > 0$ and b satisfy the divergence condition* (9) *with $b \neq 0$ on M_1, and let $X_t(x_0)$ be the corresponding autonomous flow* (6) *extended by freezing as above. Viewed as a flow map of the initial condition $x \in \mathbb{R}^d$, $X_t(x)$ satisfies the* Eulerian *equation*

$$\partial_t X_t(x) = b(x) \cdot \nabla X_t(x), \quad X_0(x) = x, \quad \textit{for } x \notin M_1, \qquad X_t(x) = x \quad \textit{for } x \in M_1. \tag{13}$$

Proof. For $x \notin M_1$ and $t < \tau(x)$, the semigroup identity (12) gives $X_{t+s}(x) = X_t(X_s(x))$. Differentiating in s at $s = 0$ and using the chain rule yields $\partial_t X_t(x) = \nabla X_t(x) \cdot b(x)$, which is the Eulerian PDE. The boundary condition $X_t(x) = x$ for $x \in M_1$ is the freezing convention. □

By Theorem 1, $X_t \to T$ as $t \to \tau$, so T is the steady state of (13): setting $\partial_t X_t = 0$ recovers the conservation equation $b \cdot \nabla T = 0$ of Theorem 2.

2.6 Direct learning of the transport map

From the conservation equation (11) we could in principle learn T by minimizing the squared residual $\mathbb{E}[|b\cdot\nabla T|^2]$. This requires access to the drift b, which the FM regression provides. We argue, however, that even with b in hand a residual-based loss is suboptimal: it has the right stationary points but no dynamical interpretation, so a partially trained T_θ approximates nothing in particular.

A more principled approach is suggested by Proposition 2: the transport map T is the long-time limit of the autonomous flow X_t, which itself solves the Eulerian equation (13). One option is to learn the full map X_t using existing methods such as consistency or distillation losses (e.g. [Song et al., 2023, Boffi et al., 2025a, Geng et al., 2025]). We propose instead to discard intermediate X_t and learn only their *updates*: starting from $X^{(0)} = \mathrm{id}$, an explicit-Euler step of (13) with time step $\eta > 0$ reads

$$X^{(k+1)}(x) = X^{(k)}(x) + \eta\, b(x)\cdot\nabla X^{(k)}(x). \tag{14}$$

Reading (14) as a regression target, we train a single network T_θ to predict $X^{(k+1)}$ from $X^{(k)}$, treating $X^{(k)}$ as fixed. This is similar in spirit to the Drifting framework of Deng et al. [2026] but without the need of a kernel, leading to more stable training dynamics, see Appendix I for a comparison. The fixed-target convention is implemented by a stop-gradient on the right-hand side of (14); what gets backpropagated is only the predicted next iterate. Iterating until convergence in k, $T_\theta \to T$.

In the FM setup we do not need b explicitly: the regression identity $b(x) = \mathbb{E}_{t,x_0,x_1}[\dot I_t \mid I_t = x]$ allows us to replace $b(I_t)\cdot\nabla T(I_t)$ in (14) by the conditional sample $\dot I_t\cdot\nabla T(I_t)$. Adding a soft Dirichlet penalty for the BC of (13), this yields:

Theorem 3 (Map objective)**.** *Under Assumption 1 and the transverse condition $b \neq 0$ on M_1, the transport map T is the unique stationary point (modulo μ_0-null sets) of the objective*

$$\mathcal{L}(T) = \mathbb{E}_{t,x_0,x_1}\left[\left|T(I_t) - \mathrm{sg}\big(T(I_t) + \dot I_t\cdot\nabla T(I_t)\big)\right|^2\right] + \lambda\,\mathbb{E}_{x_1}[|T(x_1)-x_1|^2]. \tag{15}$$

Proof. At a stationary point, the gradient through the unstopped $T(I_t)$ on the left vanishes, requiring $\mathbb{E}[\dot I_t\cdot\nabla T(I_t)\mid I_t = x] = b(x)\cdot\nabla T(x) = 0$, which is the conservation equation (11). The boundary penalty enforces $T = \mathrm{id}$ on M_1. Uniqueness follows from Theorem 2. □

Iterative convergence of the transport map. Once T is trained, sampling at inference is a single forward pass: $x_1 = T(x_0)$ for $x_0 \sim \mu_0$. If training has not fully converged, the partially trained T approximates the time-t flow X_t for some moderate $t > 0$, and Proposition 2 gives a way to compensate: by the semigroup property, $T^{\circ k} \approx X_{kt}$, which equals T for k large enough that $kt \geq \tau(x_0)$. Iterating the network at inference therefore trades training time for forward passes. This is illustrated in Figure 2, in which we learn the autonomous map for

2.7 Equilibrium Matching as Approximate Self-Stopping BTM

Wang and Du [2025] proposed Equilibrium Matching via a loss that resembles (5) but pairs the linear interpolant $I_t = (1-t)x_0 + tx_1$ with a target $c_t(x_1 - x_0)$ scaled by a non-trivial schedule:

$$\mathbb{E}_{t,x_0,x_1}\left[|b((1-t)x_0 + tx_1) - c_t(x_1 - x_0)|^2\right]. \tag{16}$$

in hope to send the b to 0 as we approach the support $\mathrm{supp}(\mu_1)$. Sampling is then done by following the drift over an indefinite amount of time.

This coincides with the loss in (5) only when $c_t \equiv 1$ (where $\dot I_t = x_1 - x_0$ for the linear interpolant). For other choices of c_t, the loss (16) does not correspond to $\mathbb{E}[|b(I_t) - \dot I_t|^2]$ for any interpolant and as a result the flow defined by $\dot{\tilde X}_t(x_0) = \tilde b(\tilde X_t(x_0))$ converges to M_1 but introduces a bias: $(\tilde X_\tau)_\sharp\mu_0 \neq \mu_1$ (see Section 3).

The resolution, which is what (5) implements, is to treat the schedule as part of the interpolant rather than as a separate weight on the target. To satisfy the desired tail behaviour, one may choose α_t and β_t such that $\dot I_1 = 0$. This results a self-stopping autonomous flow where the drift is 0 on the

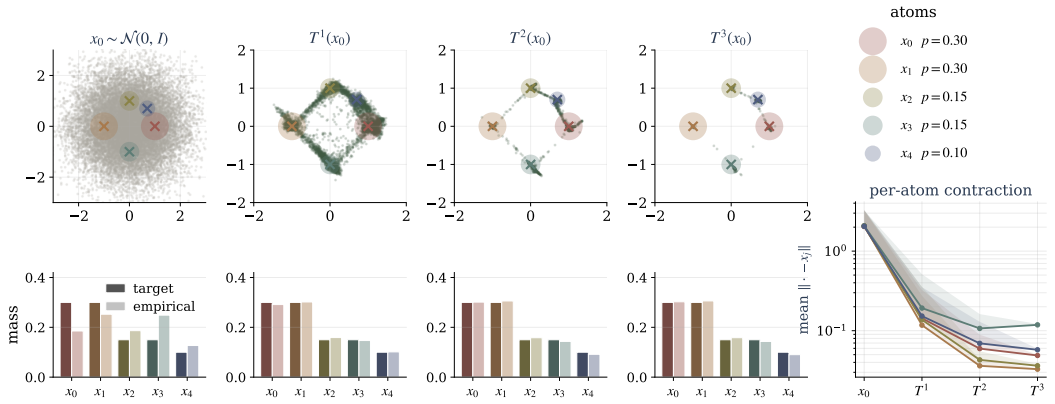


Figure 2: Autonomous transport map learning on 5-mode atomic distribution learned with (15). Iterative convergence of the map to the singular distribution shows that total variation distance of samples on each mode is maintained while clusters continue to converge upon iteration.

support of μ_1. Our experiment in Section 3 explores one such choice and confirms that this correction improves the quality reported by Wang and Du [2025] at the same training cost.

Algorithm 1: Learning the autonomous drift b

Input: Samples $\{x_0^{(n)}\} \sim \mu_0$, $\{x_1^{(m)}\} \sim \mu_1$; interpolant (α_t, β_t); network b_θ; learning rate η
for iteration $= 1, 2, \ldots$ **do**
  Sample a minibatch $\mathcal{B}$;
  for $k \in \mathcal{B}$ **do**
    Sample $t_k \sim U([0,1])$, $x_0^k \sim \mu_0$, $x_1^k \sim \mu_1$; set $I_k = \alpha_{t_k} x_0^k + \beta_{t_k} x_1^k$, $\dot{I}_k = \dot{\alpha}_{t_k} x_0^k + \dot{\beta}_{t_k} x_1^k$;
  $\theta \leftarrow \theta - \eta \nabla_\theta L$, where $L = \dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{k \in \mathcal{B}} \left| b_\theta(I_k) - \dot{I}_k \right|^2$;
return b_θ

Algorithm 2: Learning the transport map T

Input: Samples $\{x_0^{(n)}\} \sim \mu_0$, $\{x_1^{(m)}\} \sim \mu_1$; interpolant (α_t, β_t); network T_θ; boundary weight λ; learning rate η
for iteration $= 1, 2, \ldots$ **do**
  Sample a minibatch $\mathcal{B}$;
  for $k \in \mathcal{B}$ **do**
    Sample $t_k \sim U([0,1])$, $x_0^k \sim \mu_0$, $x_1^k \sim \mu_1$; set $I_k = \alpha_{t_k} x_0^k + \beta_{t_k} x_1^k$, $\dot{I}_k = \dot{\alpha}_{t_k} x_0^k + \dot{\beta}_{t_k} x_1^k$, $\widehat{T}_k = T_\theta(I_k) + \dot{I}_k \cdot \nabla T_\theta(I_k)$;
  $\theta \leftarrow \theta - \eta \nabla_\theta L$, where $L = \dfrac{1}{|\mathcal{B}|} \displaystyle\sum_{k \in \mathcal{B}} \left[\left| T_\theta(I_k) - \mathrm{sg}(\widehat{T}_k) \right|^2 + \lambda \left| T_\theta(x_1^k) - x_1^k \right|^2 \right]$;
return T_θ

3 Experiments

We validate several claims experimentally: (i) the original EqM loss (16) produces biased pushforward weights, the geometry of which is directly visible in the basin structure of the learned drift, and the consistent interpolant (5) provably corrects this bias at no additional cost; (ii) on lower-dimensional targets, the learned map T allows for one-step generation, and improve upon iteration; and (iii) BTM scales to large-scale image generation, matching and modestly improving the quality reported by Wang and Du [2025].

In Appendix J, we further study the dependence of the bias on the schedule exponent, the trade-off between convergence speed and correctness controlled by a tail parameter, and the training-free Coulomb transport of Appendix C. We defer larger scale evaluation of the direct map learning algorithm to future studies, and provide preliminar results in Appendix K.

