ADAPTING NOISE TO DATA: GENERATIVE FLOWS FROM LEARNED 1D PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a general framework for constructing generative models using one-dimensional noising processes. Beyond diffusion processes, we outline examples that demonstrate the flexibility of our approach. Motivated by this, we propose a novel framework in which the 1D processes themselves are learnable, achieved by parameterizing the noise distribution through quantile functions that adapt to the data. Our construction integrates seamlessly with standard objectives, including Flow Matching and consistency models. Learning quantile-based noise naturally captures heavy tails and compact supports when present. Numerical experiments highlight both the flexibility and the effectiveness of our method.

1 INTRODUCTION

Flow-based generative models, especially score-based diffusion Sohl-Dickstein et al. (2015); Song & Ermon (2019), flow matching (FM) Albergo et al. (2023); Lipman et al. (2023); Liu (2022) and consistency models like the recently introduced inductive moment matching (IMM) Zhou et al. (2025), achieve state-of-the-art results in many applications. All these methods construct a probability flow from a simple latent distribution (noise) to a complex target (data) with a neural network trained to approximate this flow from limited target samples. In diffusion models, the *score function* directs a reverse-time SDE, while in FM, the *velocity field* is learned to compute trajectories via a flow ODE. Finally, consistency models like IMM learn to predict the jumps from noise to the data while factoring in the consistency of the flow trajectories. Usually, a Gaussian is used as latent distribution which causes difficulties when learning certain multimodal and heavy-tailed targets Hagemann & Neumayer (2021); Salmona et al. (2022), see Figure 2 for a heavy-tailed example. There exist only few approaches to learn the noising process, Bartosh et al. (2025) fit the forward diffusion process via a learned invertible map that is trained end-to-end, Kapusniak et al. (2024) use metric flow matching, i.e., a neural network to adapt the path to a underlying Riemannian metric. On the other hand Pandey et al. (2024); Zhang et al. (2024) design heavy-tailed diffusions using Student-t latent distributions, and Shariatian et al. (2025) extend the framework to the family of α-stable distributions.

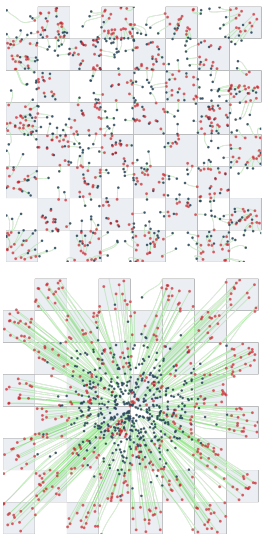


Figure 1: The learned noise (top) in conjunction with optimal coupling FM drastically shortens the transport paths compared to FM with Gaussian noise (bottom).

In this paper, we present a new approach to adapt the latent distribution to the data by *learning* from its samples. The basic idea comes from

the fact that all the above methods implicitly emerge as *componentwise* models. For example, denoting the target random variable by $\mathbf{X}_0$ and the latent by $\mathbf{X}_1 \sim \mathcal{N}(0, I_d)$, FM utilizes the process $\mathbf{X}_t = (X_t^1, \dots, X_t^d)$ with the components $X_t^i = (1-t)X_0^i + tX_1^i$ employing *one-dimensional* Gaussians $X_1^i \sim \mathcal{N}(0,1)$. This motivated us to generally construct generative models from *1D processes and their quantile functions*.

Given any appropriate 1D process we demonstrate how to learn the componentwise neural flow by the associated conditional velocity field. We give examples besides diffusion demonstrating the flexibility of our machinery, namely the Kac process arising from the 1D damped wave equation, and a process reflecting the Wasserstein gradient flow of the maximum mean discrepancy with negative distance kernel. In contrast to diffusion, assuming a compactly supported target, these processes also have a compact support, leading to a better regularity of the corresponding velocity field. This inspired us to further adapt the process to the data and to *learn* the 1D noising process rather than choosing it manually. To this end, we exploit that 1D probability measures can be equivalently described by their quantile functions $Q^i : (0,1) \to \mathbb{R}$ which are monotone functions, and consider quantile processes $X_t^i = (1-t)X_0^i + tQ^i(U^i)$, $i = 1, \dots, d$ with i.i.d. $U^i \sim \mathcal{U}[0,1]$ for $t \in [0,1]$. We learn the individual quantile functions Q_ϕ^i, $i = 1, \dots, d$ such that their componentwise concatenation $\mathbf{Q}_\phi(\mathbf{U}) := (Q_\phi^i(U^i))_{i=1}^d$ is *"close"* to the data. This inspired us to minimize

$$W_2^2(\mu_0, \mathrm{Law}(\mathbf{Q}_\phi(\mathbf{U}))), \quad \mu_0 = \mathrm{Law}(\mathbf{X}_0).$$

with the Wasserstein distance W_2. We combine the learning of the latent $\mathbf{Q}_\phi(\mathbf{U})$ with the learning of the velocity field via optimal coupling FM. This allows us to effectively exploit the learned noise and drastically shorten the transport paths, as illustrated in Figure 1.

The simplicity of quantile functions give us a flexible tool, which enables us to simultaneously learn the noising process and apply the FM framework. Our quantile perspective can further be extended to fit into consistency models.

Contributions. 1. We introduce a general construction method for neural flows by decomposing multi-dimensional flows into one-dimensional components. Ultimately, this allows us to work with *one-dimensional noising* processes in the FM framework.

2. We highlight three interesting noising processes for our framework: the Wiener process, the 1D Kac process and the 1D MMD gradient flow with negative distance kernel and uniform target measure.

3. Based on the decomposition viewpoint, we propose to describe our 1D noising processes by their *quantile functions*. Via quantile interpolants, our framework can also be incorporated into consistency models.

4. Exploiting the simplicity of quantile functions, we propose to *learn* the quantile of the 1D noise simultaneously within the FM framework, aiming to fit the noise to the data. Numerical experiments demonstrate the high flexibility of our data-adapted noise.

2 Preliminaries: Flow Matching

We start with a brief introduction of curves in Wasserstein spaces and basic ideas on flow matching. For more details we refer to Ambrosio et al. (2008) and Wald & Steidl (2025). Let $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ denote the complete metric space of probability measures with finite second moments equipped with the Wasserstein distance

$$W_2^2(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \mathrm{d}\pi(x, y)$$

Here $\Pi(\mu, \nu)$ denotes the set of all probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ having marginals μ and ν. The push-forward measure of $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ by a measurable map $\mathcal{T} : \mathbb{R}^d \to \mathbb{R}^d$ is defined by $\mathcal{T}_\sharp \mu := \mu \circ \mathcal{T}^{-1}$. Let I be an interval in $\mathbb{R}$, in this paper mainly $I = [0,1]$. A narrowly continuous curve $\mu_t : I \to \mathcal{P}_2(\mathbb{R}^d)$ is absolutely

continuous, iff there exists a Borel measurable vector field $v : I \times \mathbb{R}^d \to \mathbb{R}^d$ with $\|v_t\|_{L_2(\mathbb{R}^d,\mu_t)} \in L_2(I)$ such that (μ_t, v_t) satisfies the continuity equation

$$\partial_t \mu_t + \nabla_x \cdot (\mu_t v_t) = 0 \tag{1}$$

in the sense of distributions. If in addition $\int_I \sup_{x\in B} \|v_t(x)\| + \mathrm{Lip}(v_t, B)\,\mathrm{d}t < \infty$ for all compact $B \subset \mathbb{R}^d$, then the ODE

$$\partial_t \varphi(t,x) = v_t(\varphi(t,x)), \qquad \varphi(0,x) = x, \tag{2}$$

has a solution $\varphi : I \times \mathbb{R}^d \to \mathbb{R}^d$ and $\mu_t = \varphi(t,\cdot)_\sharp \mu_0$.