BTM corrects bias in Equilibrium Matching. We validate the proposed BTM learnt through the consistent loss (5) corrects the bias in the orignal EqM loss (16). We consider a 2d example where $\mu_0 = \mathcal{N}(0, I_2)$, and $\mu_1 = \sum_{j=1}^{5} p_j \, \delta_{x_j}$ with deliberately unequal weights $p = (0.30, 0.30, 0.15, 0.15, 0.10)$, chosen so that weight errors are immediately visible. For BTM, the five basin areas match the target weights to within measurement error ($\mathrm{MAE} = 0.005$). For EqM, the two largest basins expand while the smallest nearly disappears ($\mathrm{MAE} = 0.102$, a $20\times$ error). The mechanism is visible in the separatrices: the EqM drift satisfies $b(x_j) \approx 0$ because $c_t \to 0$ forces the regression target to vanish at $t = 1$, but this does not enforce the divergence condition (9), so the flow reaches M_1 with the wrong weights. The consistent interpolant ties the regression target to $\dot{I}_t$, encoding the source–sink balance $\mu_0 - \mu_1$; the minimizer then satisfies (9) by construction.

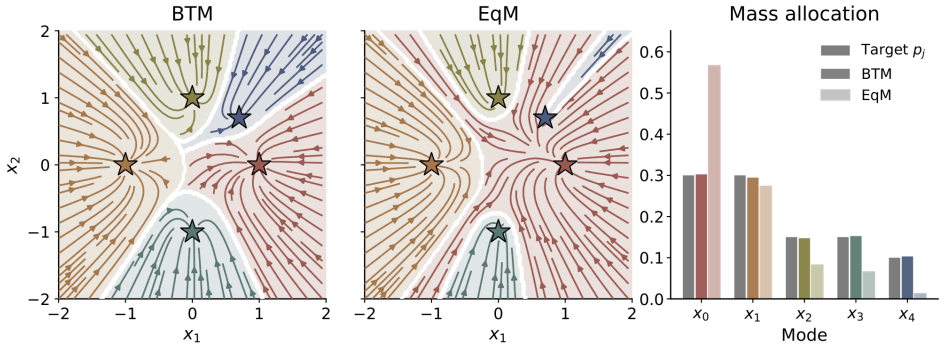


Figure 3: **Basins of attraction** for BTM (left) and EqM (right), schedule $c_t = (1-t)^{0.8}$, $J = 5$ atoms. Each color identifies the set of initial conditions that converge to the corresponding atom; white lines are the separatrices. EqM introduces a bias because of an inconsistency in its loss that BTM corrects at no cost.

Flow vs. map generation (Figure 4). We use the two-dimensional spiral distribution embedded in $d = 6$ dimensions ($\mu_0 = \mathcal{N}(0, I_6)$) using $I_t = (1-t)x_0 + tx_1$ (tranverse case) to compare generation with the autonomous flow associated with b against the one-step map T and its iterates. The autonomous flow recovers the target sharply, while the one-shot map T from a partially trained network is visibly diffuse around the spiral. Iterating the trained map three times, T^3, sharpens generation substantially and brings it close to the flow; a fully trained T matches the flow at one step. This confirms that, by the semigroup property, a partially trained T approximates X_t for some $t < \tau$, so $T^{\circ k} \approx X_{kt}$, and iteration drives trajectories closer to M_1 at the cost of (a few) additional forward passes. Crucially, inference here is one or three forward passes of T, not numerical ODE integration.

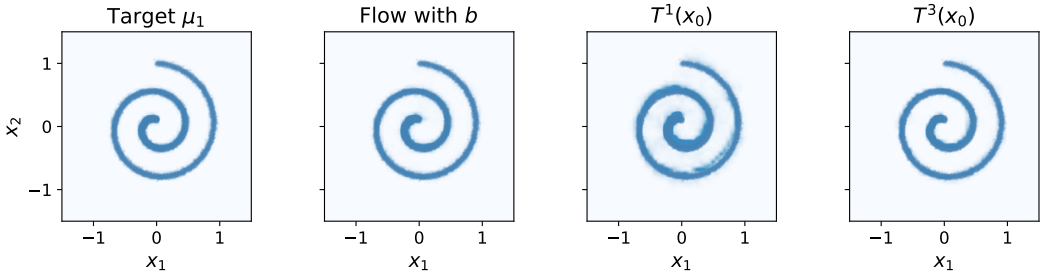


Figure 4: **Flow vs. map generation.** Spiral target embedded in $d = 6$ with $\mu_0 = \mathcal{N}(0, I_6)$. From left: target μ_1; autonomous flow $\dot{X}_t = b(X_t)$; one-step map $T^1 = T$ (1-NFE); three-fold iterate T^3 (3-NFE). Iteration sharpens a partially trained map; a fully trained T_θ matches the flow at 1-NFE.

Model	Method	FID ↓
StyleGAN-XL	GAN	2.30
VDM++	Diffusion	2.12
DiT-XL/2	Diffusion	2.27
SiT-XL/2	FM	2.06
EqM-XL/2	EqM	1.90
EqM-XL/2	BTM	**1.87**

Figure 5: **Left:** Curated class-conditional samples from EqM-XL/2 architecture on ImageNet 256×256. Each image corresponds to a distinct ImageNet class. **Right:** : FID comparison against recent baselines, All numbers use the same evaluation protocol

BTM correction improves EqM image generation for free (Figure 5). We train the consistent loss (5) on class-conditional ImageNet 256×256, using the XL/2 architecture and identical training budget as Wang and Du [2025] with a self-stopping interpolant detailed in Appendix H. The corrected model (EqM-XL/2) achieves FID $= \mathbf{1.87}$ vs. FID $= 1.90$ for the uncorrected EqM-XL/2 under the same conditions. The gain is modest but consistent with the bias being a weight-allocation error that is small when the class distribution is nearly uniform, on the other hand it is free: no change of architecture, training cost, or inference procedure is required. Curated samples are shown in Figure 5.

BTM enables one-step image generation. We train a one-step map using loss (15) with an XL/2 architecture on an SD-VAE latent space, omitting time embeddings. Training is stabilized via adaptive weighting [Geng et al., 2025], and loss terms are balanced dynamically using the final layer's gradient norm (details in Appendix K.2). As shown in Table 1, under the same budget as Wang and Du [2025], BTM achieves a 17.58 FID. While modest, competing one-step methods rely heavily on classifier-free guidance (CFG) for sample quality. Extending CFG to BTM remains an open direction for future work.

Model	NFE	Guidance	FID ↓
iCT [2024]	1	None	30.10
BTM (Ours)	**1**	**None**	**17.58**
Shortcut Models	1	CFG	10.60
SiT (w/o CFG)	250	None	8.30
MeanFlow	1	CFG	3.43
SiT (w/ CFG)	2x250	CFG	2.06

Table 1: FID comparison against recent baselines. All numbers use the same evaluation protocol.

4 Concluding remarks

We have introduced *Beckmann Transport Models*, a framework for generative modeling built on time-independent drifts b satisfying the divergence equation $\nabla \cdot (\nu b) = \mu_0 - \mu_1$. The framework recovers Equilibrium Matching and the Poisson Flow generative model as instances at different choices of ν, and admits a clean characterization of the resulting one-step map T via a conservation equation, which can be discretized to learn T directly from samples without ODE integration at inference. The framework requires the target μ_1 to be supported on a lower-dimensional manifold; this is not a real restriction, since any target on $\mathbb{R}^{d_1}$ can be embedded into $\mathbb{R}^{d_1+d_2}$ by zero-padding.

Open directions. Since any pair (ν, b) satisfying the divergence equation (9) gives a valid transport, one could optimize jointly over (ν, b) subject to $\nabla \cdot (\nu b) = \mu_0 - \mu_1$. This could be done in the Beckmann sense (to minimize the action $\int_{\mathbb{R}^d} |b|^2 \nu \, dx$ and tighten the bound (10)) or under other criteria targeting sample quality, computational cost, or training stability.

The framework is also well-suited to text generation via embedding on the probability simplex: text tokens correspond to vertices of the simplex (atomic distributions), so the singular-target assumption is satisfied natively. The conservation equation then characterizes a one-step map directly to a discrete output, an avenue worth exploring as discrete generative modeling continues to develop.

Finally, the iterative-inference scheme of Section 2.6 suggests a continuous trade-off between training-time and inference-time compute that has not been characterized. A network trained for k steps of effective flow time can be applied $\lceil \tau/k \rceil$ times at inference to recover T; understanding the optimal allocation in this trade-off is an open problem with practical relevance.

Acknowledgments and Disclosure of Funding

We thank Sophia Tang for extensive experimental discussions, and Shiyi Wang for discussion on presentation. PP is supported by the EPSRC CDT in Modern Statistics and Statistical Machine Learning [EP/S023151/1], a Google PhD Fellowship, and an NSERC Postgraduate Scholarship (PGS D). MSA is supported by a Junior Fellowship at the Harvard Society of Fellows as well as the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions). This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Martin Beckmann. A continuous model of transportation. *Econometrica*, 20(4):643–660, 1952.

Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025a. arXiv:2406.07507.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b. arXiv:2505.18825.

Yu Cao and Eric Vanden-Eijnden. Learning optimal flows for non-equilibrium importance sampling. In *Advances in Neural Information Processing Systems*, volume 35, 2022. arXiv:2206.09908.

Mingyang Deng, He Li, Tianhong Li, Yilun Du, and Kaiming He. Generative modeling via drifting. *arXiv preprint arXiv:2602.04770*, 2026.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations (ICLR)*, 2025. Oral; arXiv:2410.12557.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. Oral; arXiv:2505.13447.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers, 2024. URL `https://arxiv.org/abs/2401.08740`.

Grant M. Rotskoff and Eric Vanden-Eijnden. Dynamical computation of the density of states and Bayes factors using nonequilibrium importance sampling. *Physical Review Letters*, 122(15): 150602, 2019. arXiv:1809.11132.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Birkhäuser, 2015.

Yang Song and Prafulla Dhariwal. Improved techniques for training consistency models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=WNzy9bRDvG.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML, 2023.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Yilun Xu, Ziming Liu, Max Tegmark, and Tommi Jaakkola. Poisson flow generative models. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2209.11178.

Yilun Xu, Ziming Liu, Yonglong Tian, Shangyuan Tong, Max Tegmark, and Tommi Jaakkola. PFGM++: Unlocking the potential of physics-inspired generative models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML, 2023. URL https://arxiv.org/abs/2302.04265.

A Related work

Flow matching, diffusion, and equilibrium matching. Flow matching [Lipman et al., 2023, Albergo and Vanden-Eijnden, 2023, Albergo et al., 2023, Liu et al., 2023] and diffusion models [Ho et al., 2020, Song et al., 2021] construct differential equations with time-dependent drifts that transport between base and target distributions. The BTM drift minimizes the same FM loss, but over time-independent fields. BTM provides the mathematical justification for Equilibrium Matching [Wang and Du, 2025], which also uses time-independent drifts, and corrects the loss originally proposed there. It also recovers Poisson Flow [Xu et al., 2022] from the choice $\nu = 1$, with PFGM++ [Xu et al., 2023] softening the resulting Coulomb singularity via an extra dimension parameter — a different design knob from the choice of ν in BTM.

Relationship with optimal transport. The divergence condition is classical in OT [Santambrogio, 2015, Villani, 2009]. Standard flow matching realizes a transport without minimizing the Benamou–Brenier action [Benamou and Brenier, 2000]; analogously, BTM gives a dynamical realization of Beckmann's flux constraint [Beckmann, 1952] without minimizing the associated cost. The variational form of the divergence equation also appears in NEIS [Cao and Vanden-Eijnden, 2022, Rotskoff and Vanden-Eijnden, 2019] for partition-function estimation.