Starting in the target distribution μ_0 and ending in a simple latent distribution μ_1, as usual in diffusion models, we can reverse the flow from the latent to the target distribution using just the opposite velocity field $-v_{1-t}$ in the ODE (2). Thus, if somebody provides us with the velocity field v_t, we can sample from a target distribution by starting in a sample from the latent one and then applying our favorite ODE solver.

If we do not have a velocity field donor, we can try to approximate (learn) the velocity field by a neural network v_t^θ. Clearly, a desirable loss function would be

$$\mathcal{L}(\theta) \coloneqq \mathbb{E}_{t\sim\mathcal{U}(0,1),\, x\sim\mu_t} \left[\left\| v_t^\theta(x) - v_t(x) \right\|^2 \right].$$

Unfortunately this loss function is not helpful, since we do not know the exact velocity field v_t nor can sample from μ_t in the empirical expectation. However, employing the law of total probabilities, as done, e.g. in Lipman et al. (2023), we see that $\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{CFM}}(\theta) + \text{const}$ with a constant not depending on θ and the Conditional Flow Matching (CFM) loss

$$\mathcal{L}_{\mathrm{CFM}}(\theta) \coloneqq \mathbb{E}_{x_0\sim\mu_0,\, t\sim\mathcal{U}(0,1),\, x\sim\mu_t(\cdot|x_0)} \left[\left\| v_t^\theta(x) - v_t(x|x_0) \right\|^2 \right]. \tag{3}$$

The key difference is the use of the *conditional flow* $v_t(x|x_0)$ with respect to a fixed sample x_0 from our target distribution. To summarize, all you need is a *conditional* flow model with accessible velocity field $v_t(x|x_0)$ (at least along the flows trajectory), where you can easily sample from. Then you can indeed learn the velocity field v_t of the general (non-conditional) flow and finally sample from the target by the reverse ODE (2).

3 Multi-dimensional flows via their one-dimensional components

We begin by outlining a *general* framework based on stochastic processes for flow–based sampling from a given data distribution μ_0, see e.g. Albergo et al. (2023). Then we restrict ourselves to componentwise independent noising processes and show how they integrate into the framework. Finally, we recast the construction from a one-dimensional viewpoint using quantile interpolants.

3.1 Construction via stochastic processes

Consider a (noising) process $(\mathbf{Y}_t)_t$ with $\mathbf{Y}_0 \equiv 0 \in \mathbb{R}^d$ with associated velocity field $v_t = v_t^{\mathbf{Y}}(\cdot \mid 0)$ such that the pair $(\mu_t^{\mathbf{Y}}, v_t^{\mathbf{Y}})$ satisfy the continuity equation (1), where $\mu_t^{\mathbf{Y}}$ is the law of $(\mathbf{Y}_t)_t$. To construct a generative model we need to create a process $(\mathbf{X}_t)_t$ which can start in any sample x_0 from the target measure μ_0. Let $\mathbf{X_0} \sim \mu_0$. Following the lines in Duong et al. (2025), we define the *mean-reverting* process by

$$\mathbf{X}_t \;\coloneqq\; f(t)\,\mathbf{X_0} \;+\; \mathbf{Y}_{g(t)}, \quad t \in [0,1], \tag{4}$$

with smooth *scheduling functions* f, g

$$f(0) = 1, \quad f(1) = 0 \quad \text{and} \quad g(0) = 0, \quad g(1) = 1. \tag{5}$$

Then we have $\mathbf{X}_1 = \mathbf{Y}_1$, and by abuse of notation, the process $\mathbf{X}_t$ starts in $\mathbf{X}_0 = \mathbf{X}_0$. Differentiation of (4) results in

$$\dot{\mathbf{X}}_t = \dot{f}(t)\,\mathbf{X_0} \;+\; \dot{g}(t)\,\dot{\mathbf{Y}}_{g(t)}.$$

Hence the conditional velocity field of $\mathbf{X}_t$ is given by (see Lipman et al. (2023))

$$\begin{aligned} v_t^{\mathbf{X}}(x \mid x_0) &= \mathbb{E}\big[\dot{\mathbf{X}}_t \mid \mathbf{X}_t = x,\ \mathbf{X_0} = x_0\big] \\ &= \mathbb{E}\big[\dot{f}(t)\,x_0 + \dot{g}(t)\,\dot{\mathbf{Y}}_{g(t)} \;\big|\; \mathbf{Y}_{g(t)} = x - f(t)x_0\big] \\ &= \dot{f}(t)\,x_0 \;+\; \dot{g}(t)\,v_{g(t)}^{\mathbf{Y}}\big(x - f(t)x_0 \mid 0\big). \end{aligned} \tag{6}$$

Now, the conditional flow matching loss (3) can be minimized regarding $\mathbf{X_0} \sim \mu_0$ and $\mathbf{X}_t \sim \mu_t$. Note that given a sample $x \sim (\mathbf{X}_t \mid \mathbf{X}_0 = x_0)$, we have[1] $v_t^{\mathbf{X}}(x \mid x_0) = \dot{f}(t)\,x_0 \;+\; \dot{g}(t)\,v_{g(t)}^{\mathbf{Y}}\big(\mathbf{Y}_{g(t)} \mid 0\big)$.

Remark 1 (Relation to FM and diffusion)**.** *Consider the stochastic process*

$$\mathbf{X}_t^{\mathrm{FM}} = \alpha_t \mathbf{X}_0 + \sigma_t \mathbf{X}_1, \qquad \mathbf{X}_1 \sim \mathcal{N}(0, I_d). \tag{7}$$

Choosing $f(t) \coloneqq \alpha_t$, $g(t) \coloneqq \sigma_t^2$ and the standard Brownian motion $\mathbf{Y}_t = \mathbf{W}_t$, it holds the equality in distribution

$$\mathbf{X}_t^{\mathrm{FM}} \overset{d}{=} f(t)\mathbf{X_0} + \mathbf{W}_{g(t)} = \mathbf{X}_t.$$

Then $f(t) \coloneqq 1 - t$, $g(t) \coloneqq t^2$ yields (independent) FM Lipman et al. (2023), and $f(t) \coloneqq \exp\left(-\frac{h(t)}{2}\right)$, $g(t) \coloneqq 1 - \exp\left(-h(t)\right)$, where $h(t) \coloneqq \int_0^t \beta_{\min} + s(\beta_{\max} - \beta_{\min})\,\mathrm{d}s$ with, e.g., $\beta_{\min} = 0.1$, $\beta_{\max} = 20$, corresponds to processes used in score-based generative models Song et al. (2021), see Appendix B.

Motivated by the fact that a multi-dimensional Wiener process $\mathbf{W}_t \in \mathbb{R}^d$ consists of *independent* (and identically distributed) 1D components $\mathbf{W}_t = (W_t^1, ..., W_t^d)$, we propose to construct a d-dimensional flow $\mathbf{Y}_t$ componentwise, based on independent one-dimensional processes Y_t^i.

3.2 Construction of Componentwise Flows

Restricting ourselves to processes $\mathbf{Y}_t$ that decompose into one-dimensional components allows us to propose our **general construction method** for accessible *conditional* flows in FM. Let $Y_t^1, \ldots, Y_t^d$ be a family of independent one-dimensional stochastic processes with time dependent laws $\mu_t^i \in \mathcal{P}_2(\mathbb{R})$. For each $i = 1, \ldots, d$, let $v_t^i \colon \mathbb{R} \to \mathbb{R}$ be the associated velocity field such that the pair (μ_t^i, v_t^i) satisfies the one-dimensional continuity equation (1). Define the product measure $\mu_t \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$\mu_t(x) \;=\; \prod_{i=1}^{d} \mu_t^i(x^i), \qquad x = (x^1, \ldots, x^d) \in \mathbb{R}^d. \tag{8}$$

For the d-dimensional process $\mathbf{Y}_t \coloneqq (Y_t^1, \ldots, Y_t^d)$, independence implies that its law is exactly μ_t. Moreover, by the following proposition, the corresponding d-dimensional velocity field is given componentwise, see Duong et al. (2025).