Consistency models, flow maps, and one-step generation. One-step or few-step generation has been pursued through consistency models [Song et al., 2023], rectified flow [Liu et al., 2023], Flow Map Matching [Boffi et al., 2025a,b], MeanFlow [Geng et al., 2025], and Shortcut Models [Frans et al., 2025]. These methods start from an underlying time-dependent flow and distill or self-distill it into a one-step network. Drifting [Deng et al., 2026] learns a one-step map directly, by evolving the network's pushforward during training, using an MMD distance to measure progress. BTM also learns the map directly, but exploits a different structural fact: because the underlying flow is autonomous, the map is the unique solution of a stationary conservation equation, and training reduces to a residual loss with no kernel needed.

B Proof of Proposition 1 and Theorem 1

We begin with Theorem 1 since some steps of its proof are used in the proof of Proposition 1.

Proof of Theorem 1. The proof has two parts: showing that trajectories converge to M_1, and accounting for the mass that lands on each subset of M_1.

(i) Convergence to M_1. Since $b = \nabla\phi$, along any trajectory of $\dot{X}_t = b(X_t)$ we have

$$\frac{d}{dt}\phi(X_t) = \nabla\phi(X_t) \cdot \dot{X}_t = |\nabla\phi(X_t)|^2 \;\geq\; 0,$$

so ϕ is non-decreasing along trajectories. Away from M_1, the divergence equation (9) reads $\nabla \cdot (\nu\nabla\phi) = \rho_0 > 0$, where ρ_0 is the density of μ_0 (Assumption 1). The strong maximum principle for the weighted elliptic operator $u \mapsto \nabla \cdot (\nu\nabla u)$ then forbids ϕ from attaining a local maximum at any interior point of $\mathbb{R}^d \setminus M_1$. Combined with the decay of $\nabla\phi$ at infinity (which excludes escape), this implies that μ_0-almost every trajectory converges to M_1: any limit point would be a local maximum of ϕ along the trajectory, hence must lie on M_1. The hitting time $\tau(x_0)$ is finite when $b \neq 0$ on M_1, since $|\dot{X}_t| = |b(X_t)|$ is then bounded away from zero in a neighborhood of M_1; conversely, if $b = 0$ on M_1 (i.e. M_1 is a critical set of ϕ), the drift vanishes as $X_t \to M_1$ and the trajectory approaches M_1 asymptotically, giving $\tau(x_0) = \infty$.

(ii) Mass accounting via the divergence theorem. For any measurable $A \subset M_1$, define the basin

$$B_A \;:=\; \{x_0 \in \mathbb{R}^d : T(x_0) \in A\} = T^{-1}(A),$$

i.e. the set of initial conditions whose trajectory lands in A. We show that $\mu_0(B_A) = \mu_1(A)$, which gives $T_\sharp\mu_0 = \mu_1$.

Let $T_\delta(A) := \{x \in \mathbb{R}^d : \mathrm{dist}(x, A) < \delta\}$ be the open δ-tube around A, and set $\Omega_\delta := B_A \setminus T_\delta(A)$. By construction, $\Omega_\delta \cap M_1 = \emptyset$, so on Ω_δ the divergence equation reduces to $\nabla \cdot (\nu\, b) = \rho_0$. Integrating over Ω_δ and applying the divergence theorem,

$$\int_{\Omega_\delta} \rho_0 \, dx = \int_{\partial\Omega_\delta} \nu\, b \cdot \hat{n}\, d\sigma = \int_{\partial B_A} \nu\, b \cdot \hat{n}_{\partial B_A}\, d\sigma \;-\; \int_{\partial T_\delta(A)} \nu\, b \cdot \hat{n}_{\partial T_\delta(A)}\, d\sigma,$$

where $\hat{n}_{\partial B_A}$ is the outward normal to B_A, and $\hat{n}_{\partial T_\delta(A)}$ is the outward normal to $T_\delta(A)$ (pointing away from A); the sign on the second term reflects that the outward normal of Ω_δ on its inner boundary points away from the tube, i.e. in the direction $-\hat{n}_{\partial T_\delta(A)}$.

The outer-boundary integral over ∂B_A vanishes: basin boundaries are stable manifolds of the gradient flow and trajectories are tangent to them, so $b \cdot \hat{n}_{\partial B_A} = 0$. For the inner boundary, the flux $\int_{\partial T_\delta(A)} \nu\, b \cdot \hat{n}_{\partial T_\delta(A)}\, d\sigma$ measures the net rate at which mass leaves $T_\delta(A)$; since trajectories enter $T_\delta(A)$ rather than leave it, this integral is negative, and in the limit $\delta \to 0$ it converges to $-\mu_1(A)$ by the singular-source structure of the divergence equation at M_1. The left-hand side converges to $\mu_0(B_A)$, since ρ_0 is the density of μ_0 and $T_\delta(A)$ has vanishing μ_0-mass as $\delta \to 0$. Combining,

$$\mu_0(B_A) = \mu_1(A).$$

Since $(T_\sharp\mu_0)(A) = \mu_0(T^{-1}(A)) = \mu_0(B_A)$, this gives $T_\sharp\mu_0 = \mu_1$. □

Proof of Proposition 1. The strategy is to identify the FM-derived autonomous drift as a particular instance of the divergence-equation framework of Theorem 1, then check the additional structure specific to the FM construction — escape at infinity, and the $\dot{I}_1$ dichotomy for the hitting time.

Step 1: Derivation of the divergence equation. The interpolant law μ_t satisfies the continuity equation

$$\partial_t\mu_t + \nabla \cdot (b_t\mu_t) = 0, \qquad \mu_{t=0} = \mu_0, \quad \mu_{t=1} = \mu_1, \tag{17}$$

where b_t is the time-dependent FM drift. Integrating (17) in t over $[0, 1]$ gives the divergence identity

$$\nabla \cdot j = \mu_0 - \mu_1, \qquad j := \int_0^1 b_t\mu_t \, dt, \tag{18}$$

in the sense of distributions on $\mathbb{R}^d$. For $t \in [0,1)$, μ_t has a density ρ_t (since the interpolant $I_t = \alpha_t x_0 + \beta_t x_1$ is the convex combination of $x_0 \sim \mu_0$, which has a density, and $x_1 \sim \mu_1$, with $\alpha_t > 0$ for $t < 1$). On $\mathbb{R}^d \setminus M_1$ both integrals

$$j(x) = \int_0^1 b_t(x)\rho_t(x)\,dt, \qquad \nu(x) = \int_0^1 \rho_t(x)\,dt \tag{19}$$

are therefore well-defined densities, with $\nu(x) > 0$ pointwise on $\mathbb{R}^d \setminus M_1$ since $\rho_t > 0$ for all $t < 1$.

Step 2: The minimizer of the time-independent FM loss is $b = j/\nu$. The minimizer of the time-independent FM loss (5) is, by orthogonality of the regression,

$$b(x) = \mathbb{E}_{t,x_0,x_1}[\dot{I}_t \mid I_t = x] = \frac{\mathbb{E}_{t,x_0,x_1}[\dot{I}_t\,\delta_{I_t}]}{\mathbb{E}_{t,x_0,x_1}[\delta_{I_t}]}. \tag{20}$$

Both numerator and denominator are formal: the numerator equals $j(x)$ (Bayes' rule applied to the joint density of (t, I_t), then marginalizing t gives $\mathbb{E}[\dot{I}_t \mid I_t = x]\rho(x) = \int_0^1 b_t(x)\rho_t(x)\,dt = j(x)$), and the denominator equals $\nu(x)$ (the marginal density of I_t when $t \sim U[0,1]$). Hence

$$b(x) = \frac{j(x)}{\nu(x)}, \qquad x \in \mathbb{R}^d \setminus M_1. \tag{21}$$

Substituting into (18) gives the divergence condition

$$\nabla \cdot (\nu b) = \mu_0 - \mu_1. \tag{22}$$

Step 3: Applying Theorem 1 via Helmholtz decomposition. The drift $b = j/\nu$ defined by (21) is not in general a gradient. However, it admits a ν-weighted Helmholtz decomposition

$$\nu b = \nu \nabla \phi + q^{\perp}, \qquad \nabla \cdot q^{\perp} = 0, \tag{23}$$

on $\mathbb{R}^d \setminus M_1$, where $\nabla\phi$ is the ν-weighted gradient projection of b and $q^{\perp}$ is the divergence-free remainder. Substituting (23) into the divergence condition (22) gives

$$\nabla \cdot (\nu b) = \nabla \cdot (\nu \nabla \phi) + \nabla \cdot (q^{\perp}) = \nabla \cdot (\nu \nabla \phi) = \mu_0 - \mu_1,$$

so the gradient pair $(\nu, \nabla\phi)$ satisfies the divergence equation on its own. The basin structure of the autonomous flow $\dot{X}_t = b(X_t)$ coincides with that of the gradient flow $\dot{X}_t = \nabla\phi(X_t)$ on μ_0-a.e. initial conditions: the divergence-free part $q^{\perp}/\nu$ generates rotations within changing the mass transport. Theorem 1 applied to the gradient pair $(\nu, \nabla\phi)$ then gives $T_\sharp \mu_0 = \mu_1$, where $T(x_0) = \lim_{t \uparrow \tau(x_0)} X_t(x_0) \in M_1$ is the limit of the autonomous trajectory.

Step 4: The $\dot{I}_1$ dichotomy. On the target manifold, $b(x) = \mathbb{E}_{t,x_0,x_1}[\dot{I}_t \mid I_t = x]$ for $x \in M_1$ reduces to a conditional expectation that concentrates at $t = 1$ (since $I_t \in M_1$ in the limit $t \to 1$), giving $b(x) \propto \mathbb{E}[\dot{I}_1 \mid I_1 = x]$ on M_1. Two cases:

- If $\dot{I}_1 \neq 0$ (e.g., α_t, β_t have nonzero derivatives at $t = 1$, the standard linear-interpolant case), then $b \neq 0$ on M_1. By Theorem 1, $\tau(x_0) < \infty$ for μ_0-a.e. x_0 and the trajectory crosses M_1 transversely.
- If $\dot{I}_1 = 0$ (e.g., $\alpha_t = \cos(\pi t/2)$ or any schedule with $\dot{\alpha}_1 = \dot{\beta}_1 = 0$), then $b = 0$ on M_1, which becomes a critical set of ϕ. The trajectory approaches M_1 asymptotically and $\tau(x_0) = \infty$.