Proposition 2. *Let μ_t be given by* (8)*, where the μ_t^i are absolutely continuous curves in $\mathbb{R}$ with velocity fields v_t^i. Then μ_t satisfies a continuity equation* (1) *with a velocity field which decomposes into the univariate velocities*

$$v_t(x) \coloneqq \left(v_t^1(x^1), \ldots, v_t^d(x^d)\right).$$

Using these insights on componentwise flows, we propose the following guide for constructing neural flows.

[1] In general, $v^{\mathbf{Y}}$ might not be tractable, and only given as an conditional expectation of the time derivative $\dot{\mathbf{Y}}$. Yet, through our componentwise construction below, we will obtain easier access to it via its 1D components.

General construction method for accessible conditional flows in FM

1. *One-dimensional noise:* Start with an appropriate absolutely continuous measure curve μ_t starting in $\mu_0 = \delta_0$, $0 \in \mathbb{R}$, where you can compute its velocity field v_t in the 1D continuity equation
$$\partial_t \mu_t + \partial_x(\mu_t v_t) = 0, \qquad \mu_0 = \delta_0. \tag{9}$$
Appropriate 1D noising processes are provided in Section 4.
2. *Multi-dimensional noise:* Set up a multi-dimensional conditional flow model starting in $\mu_0 = \delta_0$, $0 \in \mathbb{R}^d$ with possibly different, but independent 1D processes as described in Section 3.2.
3. *Incorporating the data:* Construct a multi-dimensional conditional flow model starting in $\mu_0 = \delta_{x_0}$ for any data point $x_0 \sim \mu_0$ by mean-reversion as shown in Section 3.1.

3.3 Quantile Processes

The restriction to componentwise noising processes $\mathbf{Y}_t$ in (4) [2] allows us to use the quantile functions of the 1D components. Recall that the *cumulative distribution function* (CDF) R_μ of $\mu \in \mathcal{P}_2(\mathbb{R})$ and its *quantile function* Q_μ are given by

$$R_\mu(x) \coloneqq \mu\big((-\infty, x]\big), \quad x \in \mathbb{R} \quad \text{and} \quad Q_\mu(u) \coloneqq \min\{x \in \mathbb{R} : R_\mu(x) \geq u\}, \quad u \in (0,1). \tag{10}$$

In Figure 7 we exemplify the CDF and quantile of a standard Gaussian. The quantile functions form a closed, convex cone $\mathcal{C} \coloneqq \{f \in L_2(0,1) : f \text{ increasing } a.e.\}$ in $L_2(0,1)$. The mapping $\mu \mapsto Q_\mu$ is an isometric embedding of $(\mathcal{P}_2(\mathbb{R}), W_2)$ into $(L_2(0,1), \|\cdot\|_{L_2})$, meaning that

$$W_2^2(\mu, \nu) = \int_0^1 \big|Q_\mu(s) - Q_\nu(s)\big|^2 \, \mathrm{d}s$$

and $\mu = Q_{\mu,\sharp}\mathcal{L}_{(0,1)}$. Let $U \sim \mathcal{U}[0,1]$ be uniformly distributed on $[0,1]$. Now, any probability measure flow μ_t can be described by their respective quantile flow $Q_t \coloneqq Q_{\mu_t}$, such that $\mu_t = Q_{t,\sharp}\mathcal{L}_{(0,1)}$ and $Q_t \circ U$ is a stochastic process with marginals μ_t.

We can therefore model any *multi-dimensional* noising process, that decomposes into its components, via quantile functions. Namely let X_0 be any component $\mathbf{X}_0^i$ of $\mathbf{X}_0 \sim \mu_0$, and $f, g : [0,1] \to \mathbb{R}$ smooth schedules fulfilling (5). We assume that we are given a flow $(Q_t)_t$ of quantile functions $Q_t : (0,1) \to \mathbb{R}$, $t \in [0,1]$, which fulfill $Q_0 \equiv 0$ and are invertible on their respective image with the inverse given by the CDF $R_t : Q_t(0,1) \to \mathbb{R}$. We introduce the *quantile process*

$$Z_t = f(t)X_0 + Q_{g(t)}(U), \quad U \sim \mathcal{U}(0,1),\, t \in [0,1]. \tag{11}$$

The quantile process coincides (in distribution) with the components of the mean-reverting process (4), where the noising term is represented as $\mathbf{Y}^i_{g(t)} \overset{d}{=} Q_{\mathrm{Law}(\mathbf{Y}^i_{g(t)})}(U)$. In particular, the components of the process (7) are obtained via (11) using the quantile Q_t of a standard Brownian motion W_t and $f(t) \coloneqq \alpha_t$, $g(t) \coloneqq \sigma_t^2$.

Quantile Interpolants. Let us briefly mention how our setting fits into the framework of consistency models. To this end, we define the *quantile interpolants*

$$I_{s,t}(x,y) = f(s)x + Q_{g(s)}\big(R_{g(t)}(y - f(t)x)\big), \quad s,t \in [0,1] \tag{12}$$

which generalize the interpolants used in Denoising Diffusion Implicit Models (DDIM), see Remark 9.

[2] Besides componentwise 1D processes we may also use triangular decompositions, not addressed in this paper.

Proposition 3. *For all $x, y \in \mathbb{R}$ and all $s, r, t \in [0,1]$, it holds $I_{0,t}(x,y) = x$, $I_{t,t}(x,y) = y$, and*

$$I_{s,r}(x, I_{r,t}(x,y)) = I_{s,t}(x,y).$$

Furthermore, inserting the quantile process (11) *yields $I_{s,t}(Z_0, Z_t) = Z_s$.*

The proof is given Appendix C. Proposition 3 allows us to also apply the concept of consistency models to our quantile process (11). The shared idea of these models is to predict the jumps from the process Z_t to the target X_0, while factoring in the *consistency* of the trajectory of Z_t via Z_s, $0 < s < t$. In FM, this consistency of the flow is usually neglected as only single points on the FM paths are sampled. Also, consistency models as one-step or multistep samplers usually are in no need of velocity fields. In the Appendix C, we demonstrate by means of the recently proposed *inductive moment matching* (IMM) Zhou et al. (2025), that our formulation via quantile interpolants fits seamlessly into the consistency framework.

4 One-dimensional Processes: From Prescribed to Learned

Next, we address the question of finding "good" 1D processes Y_t^i which can drive our mean-reverting process (4). Aside of the Wiener process, we highlight two other ones with accessible velocities and conditional measures in Section 4.1. These processes have characteristics very different from diffusion, notably non-exploding vector fields. This raises the question which 1D processes are best suited for certain problems. In Subsection 4.2, we present a new method for learning data-adapted processes via their quantile functions.

4.1 One-dimensional flows besides diffusion

We explore three interesting 1D (noising) processes Y_t in connection with their respective PDEs, for which our approach via reduction to one dimension is nicely applicable, namely the

- Wiener process W_t and diffusion equation,
- Kac process K_t and damped wave equation,
- Uniform process U_t and the gradient flow of the maximum mean functional $\mathcal{F}_\nu := \mathrm{MMD}_K(\cdot, \nu)$ with negative distance kernel $K(x,y) = -|x-y|$ and $\nu = \mathcal{U}(-b,b)$.

In each case, we explicitly calculate the respective conditional measure flow and its conditional velocity field in Appendix A, such that the conditional flow matching loss (3) can be minimized. Note that in contrast to the Wiener process W_t usually seen in diffusion and flow matching models, the latter two processes K_t, U_t do not enjoy a trivial analogue in multiple dimensions: in case of K_t the corresponding PDE (damped wave equation) is no longer mass-conserving in dimension $d \geq 3$, see Tautz & Lerche (2016); in case of U_t the mere existence of the MMD gradient flow in multiple dimensions is unclear by the lack of convexity of the MMD, see Hertrich et al. (2024). Our general construction method makes these 1D processes accessible for generative modeling in arbitrary dimensions.