In both cases the limit $T(x_0)$ exists and lies in M_1, so the conclusion $T_\sharp \mu_0 = \mu_1$ from Step 3 holds. □

Explicit kernels for discrete μ_1

For $\mu_0 = \mathcal{N}(0, I_d)$, $\mu_1 = \sum_j p_j \delta_{x_j}$, and a general interpolant $I_t = \alpha_t x_0 + \beta_t x_1$ (with $\alpha_t + \beta_t = 1$ assumed for brevity), the contribution of atom j to ν and j is, with $\epsilon = x - x_j \neq 0$:

$$\nu^{(j)}(x) = \frac{p_j}{(2\pi)^{d/2}} \int_0^1 \alpha_t^{-d} \exp\left(-\frac{|\epsilon + \alpha_t x_j|^2}{2\alpha_t^2}\right) dt, \tag{24}$$

$$j^{(j)}(x) = \frac{p_j}{(2\pi)^{d/2}} \int_0^1 \alpha_t^{-d} \left(\frac{\dot{\alpha}_t}{\alpha_t}(\epsilon + \alpha_t x_j) + \dot{\beta}_t x_j\right) \exp\left(-\frac{|\epsilon + \alpha_t x_j|^2}{2\alpha_t^2}\right) dt. \tag{25}$$

For $x_j = 0$, the integrals reduce via the substitution $u = |\epsilon|/\alpha_t$ (Laplace asymptotics are controlled by $\alpha_t \sim |\epsilon|$) and admit closed forms in terms of incomplete Gamma functions when α_t has a power-law structure; for the specific case $\alpha_t = (1-t)^a$, integrating gives expressions that generalize equation (11) of the original note. For $x_j \neq 0$, the cross term $\epsilon \cdot x_j/\alpha_t$ prevents a closed form for all $|\epsilon|$; the dominant asymptotic behaviour as $|\epsilon| \to 0$ is captured by the exponent $1 - 1/a$.

A notable special case: for a single atom and $\alpha_t = 1 - t$ (linear interpolant, $a = 1$), $\dot{\alpha}_t = -1, \dot{\beta}_t = 1$, and a direct calculation gives $b(x) = x_j - x$ exactly — recovering a linear attractor toward the atom. The regression loss (5) is identically zero at this b provided the parametrization allows it.

C Link with Poisson flows when $\nu = 1$

With $\nu = 1$, equation (9) is the Poisson equation $\Delta\phi = \mu_0 - \mu_1$, and $b = \nabla\phi$ is the electric field generated by μ_1 (as a charge distribution) in the background field of μ_0

$$b(x) = \underbrace{\frac{x}{\omega_d\,|x|^d}\,\frac{\gamma(d/2,\,|x|^2/2)}{\Gamma(d/2)}}_{b_0(x)} \;-\; \underbrace{\frac{1}{\omega_d}\int_M \frac{x-y}{|x-y|^d}\mu_1(dy)}_{b_1(x)}, \tag{26}$$

where $\omega_d = 2\pi^{d/2}/\Gamma(d/2)$.

In the atomic case when $\mu_1 = \sum_{j\in[N]} p_j \delta_{x_j}$, the second term in (26) is also explicit

$$b_1(x) = \frac{1}{\omega_d}\int_M \frac{x-y}{|x-y|^d}\mu_1(dy) = \frac{1}{\omega_d}\sum_{j\in[N]} p_j\,\frac{x-x_j}{|x-x_j|^d}$$

In this case b_1 can be estimated at each ODE step by a fresh i.i.d. mini-batch from μ_1 and no neural network is required. This is the basus of Poisson Flow generative model of Xu et al. [2022].

D Proof of the cost-chain inequality (10)

We establish the chain

$$W_2^2(\mu_0,\mu_1) \;\le\; \int_{\mathbb{R}^d} |b|^2\nu \;\le\; \int_0^1\int_{\mathbb{R}^d} |b_t|^2\mu_t\,dt$$

in two steps. The first inequality is Benamou–Brenier applied to the autonomous flow once we identify ν as its occupation measure. The second is Jensen's inequality applied to the FM time-averaging.

Step 1: ν is the occupation density of the autonomous flow. Let ν_t denote the distribution of the solution to the autonomous flow $\dot{X}_t = b(X_t)$, with $\nu_0 = \mu_0$ and mass progressively absorbed into M_1 as trajectories arrive there. The continuity equation for ν_t reads

$$\partial_t\nu_t + \nabla\cdot(b\nu_t) = 0 \qquad \text{on } \mathbb{R}^d \setminus M_1, \qquad t\in[0,\infty), \tag{27}$$

with mass progressively transferred from the bulk onto M_1. By Proposition 1, the total mass absorbed onto any measurable $A \subset M_1$ equals $\mu_1(A)$. Integrating (27) from $t = 0$ to $t = \infty$ and using $\nu_0 = \mu_0$ together with the absorbed-mass distribution μ_1 on M_1,

$$\nabla\cdot\left(b\int_0^\infty \nu_t\,dt\right) = \mu_0 - \mu_1. \tag{28}$$

The autonomous flow's occupation mesaure $\bar{\nu} := \int_0^{infty} \nu_t(x)\,dt$ therefore satisfies $\nabla\cdot(\bar{\nu}b) = \mu_0 - \mu_1$ on $\mathbb{R}^d \setminus M_1$, the same divergence equation that ν satisfies. Since both $\bar{\nu}$ and ν are positive on $\mathbb{R}^d \setminus M_1$, vanish at infinity, and integrate to give matching total mass against the source-sink structure of $\mu_0 - \mu_1$, they are equal:

$$\nu = \bar{\nu} \qquad \text{on } \mathbb{R}^d \setminus M_1. \tag{29}$$

The FM time-averaged density ν is thus also the occupation measure of the autonomous flow.

Step 2: Benamou–Brenier on the autonomous flow. The autonomous flow with drift b and law ν_t realizes a continuous-time path interpolating μ_0 to μ_1. By the Benamou–Brenier inequality for ODE-driven transport,

$$W_2^2(\mu_0, \mu_1) \;\le\; \int_0^\infty \int_{\mathbb{R}^d} |b|^2 \nu_t \, dt = \int_{\mathbb{R}^d} |b|^2 \bar{\nu} (= \int_{\mathbb{R}^d} |b|^2 \nu, \tag{30}$$

using (29) in the last equality. (Strictly, the BB infimum is over time-rescaled paths on $[0, 1]$; the inequality follows from a standard time-reparametrization of the autonomous trajectory that preserves the integrand.) This establishes the first inequality of the cost chain.

Step 3: Jensen on the FM time-averaging. The FM-derived drift b is the conditional expectation $b(x) = \mathbb{E}_{t,x_0,x_1}[\dot{I}_t \mid I_t = x]$, which can be written as the ratio

$$b(x) = \frac{j(x)}{\nu(x)}, \qquad j(x) = \int_0^1 b_t(x) \rho_t(x) \, dt, \qquad \nu(x) = \int_0^1 \rho_t(x) \, dt. \tag{31}$$

By the Cauchy–Schwarz inequality (a special case of Jensen's inequality applied to the convex function $u \mapsto u^2$ against the time-averaging probability measure $\rho_t(x)\, dt/\nu(x)$ on $[0, 1]$),

$$|b(x)|^2 = \left| \frac{1}{\nu(x)} \int_0^1 b_t(x) \rho_t(x) \, dt \right|^2 \le \frac{1}{\nu(x)} \int_0^1 |b_t(x)|^2 \rho_t(x) \, dt. \tag{32}$$

Multiplying both sides by $\nu(x)$ and integrating over $\mathbb{R}^d$,

$$\int_{\mathbb{R}^d} |b(x)|^2 \nu(x) \, dx \;\le\; \int_{\mathbb{R}^d} \int_0^1 |b_t(x)|^2 \rho_t(x) \, dt \, dx = \int_0^1 \int_{\mathbb{R}^d} |b_t(x)|^2 \mu_t(dx) \, dt, \tag{33}$$

using $\rho_t \, dx = \mu_t(dx)$ in the last identification. This establishes the second inequality of the cost chain.

Combining. (30) and (32) give

$$W_2^2(\mu_0, \mu_1) \;\le\; \int_{\mathbb{R}^d} |b|^2 \nu \;\le\; \int_0^1 \int_{\mathbb{R}^d} |b_t|^2 \mu_t \, dt,$$

which is (10).

E Discontinuity of b and regularity of T at M_1: two independent design questions

The autonomous-flow construction underlying BTM produces two learnable objects: the drift $b(x)$ and the map $T(x_0)$. Each can be learned via the corresponding loss in Section **??**, and each has its own regularity considerations at M_1. We show in this appendix that the regularity properties of b and T at M_1 are governed by *independent* design choices in the construction:

- The discontinuity of b at M_1 is controlled by the choice of weight ν in the divergence equation $\nabla \cdot (\nu b) = \mu_0 - \mu_1$. Adjusting ν to vanish on M_1 (equivalently, $\nu \to \infty$ on M_1) makes $b \to 0$ smoothly at M_1, easing learning of b.
- The regularity of T at M_1 is controlled by the geometry of the current $j = \nu b$, which is independent of ν. Making T smooth (rather than merely continuous) at M_1 requires the geometric current j to be transverse to M_1, which in the FM construction depends on the interpolant family — not on its time parameterization.

The two design questions are separately addressed by separate mechanisms. The schedule-invariance result (Subsection **??**) makes this separation precise: time reparameterizations of a single-schedule interpolant change ν and b but not j or T.

E.1 Constraints on j at M_1

The divergence equation $\nabla \cdot j = \mu_0 - \mu_1$ pins down the normal component of j on M_1 via the divergence theorem: the limiting flux of j into any region of M_1 equals the local density of μ_1. In the codimension-1 case, with μ_1 having smooth density ρ_1 on M_1:

$$j^\perp(x^+) - j^\perp(x^-) = -\rho_1(x), \qquad x \in M_1,$$

where x^+, x^- denote approach from the two sides of M_1. $j^\perp$ **is therefore discontinuous across** M_1 **whenever** $\rho_1 > 0$, with a jump equal to (minus) the density of μ_1 on M_1.

This jump is structural — a consequence of μ_1 being a singular sink for the divergence equation. It cannot be removed by any choice of ν or interpolant. The drift $b = j/\nu$ inherits this jump unless $\nu \to \infty$ on M_1 at a sufficient rate to cancel it.

The behavior of the drift $b = j/\nu$ at M_1 depends on whether ν is bounded or unbounded there:

- ν **bounded on** M_1 **(e.g., PFGM with** $\nu = 1$**).** $b = j/\nu$ has the same jump as j: discontinuous across M_1 with finite normal jump. Trajectories of $\dot{X} = b(X)$ approach M_1 at finite speed and reach it in finite time, but the integrator must detect the crossing and stop — otherwise the ODE will continue past M_1 in the wrong direction. Furthermore, the discontinuity of b at M_1 is a difficult target for neural-network regression: the network must learn a function with a sharp transition on a low-dimensional manifold.
- $\nu \to \infty$ **on** M_1 **(e.g., FM single-schedule with linear interpolant, where** $\nu \sim \ln(1/\mathrm{dist})$**).** $b = j/\nu \to 0$ on M_1, and b is continuous (with value zero on M_1). Trajectories slow down asymptotically as they approach M_1, but no stopping criterion is needed: integrating the ODE for sufficiently long time leaves the trajectory arbitrarily close to M_1, with no risk of overshoot since b vanishes there. Learning b via neural-network regression is also easier because the target is smooth (bounded gradients, no jump on a manifold).

The choice between these cases is purely a function of ν, which the framework treats as a design knob (Section 2.3, equation (9)). Within FM constructions, ν is determined by the interpolant; within other constructions (PFGM, custom ν), it can be chosen independently.

Implication. For both learning and inference, the FM-style construction with $\nu \to \infty$ at M_1 is preferred: it eliminates the discontinuity that would be a barrier to neural-network regression of b, and it removes the need for a stopping criterion in ODE integration. The trade-off is purely numerical: the ODE must be integrated for a longer fixed horizon to bring trajectories close to M_1.

F Analysis of j near M_1 and the regularity of T

F.1 Setup

$\mu_0 = \mathcal{N}(0, I_d)$ centered at the origin, μ_1 supported on a smooth codim-k submanifold $M_1 \subset \mathbb{R}^d$ ($1 \le k \le d-1$ for the manifold case; $k = d$ atomic, treated separately), with smooth tangential density ρ_1 on M_1. For field point $x \in \mathbb{R}^d$ near M_1:

- $y(x) := \arg\min_{z \in M_1} |x - z|$, the foot point on M_1;
- $T_{y(x)}M_1$ (tangent, dim $d-k$) and $N_{y(x)}M_1$ (normal, dim k);
- $h\xi^\perp := x - y(x) \in N_{y(x)}M_1$, with $h = |x - y(x)|$ and $\xi^\perp$ a unit vector in $N_{y(x)}M_1$;
- $y(x)^\parallel \in T_{y(x)}M_1$ the projection of $y(x)$ onto $T_{y(x)}M_1$.

We consider interpolants of the generic form $I_t = \alpha_t x_0 + f(\alpha_t)\, x_1$, where $f : [0,1] \to [0,1]$ with $f(0) = 1$, $f(1) = 0$, $f'(\alpha) \le 0$ encodes the curve geometry, and $\alpha_t : [0,1] \to [0,1]$ with $\alpha_0 = 1$, $\alpha_1 = 0$, $\dot{\alpha}_t < 0$ encodes the time parametrization. The path $\Gamma_{x_0,x_1} = \{I_t : t \in [0,1]\}$ depends on f only.

When $f'(0) = 0$, define the path-curvature parameter

$$c \;:=\; -\tfrac{1}{2} f''(0), \qquad \text{so that } 1 - f(\alpha) \sim c\,\alpha^2 \text{ near } \alpha = 0 \tag{34}$$

(for VP, $f(\alpha) = \sqrt{1-\alpha^2}$, $c = 1/2$).

F.2 Invariance: j is geometric, ν and b are kinematic

Resolving the δ in the line integral $j(x) = \mathbb{E}[\int \delta(x - I_t)\dot{I}_t\, dt]$ via $x_0 = (x - f(\alpha_t)x_1)/\alpha_t$ (Jacobian α_t^{-d}) and $\dot{I}_t = \dot{\alpha}_t[x_0 + f'(\alpha_t)x_1]$, the time-derivative $\dot{\alpha}_t$ exactly cancels the $dt = d\alpha/\dot{\alpha}_t$ Jacobian. Setting $\alpha = \alpha_t$:

$$j(x) \;=\; \int_0^1 d\alpha \int dx_1\, \frac{1}{\alpha^d}\, \rho_0\!\left(\frac{x - f(\alpha)x_1}{\alpha}\right) \rho_1(x_1)\, \left[\frac{x - f(\alpha)x_1}{\alpha} + f'(\alpha)\, x_1\right]. \tag{35}$$

This is manifestly independent of the time parametrization $\alpha_t(\cdot)$ and depends on f only.

By contrast, ν and the time-pointwise ρ_t, b_t depend on the time parametrization. Direct change of variable gives

$$\nu(x) \;=\; \int_0^1 \rho_t(x)\, dt \;=\; \int_0^1 \frac{\rho_\alpha(x)}{|\dot{\alpha}_t(\alpha)|}\, d\alpha, \tag{36}$$

where ρ_α is the marginal density relabeled by α. The factor $1/|\dot{\alpha}_t|$ is the kinematic content. Similarly the average drift $\bar{b} := j/\nu$ inherits the ν-dependence.

G Analysis of j near M_1 and the regularity of T

G.1 Setup

$\mu_0 = \mathcal{N}(0, I_d)$ centered at the origin, μ_1 supported on a smooth codim-k submanifold $M_1 \subset \mathbb{R}^d$ ($1 \le k \le d-1$ for the manifold case; $k = d$ atomic, treated separately), with smooth tangential density ρ_1 on M_1. For field point $x \in \mathbb{R}^d$ near M_1:

- $y(x) := \arg\min_{z \in M_1} |x - z|$, the foot point on M_1;
- $T_{y(x)}M_1$ (tangent, dim $d-k$) and $N_{y(x)}M_1$ (normal, dim k);
- $h\xi^\perp := x - y(x) \in N_{y(x)}M_1$, with $h = |x - y(x)|$ and $\xi^\perp$ a unit vector in $N_{y(x)}M_1$;
- $y(x)^\parallel \in T_{y(x)}M_1$ the projection of $y(x)$ onto $T_{y(x)}M_1$.

We consider interpolants of the generic form $I_t = \alpha_t x_0 + f(\alpha_t)\, x_1$, where $f : [0,1] \to [0,1]$ with $f(0) = 1$, $f(1) = 0$, $f'(\alpha) \le 0$ encodes the curve geometry, and $\alpha_t : [0,1] \to [0,1]$ with $\alpha_0 = 1$, $\alpha_1 = 0$, $\dot{\alpha}_t < 0$ encodes the time parametrization. The path $\Gamma_{x_0,x_1} = \{I_t : t \in [0,1]\}$ depends on f only.

When $f'(0) = 0$, define the path-curvature parameter

$$c \;:=\; -\tfrac{1}{2} f''(0), \qquad \text{so that } 1 - f(\alpha) \sim c\,\alpha^2 \text{ near } \alpha = 0 \tag{37}$$

(for VP, $f(\alpha) = \sqrt{1 - \alpha^2}$, $c = 1/2$).

G.2 Invariance: j is geometric, ν and b are kinematic

Resolving the δ in the line integral $j(x) = \mathbb{E}[\int \delta(x - I_t)\dot{I}_t\, dt]$ via $x_0 = (x - f(\alpha_t)x_1)/\alpha_t$ (Jacobian α_t^{-d}) and $\dot{I}_t = \dot{\alpha}_t[x_0 + f'(\alpha_t)x_1]$, the time-derivative $\dot{\alpha}_t$ exactly cancels the $dt = d\alpha/\dot{\alpha}_t$ Jacobian. Setting $\alpha = \alpha_t$:

$$j(x) \;=\; \int_0^1 d\alpha \int dx_1\, \frac{1}{\alpha^d}\, \rho_0\!\left(\frac{x - f(\alpha)x_1}{\alpha}\right) \rho_1(x_1)\, \left[\frac{x - f(\alpha)x_1}{\alpha} + f'(\alpha)\, x_1\right]. \tag{38}$$

This is manifestly independent of the time parametrization $\alpha_t(\cdot)$ and depends on f only.

By contrast, ν and the time-pointwise ρ_t, b_t depend on the time parametrization. Direct change of variable gives

$$\nu(x) \;=\; \int_0^1 \rho_t(x)\, dt \;=\; \int_0^1 \frac{\tilde{\rho}_\alpha(x)}{|\dot{\alpha}_t(\alpha)|}\, d\alpha, \tag{39}$$

where $\tilde{\rho}_\alpha = \rho_{t(\alpha)}$ is the marginal density relabeled by α. The factor $1/|\dot{\alpha}_t|$ is the kinematic content. Similarly the average drift $\bar{b} := j/\nu$ inherits the ν-dependence.

G.3 Impact of j on the regularity of T

Proposition 3. *The map T is characterized by*

$$j(x) \cdot \nabla T(x) \;=\; 0 \quad \text{on } \mathbb{R}^d \setminus M_1, \qquad T = \mathrm{id} \quad \text{on } M_1. \tag{40}$$

Proof. Trajectories of the autonomous flow $\dot{X}_t = b(X_t) = j(X_t)/\nu(X_t)$ have direction $j/|j|$, independent of $\nu > 0$, so two choices of ν give the same orbits and therefore the same map T. Equivalently, multiplying (40) by ν^{-1} recovers $b \cdot \nabla T = 0$. By (38), j depends only on f. □

The characteristics of (40) are the orbits of $j/|j|$, and M_1 carries the data $T = \mathrm{id}$. Decomposing $j(x) = j^{\perp}(x) + j^{\parallel}(x)$ with $j^{\perp} \in N_{y(x)}M_1$, $j^{\parallel} \in T_{y(x)}M_1$, two regimes arise:

- j **transverse** ($|j^{\parallel}|/|j^{\perp}|$ bounded at M_1): characteristics puncture through M_1 at finite angle. M_1 is non-characteristic for (40), the BVP is C^1-well-posed, T extends to a C^1 map across M_1 (within each basin).
- j **tangential** ($|j^{\parallel}|/|j^{\perp}| \to \infty$ at M_1): characteristics curve to align with M_1 rather than crossing it. M_1 is characteristic in the limit; the BVP is C^0-well-posed but the normal derivative of T blows up at the rate $|j^{\parallel}|/|j^{\perp}|$.

Determining the regime is the goal of the asymptotic analysis below.

G.4 Asymptotics for codim $k \geq 3$ (general ρ_1)

For $k \geq 3$, the saddle of (38) dominates the integral asymptotically as $h \to 0$, and the leading behavior follows from the joint (α, x_1) saddle.

Saddle. The action of (38) is

$$\tilde{S}(\alpha, x_1) \;=\; -\frac{|x - f(\alpha)\, x_1|^2}{2\alpha^2} \;+\; \log \rho_1(x_1) \;-\; d \log \alpha + \mathrm{const.}$$

Joint minimization over $(\alpha, x_1 \in M_1)$. The Euler–Lagrange in x_1 (projected onto $T_{x_1}M_1$):

$$\frac{f(\alpha)}{\alpha^2}\,(x - f(\alpha)x_1)^{\parallel} + \nabla_{\parallel} \log \rho_1(x_1) = 0.$$

For small α, $f \to 1$ and the kinematic term dominates:

$$x_1^*(\alpha) \;=\; y(x) + \alpha^2\,(g + c\, y(x)^{\parallel}) + O(\alpha^4), \qquad g := \nabla_{\parallel} \log \rho_1(y(x)). \tag{41}$$

Substituting back, the action in α alone is $\tilde{S}^*(\alpha) \sim -h^2/(2\alpha^2) - d \log \alpha + \log \rho_1(y(x)) + O(h^2)$, with saddle

$$\alpha^* \;=\; |h|/\sqrt{d}, \qquad x_1^* \;=\; y(x) + O(h^2). \tag{42}$$

Bracket at saddle. Decompose $[(x - f(\alpha)x_1)/\alpha + f'(\alpha)\, x_1]$:

- Normal: $h\xi^{\perp}/\alpha^* = \sqrt{d}\,\mathrm{sgn}(h)\,\xi^{\perp}$ — $O(1)$.
- Tangent (using $f'(0) = 0$, the score correction (41), and $1 - f(\alpha) \sim c\alpha^2$): $-(|h|/\sqrt{d})\,(g + 2c\, y(x)^{\parallel}) + O(|h|^3)$ — $O(|h|)$.

The tangent bracket is one power smaller in $|h|$ than the normal: the $f'(0) = 0$ assumption forces the kinematic and score pieces to nearly cancel.