4.2 Learning 1D Processes via Quantile Functions

The choice of the noise can have a significant impact on the sampling performance, see Figure 1 for the checkerboard distribution and Figure 2 for a heavy-tailed one. Now we adopt the quantile process view from Section 3.3 to learn data–adapted noise. We pose the following requirements on the latent distribution ν: i) absolute continuity, ii) data–independence, and iii) independence of components (to fit our 1D construction).

[4]Note that we used the independent coupling for training of these models. We also used z-score normalization.

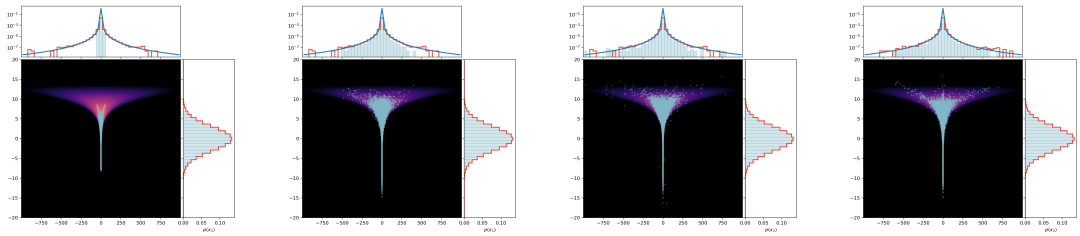


Figure 2: Sampling of Neal's funnel with different latent distributions.[4]From left to right with uniform ($[-1, 1]$), standard Gaussian, Student-T (with parameters $(20, 4)$ inspired by the choice in Pandey et al. (2024)) and our learned distribution. The last two heavy-tailed noises perform significantly better.

Under these assumptions the latent class reduces to the set $S \coloneqq \{\nu \in \mathcal{P}_2(\mathbb{R}^d) : \nu = \rho\,\mathrm{d}x \text{ and } \rho = \Pi_{i=1}^d \rho^i\}$, i.e. considering quantile processes of the form

$$X_t^i \;=\; (1-t)\,X_0^i \;+\; tQ^i(U^i),\; i = 1, \ldots, d,\; t \in [0,1],$$

we have $\nu = \mathbf{Q}_{\#}\,\mathcal{U}([0,1]^d)$ with $\mathbf{Q}(u) \coloneqq (Q^1(u^1), \ldots, Q^d(u^d))$. In particular, in our framework the quantile family determines the scales and tails of $\mathbf{Q}(\mathbf{U})$, thereby influencing the difficulty and inductive bias of predicting the conditional velocity $v_t(\mathbf{X}_t) = \mathbf{Q}(\mathbf{U}) - \mathbf{X}_0$ along the linear paths $\mathbf{X}_t = (1-t)\mathbf{X}_0 + t\mathbf{Q}(\mathbf{U})$.

We now describe how we learn the quantile maps $\mathbf{Q}_\phi$. The core idea is that besides our requirements i)-iii) as well as being a valid quantile function, we would like our noise to be *"close"* to the data. We learn $\mathbf{Q}_\phi$ by minimizing a statistical discrepancy, e.g. the Wasserstein distance, between μ_0 and ν_ϕ,

$$\mathcal{E}(\phi) \;=\; W_2^2\big(\mu_0, \nu_\phi\big), \quad \nu_\phi \coloneqq (\mathbf{Q}_\phi)_{\#}\,\mathcal{U}([0,1]^d). \tag{13}$$

Note that due to the restriction of our quantiles to the class S, the minimizer of (13) is in general *not* μ_0.