Asymptotics. Combining the bracket with the post-Laplace integrand structure $\sim 1/\alpha^k \cdot e^{-h^2/(2\alpha^2)}$ and the standard $u = h/\alpha$ substitution,

$$j^{\perp}(x) \;\sim\; -\,\rho_1(y(x))\,\frac{2^{(k-2)/2}\,\Gamma(k/2)}{(\sqrt{2\pi})^k}\,|h|^{1-k}\,\xi^{\perp}, \tag{43}$$

$$j^{\parallel}(x) \;\sim\; \rho_1(y(x))\,[g + 2c\, y^{\parallel}]\,\frac{2^{(k-4)/2}\,\Gamma((k-2)/2)}{(\sqrt{2\pi})^k}\,|h|^{2-k}. \tag{44}$$

For the kinematic ν, an extra $1/|\dot{\alpha}_1|$ appears:

$$\nu(x) \;\sim\; \frac{\rho_1(y(x))\, 2^{(k-3)/2}\, \Gamma((k-1)/2)}{(\sqrt{2\pi})^k\, |\dot{\alpha}_1|}\, |h|^{1-k}. \tag{45}$$

Conclusion (for $k \geq 3$, $f'(0) = 0$). Both currents diverge with $h \to 0$, but the ratio

$$\frac{|j^{\parallel}|}{|j^{\perp}|} \;\sim\; |h| \;\to\; 0 : \tag{46}$$

streamlines approach M_1 *normally*, $T \in C^1$ across M_1. The pre-factor in (46) is irrelevant for the regime determination — only the scaling matters.

Why the saddle fails for $k \leq 2$. The post-Laplace integrand for $j^{\parallel}$ behaves as $1/\alpha^{k-1} \cdot e^{-h^2/(2\alpha^2)}$ near $\alpha = 0$. For $k \geq 3$, the $1/\alpha^{k-1}$ factor concentrates the integrand at $\alpha \sim |h| \to 0$, where the saddle approximation is accurate. For $k = 2$ the integrand $\propto 1/\alpha$ has a logarithmic tail that the Gaussian-around-saddle misses; the actual scaling is $\log(1/|h|)$, not $|h|^0$. For $k = 1$ the integrand $\propto 1$ is bulk-dominated: the integral samples the slowly-varying factors over $\alpha \in (|h|, 1)$, and replacing them by their values at $\alpha = 0$ gives the wrong constant. The exact Gaussian benchmark below makes this concrete.

G.5 The drift $b = j/\nu$ remains bounded

Although $j^{\perp}$, $j^{\parallel}$, and ν may individually diverge at M_1 (for $k \geq 2$), the autonomous-flow drift $b = j/\nu$ stays bounded in every regime. From (43)–(45):

$$b^{\perp}(x) \;=\; \frac{j^{\perp}(x)}{\nu(x)} \;\to\; -|\dot{\alpha}_1|\,\sqrt{2}\,\frac{\Gamma(k/2)}{\Gamma((k-1)/2)}\,\xi^{\perp} \quad (h \to 0,\, k \geq 2,\, f'(0) = 0), \tag{47}$$

$$b^{\parallel}(x) \;=\; \frac{j^{\parallel}(x)}{\nu(x)} \;\sim\; -|\dot{\alpha}_1|\,(g + 2c\,y^{\parallel})\,\frac{\sqrt{2}\,\Gamma((k-2)/2)}{2\,\Gamma((k-1)/2)}\,|h| \;\to\; 0 \quad (h \to 0,\, k \geq 3,\, f'(0) = 0). \tag{48}$$

The Coulomb-type divergences of j and ν cancel exactly in the ratio: $b^{\perp}$ approaches a finite, nonzero constant, and $b^{\parallel}$ vanishes linearly in $|h|$. This is consistent with the streamline regime read off directly from b:

- **Normal streamlines**: $b^{\perp} \to$ finite, $b^{\parallel} \to 0$ — codim ≥ 2 with $f'(0) = 0$.
- **Oblique**: both $b^{\perp}, b^{\parallel}$ finite — codim ≥ 2 with $f'(0) \neq 0$, or codim 1 with $f'(0) = 0$ (where $\nu \sim \log(1/|h|)$ and $j^{\perp}$ finite give $b^{\perp} \to 0$, but the ratio $b^{\parallel}/b^{\perp}$ stays $O(1)$).
- **Tangent**: $b^{\perp} \to 0$, $b^{\parallel} \to$ finite — only codim 1 with $f'(0) \neq 0$. Particles entering the boundary slow down in the normal direction faster than tangentially, sweeping along M_1 rather than crossing it.

G.6 Gauss-law consistency check

The current satisfies $\nabla \cdot j = \mu_0 - \mu_1$, so integrating over $\Omega_\epsilon = \{x : \mathrm{dist}(x, M_1) > \epsilon\}$ and applying the divergence theorem,

$$\int_{\partial\Omega_\epsilon} j \cdot n\, d\sigma \;=\; \int_{\Omega_\epsilon} \mu_0\, dx \;-\; 0 \;\to\; 1 \quad (\epsilon \to 0),$$

with $n = -\xi^{\perp}$ outward from Ω_ϵ (toward M_1). Only $j^{\perp}$ contributes. The cross-section of $\partial\Omega_\epsilon$ at distance ϵ from M_1 scales as

$$\text{cross-section area} \;\sim\; \begin{cases} 2 & (k = 1,\ \text{two sheets}) \\ \epsilon^{k-1}\,|S^{k-1}| & (k \geq 2,\ \text{tube}). \end{cases}$$

For the flux to have a finite, ϵ-independent limit, $j^{\perp}$ must scale as

$$|j^{\perp}| \;\sim\; \rho_1(y(x)) \cdot \frac{1}{\text{cross-section area}} \;\sim\; \rho_1(y)\,|h|^{1-k} :$$

- Codim 1: $|j^\perp| = \rho_1/2$ on each side, two-sheet flux $= \rho_1$ per unit area, integrating to 1.
- Codim ≥ 2: $|j^\perp| \sim \rho_1\,|h|^{1-k}$ Coulomb scaling.

This is forced by the conservation law alone, independent of any approximation. The constant in (43) can be checked: $C_k\,|S^{k-1}| = [2^{(k-2)/2}\Gamma(k/2)/(\sqrt{2\pi})^k] \cdot [2\pi^{k/2}/\Gamma(k/2)] = 1$, giving total flux $\int_{M_1} \rho_1\, d\sigma = 1$. ✓

Why M_1 is a dividing surface only in codim 1. $\mathbb{R}^d \setminus M_1$ is disconnected for $k = 1$ (two sides) and connected for $k \geq 2$. So the "two-sided" structure of $j^\perp$ exists only in codim 1; for $k \geq 2$, the field threads a vanishing ϵ^{k-1} cross-section, and Gauss's law forces it to diverge at M_1. Electrostatic analogy: codim-k charged surface in $\mathbb{R}^d$ produces $|h|^{1-k}$ field — uniform for a charged hyperplane ($k = 1$), $1/r$ for a charged line ($k = 2$), $1/r^2$ for a point charge ($k = d$).

G.7 Direct calculation for the Gaussian case (any k)

Take $\mu_0 = \mathcal{N}(0, I_d)$ and $\mu_1 = \mathcal{N}(0, v_1\, I_{d-k}) \otimes \delta_0$ on $M_1 = \{x^\perp = 0\}$, a planar codim-k subspace through the origin. Everything is jointly Gaussian. Define

$$V(\alpha) \;:=\; \alpha^2 + f(\alpha)^2\, v_1. \tag{49}$$

The marginal density factorizes:

$$\rho_t(x) \;=\; (2\pi V)^{-(d-k)/2}\,(2\pi\alpha_t^2)^{-k/2}\,\exp\Big(-\frac{|x^\parallel|^2}{2V} - \frac{|x^\perp|^2}{2\alpha_t^2}\Big), \tag{50}$$

and the drifts read

$$b_t^\parallel(x) \;=\; \frac{\dot V_t}{2\,V_t}\,x^\parallel, \qquad b_t^\perp(x) \;=\; \frac{\dot\alpha_t}{\alpha_t}\,x^\perp. \tag{51}$$

Changing variable to $\alpha = \alpha_t$ (with the orientation reversal $\alpha : 1 \to 0$ as $t : 0 \to 1$):

$$j^\parallel(x) \;=\; -\frac{x^\parallel}{2}\int_0^1 \frac{V'(\alpha)}{V(\alpha)}\,\rho_\alpha(x)\,d\alpha, \qquad j^\perp(x) \;=\; -x^\perp\int_0^1 \frac{1}{\alpha}\,\rho_\alpha(x)\,d\alpha, \tag{52}$$

$$\nu(x) \;=\; \int_0^1 \frac{\rho_\alpha(x)}{|\dot\alpha_t(\alpha)|}\,d\alpha. \tag{53}$$

These are exact for the Gaussian benchmark, depend on f alone (geometric, in the case of j), and serve as the baseline for testing the codim-k asymptotics.

Asymptotics $h \to 0$. The integrand structure for $j^\parallel$ near $\alpha = 0$ (using $V'(0) = 2f(0)f'(0)v_1 = 0$ for $f'(0) = 0$, hence $V'(\alpha) \sim 2c\alpha\cdot$ stuff and $V'/\alpha^k \sim 1/\alpha^{k-1}$):

codim k	$\nu(x)$	$j^\perp(x)/\xi^\perp$	$j^\parallel(x)$
1	$\rho_1\,\log(1/\lvert h\rvert)/(\sqrt{2\pi}\,\lvert\dot\alpha_1\rvert)$	$-\mathrm{sgn}(h)\,\rho_1/2$	$\int_0^1$ integral, finite
2	$\rho_1 \cdot C_2^\nu \cdot \lvert h\rvert^{-1}$	$-\rho_1 \cdot C_2^\perp \cdot \lvert h\rvert^{-1}$	$\rho_1\,[g + 2cy^\parallel] \cdot C_2^\parallel \cdot \log(1/\lvert h\rvert)$
≥ 3	$\rho_1 \cdot C_k^\nu \cdot \lvert h\rvert^{1-k}$	$-\rho_1 \cdot C_k^\perp \cdot \lvert h\rvert^{1-k}$	$\rho_1\,[g + 2cy^\parallel] \cdot C_k^\parallel \cdot \lvert h\rvert^{2-k}$

with explicit constants $C_k^\nu = 2^{(k-3)/2}\Gamma((k-1)/2)/((\sqrt{2\pi})^k|\dot\alpha_1|)$, $C_k^\perp = 2^{(k-2)/2}\Gamma(k/2)/(\sqrt{2\pi})^k$, $C_k^\parallel = 2^{(k-4)/2}\Gamma((k-2)/2)/(\sqrt{2\pi})^k$ (for $k \geq 3$). The codim-1 entry for $j^\parallel$ is the explicit integral

$$j^\parallel(x) \;=\; \frac{y^\parallel(v_1 - 1)}{(2\pi)^{d/2}}\int_0^1 \frac{e^{-|y^\parallel|^2/(2V(\alpha))}}{V(\alpha)^{(d+1)/2}}\,d\alpha \quad (h = 0,\ \text{VP}), \tag{54}$$

which depends on the full $V(\alpha)$ profile and does *not* reduce to the saddle form $\rho_1[g + 2cy^\parallel]/\sqrt{2\pi}$ (which would replace $V \to v_1$ throughout, missing the bulk variation of V).

Cancellation. For Gaussian μ_1, $j^\parallel \equiv 0$ everywhere (not just at M_1) iff the integrand of (52) vanishes, i.e., $V'(\alpha) \equiv 0$, i.e., $V(\alpha) \equiv$ const. Since $V(0) = v_1$ and $V(1) = v_0 = 1$, this requires $v_1 = 1$: the matched-marginals case.

G.8 Conclusions

- **Invariance.** j and T depend on the path geometry f alone. The time parametrization α_t enters ν and b but cancels from T.
- **Streamline regime, $k \geq 3$, $f'(0) = 0$.** The saddle analysis gives $|j^{\parallel}|/|j^{\perp}| \sim |h| \to 0$ unconditionally: streamlines normal at M_1, $T \in C^1$. The pre-factor is irrelevant for this conclusion — the codim-≥ 3 regime is robust.
- **Streamline regime, $k = 2$, $f'(0) = 0$.** $|j^{\parallel}|/|j^{\perp}| \sim |h| \log(1/|h|) \to 0$: streamlines normal, $T \in C^1$.
- **Streamline regime, $k = 1$.** The saddle approximation fails; one must compute $j^{\parallel}$ explicitly. For Gaussian benchmark with VP, (54): $j^{\parallel}$ is finite, generically nonzero, and the ratio $|j^{\parallel}|/|j^{\perp}|$ is $O(1)$ — streamlines *oblique*, $T \in C^1$. With $f'(0) \neq 0$, $j^{\parallel} \sim \log(1/|h|)$ diverges and the ratio blows up: streamlines tangent, T fails C^1.
- **The pathology — T failing C^1 — occurs only in codim 1 with $f'(0) \neq 0$.** The geometric condition $f'(0) = 0$ (equivalently $\dot{\beta}_1 = 0$, since $\dot{\beta}_t = f'(\alpha_t)\dot{\alpha}_t$) is the unique fix. In all other regimes the BVP (40) is C^1-well-posed.
- **Cancellation.** For Gaussian μ_1, $j^{\parallel} \equiv 0$ iff $v_0 = v_1$ (matched tangential variances). For general ρ_1, an analogous condition would require the V_t-analog of (52) to be constant in t.

G.9 What if $f'(0) \neq 0$?

The analysis of Section G.4 assumed $f'(0) = 0$, which produced an $O(\alpha)$ tangential bracket at the saddle and the regularity conclusion $T \in C^1$. With $f'(0) \neq 0$, two leading-order changes occur.

x_1-saddle and bracket-par. The EL equation in x_1 now balances $f(\alpha)/\alpha^2 \cdot (x - f(\alpha)x_1)^{\parallel}$ against $\nabla_{\parallel} \log \rho_1$ at a different scale, giving

$$x_1^*(\alpha) \;=\; y(x) - \alpha\, f'(0)\, y(x) + O(\alpha^2), \tag{55}$$

shifted by $O(\alpha)$ rather than $O(\alpha^2)$. Computing the bracket at this saddle:

$$[\,\text{bracket-par}\,]^*(\alpha) \;=\; f'(0)\, y(x)^{\parallel} + O(\alpha) \quad (f'(0) \neq 0). \tag{56}$$

Nonzero at $\alpha = 0$ — $O(1)$, versus $O(\alpha)$ when $f'(0) = 0$. The kinematic and score pieces no longer cancel to leading order, leaving a residual proportional to $f'(0)y^{\parallel}$.

Consequence. The post-Laplace integrand for $j^{\parallel}$ becomes $\sim 1/\alpha^k \cdot e^{-h^2/(2\alpha^2)}$ — same α-power as $j^{\perp}$, rather than one power softer. Repeating the saddle/exact analysis:

codim k	$\lvert j^{\perp}\rvert$	$\lvert j^{\parallel}\rvert$	$\lvert j^{\parallel}\rvert/\lvert j^{\perp}\rvert$	streamlines	T at M_1
1	finite	$\log(1/\lvert h\rvert)$	$\to \infty$	**tangent**	**fails** C^1
2	$\lvert h\rvert^{-1}$	$\lvert h\rvert^{-1}$	finite	oblique	C^1
≥ 3	$\lvert h\rvert^{1-k}$	$\lvert h\rvert^{1-k}$	finite	oblique	C^1

Three regimes.

- $k = 1$, $f'(0) \neq 0$: the unique pathological case. $j^{\parallel}$ acquires a log divergence — the bulk-dominated integrand $1/\alpha \cdot e^{-h^2/(2\alpha^2)}$ replaces the constant integrand of the $f'(0) = 0$ case — while $j^{\perp}$ stays finite. Ratio $\to \infty$; streamlines align with M_1; T fails C^1.
- $k \geq 2$, $f'(0) \neq 0$: $j^{\parallel}$ and $j^{\perp}$ scale identically (both Coulomb $|h|^{1-k}$). Ratio finite; streamlines oblique; $T \in C^1$.
- $k \geq 2$, $f'(0) = 0$: $|j^{\parallel}|/|j^{\perp}| \to 0$, streamlines normal; the strongest regularity.

Summary. Setting $f'(0) = 0$ is *necessary* only in codim 1 — without it, T fails C^1. In codim ≥ 2 it is not required: $T \in C^1$ for any f. The condition's role in higher codim is to promote oblique streamlines to normal — a stronger but inessential form of regularity.

G.10 Atomic case ($k = d$)

$T_{y(x)}M_1 = \{0\}$, $y(x)^{\parallel} \equiv 0$, hence $j^{\parallel} \equiv 0$ trivially. The codim-d formula (43) for $j^{\perp}$ applies.

H Self-stopping interpolant for BTM image experiments

Motivated by the empirical success of Wang and Du [2025], we consider a class of interpolant where the minimizer of the consistent loss (5) results in a self-stopping autonomous flow, that is a flow with drift such that $b = 0$ on M_1. In words, as a sample evolves along the dynamics of b, as $t \to \infty$, the sample lands on a point on $\mathrm{supp}(\mu_1)$ and stops moving.

We consider a piecewise interpolant where α_t is chosen as:

$$\alpha_t = \begin{cases} 1 - \dfrac{2\,t}{1+t_c} & 0 \le t \le t_c, \\ \dfrac{(1-t)^2}{1-t_c^2} & t_c < t \le 1. \end{cases} \tag{57}$$

The interpolant inherits the best of both regimes. In the bulk $t \le t_c$, α_t decreases linearly and $\dot{I}_t$ is at full strength: transport is fast. In the tail $t > t_c$, α_t vanishes quadratically, so $a = 2$, $\dot{I}_1 = 0$, and $|b(x)| \sim |x - x_j|^{1/2}$ near each atom — trajectories stop on their own, no stopping criterion needed. The breakpoint $t_c \in (0, 1)$ is a single tunable scalar: $t_c \to 1$ recovers the linear interpolant ($a = 1$, transverse crossing, requires a stopping criterion), $t_c \to 0$ the purely quadratic one (slow asymptotic), intermediate values retain both virtues. Other tail exponents $a > 1$ are obtained by replacing the tail branch with $\alpha_t = a_c(1-t)^a/(1-t_c)^a$ and adjusting a_c to preserve C^1 matching.

I Comparison with Drifting

While the Beckmann transport model (BTM) and the drifting framework Deng et al. [2026] both approximate the fixed points of transport maps, they differ fundamentally in implementation. Drifting relies on learning the fixed-point of a McKean-Vlasov dynamics, driving samples from measure μ_0 to a target probability measure μ_* as $t \to \infty$. In equation, this is written as,

$$\dot{x}_t = V[\mu_t, \mu_*](x_t) := \frac{\mathbb{E}[\mathcal{K}(x_t, X_t)X_t]}{\mathbb{E}[\mathcal{K}(x_t, X_t)]} - \frac{\mathbb{E}[\mathcal{K}(x_t, X_*)X_*]}{\mathbb{E}[\mathcal{K}(x_t, X_*)]}, \qquad x_0 \sim \mu_0 \tag{58}$$

where $\mathcal{K}$ is a kernel of choice, $X_t \sim \mu_t$, and $X_* \sim \mu_*$. As the expectation is not available in an analytical form, in training, the expectation is approximated through a biased estimate from a mini-batch of samples.

Kernel-based approaches are notoriously fragile in high dimensions, as kernel performance is hypersensitive to the target's unknown scale and geometry, often necessitating averaging a mixture of kernels and temperature tuning. In figure 6, we show that the canonical linear transport, and the piecewise self-stopping interpolant is sufficient to obtain good samples without tuning whereas drifting can fail easily depending on choice of kernel and its derived hyperparameters.

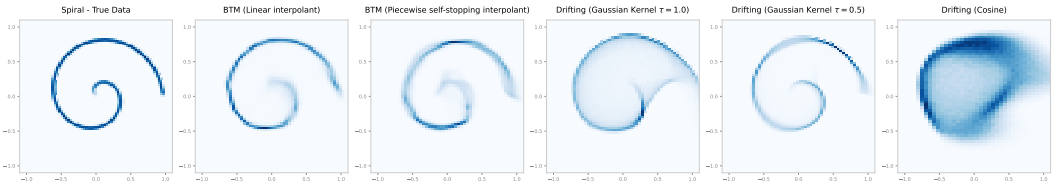


Figure 6: **Comparison between the proposed Beckmann transport models (BTMs) and Drifting Deng et al. [2026].** Depending on the choice of kernel the training of drifting model can be unstable, whereas the proposed Beckmann transport model trains stably across common choices of interpolants.

J Additional experiments

Bias grows monotonically with κ (Figure 7). At $\kappa = 0$ the two losses are algebraically identical and both give $\mathrm{MAE} \approx 0.007$. As κ increases the Wang–Du loss degrades monotonically (up to $25\times$ at $\kappa = 0.9$), while the consistent loss remains flat, confirming that the consistent interpolant decouples schedule choice from transport accuracy. Practitioners may therefore tune the tail behaviour freely to control ODE convergence speed without any accuracy penalty.

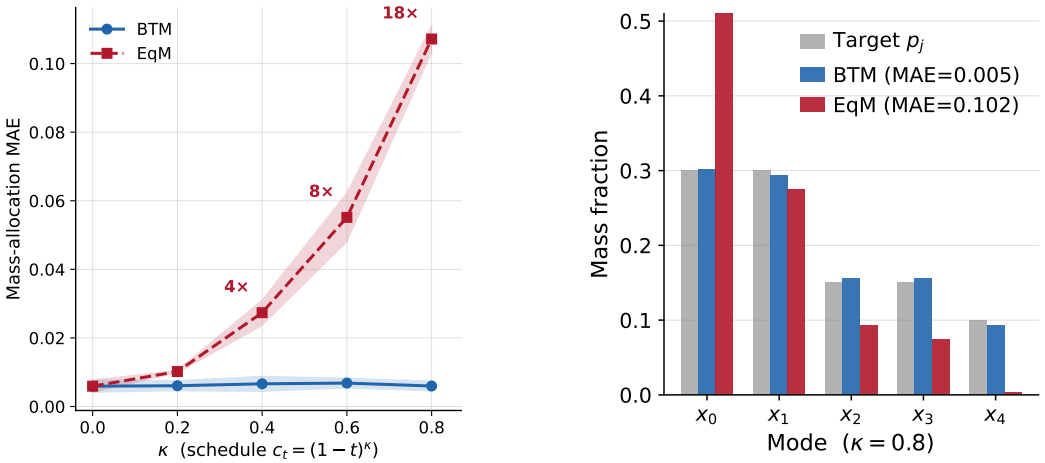


Figure 7: Left: mass-allocation MAE vs. schedule exponent κ (mean $\pm$ std over 3 seeds). Right: per-atom mass fractions at $\kappa = 0.8$. The consistent loss (blue) is invariant to κ; the EqM loss (red) degrades up to $25\times$.

Convergence speed vs. correctness (Figure 8). Both the linear interpolant ($a = 1$, transverse crossing, $\tau < \infty$) and the quadratic interpolant ($a = 2$, self-stopping, $\tau = \infty$), trained with the consistent loss, achieve correct mass allocation ($\mathrm{MAE} \approx 0.005$). The difference is purely in convergence speed: 99% of particles freeze by $T \approx 4$ for $a = 1$ and by $T \approx 6$ for $a = 2$.

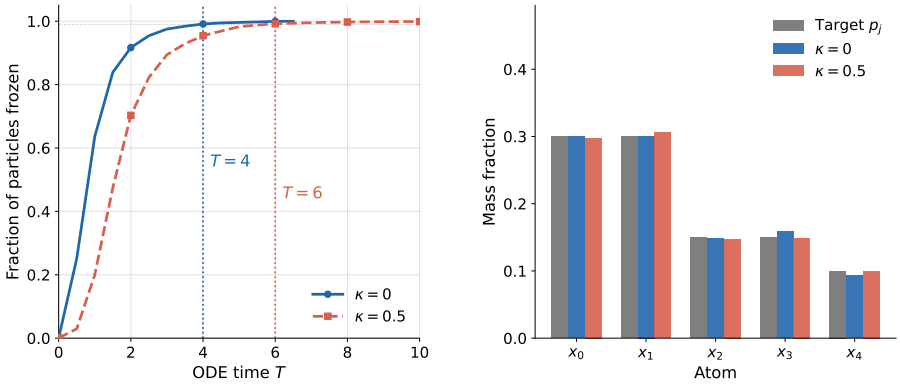


Figure 8: Left: fraction of frozen particles vs. ODE time for $a = 1$ and $a = 2$. Right: both achieve correct mass allocation ($\mathrm{MAE} \approx 0.005$); only convergence speed differs.

Training-free Coulomb transport (Figure 9). For completeness, we verify that the training-free Coulomb field ($\nu = 1$, Theorem 1) transports a Gaussian cloud onto a continuous target without any neural network or training loop. Figure 9 shows this for a Swiss-roll target in $d = 5$: particles are transported onto the spiral manifold by integrating the Coulomb drift with mini-batch estimates. At large d the $1/r^{d-1}$ singularity sharpens, concentrating mass onto $\mathrm{supp}(\mu_1)$ faster.

K Image Experiments for Direct Map Learning

In this appendix, we detail our experimental finding of the direct-map learning algorithm. We consider two experiments: 1) A conditional MNIST digit generation and 2) Conditional ImageNet $256x256$ generation in a latent space. The first acts as a sanity check for the feasibility of the algorithm for images, while the second act as a testbed to rigorously ablate components of the method and its scaling properties.

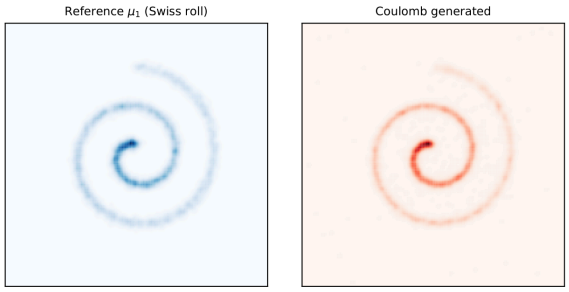
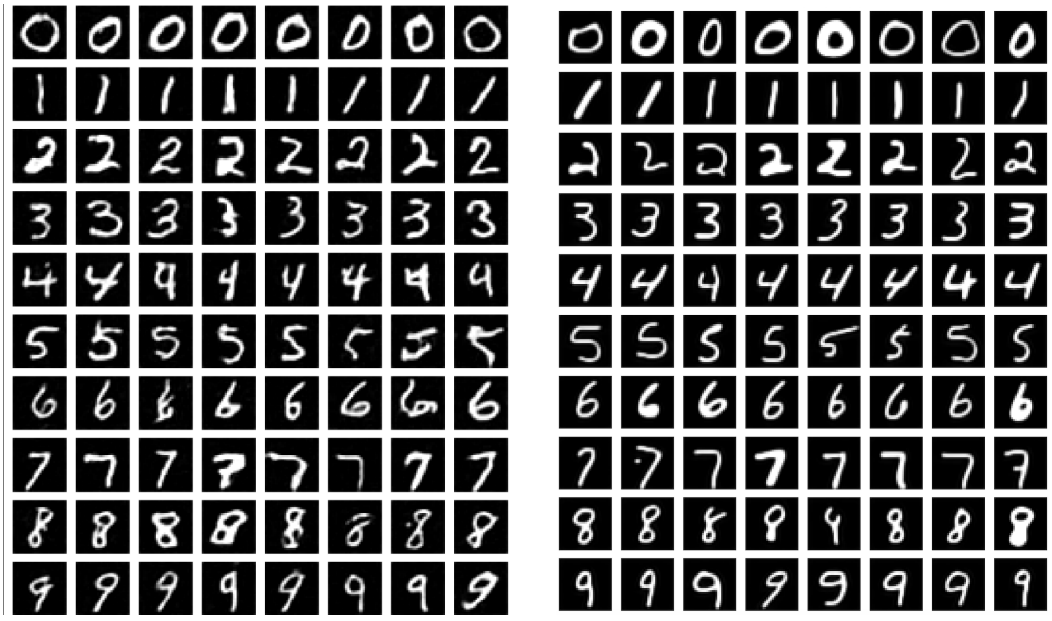


Figure 9: KDE of reference μ_1 (left) and Coulomb-transported particles (right). $\mu_0 = \mathcal{N}(0, I_5)$, $N = 5000$ particles, $M = 1000$ mini-batch, $T = 30$. No training, no score function.

K.1 MNIST digit generation

We consider a class-conditional MNIST digit generation task, where the transport map T is approximated using a standard diffusion U-Net of 23 million parameters with the time conditioning removed from the architecture. Figure 10 shows high quality MNIST digits generated from the learnt transport T. Notably, we observe that repeated application of the learnt transport T removes visual artifiacts and sharpens the edges of the generated digits.



(a) One-step generated results

(b) Two-step generated results

Figure 10: **Generated MNIST digits.** From left, one-step and two-step generated results from the learnt transport map T.

K.2 Latent ImageNet 256×256 generation

We now detail the experimental setup and ablation results for our image experiments. We utilize the SiT architecture Ma et al. [2024] without time embeddings. The models operate entirely within the SD-VAE latent space. We conduct our initial ablations using a B/2 model with 131M parameters. We then scale our model to a 637M parameters XL/2 model. The architectural configurations (layers, hidden dimension, and attention heads) for the Base and Extra-Large variants strictly follow the standard SiT configurations.

Optimization and Stabilization Techniques Scaling the direct map learning algorithm on ImageNet requires careful stabilization of the objective. We train both models using the Muon optimizer with a constant learning rate of 1×10^{-3}, and employ the stabilization techniques:

- **Adaptive Weighting:** Following MeanFlow, we apply an adaptive weight to the transport term of the loss. The adaptively weighted loss is defined as $\text{sg}(w) \cdot \mathcal{L}$, where $\mathcal{L} = ||\Delta||_2^2$ denotes the regression error, and $\text{sg}(\cdot)$ is the stop-gradient operator. The weight w is computed as:
$$w = \frac{1}{(||\Delta||_2^2 + c)^p} \tag{59}$$
We set the hyperparameters to $p = 1$ and $c = 0.01$. This mechanism down-weights samples with excessively large regression errors, stabilizing early training dynamics. While adaptive weighting is highly effective for the transport term, we utilize a plain L_2 loss for the boundary condition. We observe applying adaptive weights to the boundary condition resulted in blurry generated images.
- **Gradient Adaptive Weighting (Term Balancing):** To ensure neither the boundary loss nor the transport loss dominates the updates, we dynamically balance their gradients. Following the adaptive weight mechanism of VQGAN, we weight the transport term by the ratio of their gradient norms with respect to the parameters of the last layer (L):
$$\lambda = \frac{||\nabla_L \mathcal{L}_{\text{boundary}}||}{||\nabla_L \mathcal{L}_{\text{transport}}||} \tag{60}$$
- **Anchoring the slightly noisy image:** To increase the learning signal from the boundary loss term, we find it useful to apply the loss not only for time $I_{s=1}$, but for $I_{s>\tau}$ where τ is a prespecified threshold. The motivation is that when a small amount of noise is injected, the clean image should be easily recoverable. We find $\tau = 0.8$ to be effective.

Training Configurations and Hyperparameters For hyperparameter selection and objective stabilization, we ablate using the BTM B/2 model, which is trained for a total of 80 epochs. To achieve optimal generation quality and convergence, the full BTM XL/2 model is trained for an extended period of 1280 epochs. A comprehensive summary of our architectural details and training hyperparameters is provided in Table 3.

Ablation and Scaling Results Using the stabilized objective, we present our ablation and scaling results below. We first evaluated the BTM B/2 model trained for 80 epochs against the MeanFlow B/2 model (Table 2a); we observe that BTM is competitive with MeanFlow without guidance (CFG). We additionally ablate the effect of scaling the model architecture (Table 2b). Increasing the parameter count to an XL/2 size yields significant stabilization and improvements in early training.

Iterated Composition We also explore the iterated application of the map. Repeated application does not require re-noising, time conditioning, or a schedule, but rather plain composition, which primarily serves to remove visual artifacts. As shown in Table 2c, increasing the number of function evaluations (NFE) through iteration yields a modest improvement in FID.

Method	**FID**
MeanFlow B/2	61.06
BTM B/2	62.14

(a) 80-Epoch Baseline Comparison (Without CFG)

Method	**FID**
BTM B/2	62.14
BTM XL/2	47.58

(b) Effect of Model Scaling at 80 Epochs

NFE	1	2	5
FID	17.58	17.04	16.53

(c) Iterated application of the direct map for BTM.

Configuration	**BTM B/2 (Ablation)**	**BTM XL/2 (Full Scale)**
Generator Architecture		
Architecture	SiT-B	SiT-XL
Latent Space	SD-VAE	SD-VAE
Input size	$32 \times 32 \times 4$	$32 \times 32 \times 4$
Patch size	2×2	2×2
Hidden size d	768	1152
Layers N	12	28
Attention Heads	12	16
Optimizer Settings		
Optimizer	Muon	Muon
Learning Rate	1×10^{-3}	1×10^{-3}
Training Epochs	80	1280
Loss Stabilization		
Adaptive Weight p	1	1
Adaptive Weight c	0.01	0.01
Boundary Loss	Plain L_2	Plain L_2
Gradient Balancing	Yes (last layer)	Yes (last layer)

Table 3: Hyperparameter Configurations for BTM Models. Architectural dimensions match the standard SiT configurations.