While our quantiles can be trained independently, in order to provide an aligned training signal for the velocity field, we propose to also train $\mathbf{Q}_\phi$ *jointly* with the velocity v_θ. Hence, we aim to minimize the loss

$$\begin{aligned} \mathcal{L}(\theta;\phi) \;&=\; \mathcal{E}_{\mathsf{CFM}}(\theta;\phi) + \lambda\,\mathcal{E}(\phi), \quad \lambda > 0, \\ \text{with} \quad \mathcal{E}_{\mathsf{CFM}}(\theta;\phi) \;&=\; \mathbb{E}_{t\sim\mathcal{U}(0,1),(x,y)\sim\pi_\phi}\Big[\big\|v_\theta\big((1-t)x + ty,\, t\big) - (y-x)\big\|_2^2\Big], \end{aligned}$$

where $\pi_\phi \in \Pi_o(\mu_0, \nu_\phi)$ is an optimal coupling between μ_0 and ν_ϕ.

In practice, we optimize the empirical expectation via minibatches; see Appendix D.4. A pseudo-algorithm is provided in Algorithm 1. In particular, we compute a mini batch optimal transport map T that minimizes $\sum_{j=1}^B \|\mathbf{x}_0^{(j)} - \mathbf{y}^{(T(j))}\|_2^2$ for batches of data $\{\mathbf{x}_0^{(j)}\}_{j=1}^B$, $\{\mathbf{y}^{(j)}\}_{j=1}^B$ from $\mathbf{X}_0$ and $\mathbf{Q}_\phi(\mathbf{U})$, respectively. This minibatch map T is reused below for flow matching to keep the targets consistent across the two terms.

5 Experiments

To validate our proposed method, we conduct experiments on synthetic and imaging datasets. We parametrize the latent distribution's quantile function using Rational Quadratic Splines (RQS) Durkan et al. (2019). This choice is motivated by several factors: RQS enforce monotonicity by construction, are parameter-efficient, and provide access to analytic derivatives. For our experimental setup, see Table 1.

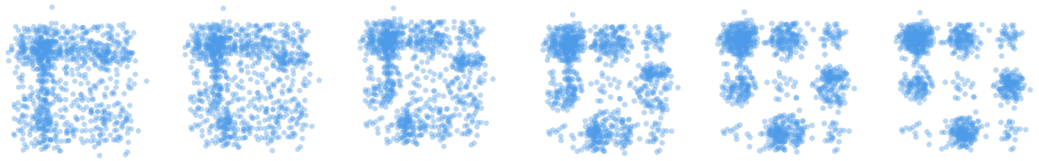


Figure 3: A generated trajectory from the learned quantile latent (left) to the unevenly weighted Gaussian mixture target (right). The adapted latent is already close to the target distribution.

5.1 Analysis on Synthetic Datasets

We begin by qualitatively analyzing our algorithm on several synthetic 2D distributions (see Appendix D.3), each designed to highlight a specific aspect of our approach.

Gaussian Mixture Model (GMM). We first consider a 2D GMM with nine unevenly weighted modes, as visualized in Figure 3. Due to the independence assumption inherent in our factorized quantile function, the learned latent cannot perfectly replicate the target's joint distribution and is not the product of the correct marginals; see also D.1. Instead, it approximates a distribution where the components cannot further independently improve the transport cost to the target.

Funnel Distribution. The funnel distribution, shown in Figure 2, presents a challenge due to its heavy-tailed, conditional structure. This experiment highlights the importance of matching the latent's tail behavior to that of the target. We observe that our learned latent successfully adapts to the target's heavy tails, see also the visualization in Figure 10. This enables the flow matching model to generate high fidelity samples across the distribution. Note that due to the high variance signal when training on the funnel distribution, we pre-train our quantile.

Checkerboard Distribution. In contrast to the funnel, the checkerboard distribution (Figure 1) features a compact support. Here, we demonstrate the synergy between our learned latent and an Optimal Transport (OT) coupling. Our method learns a latent that approximates a uniform distribution over the target's support. When this adapted latent is combined with an OT coupling for flow matching, the resulting transport paths are substantially shorter (Figure 8) than those originating from a standard Gaussian, and the vector field training converges much faster (Figure 9). This result underscores our central claim: combining a data-dependent latent with a data-dependent coupling has the potential to significantly improve model performance.

Next we analyze our method on standard image generation benchmarks. In high-dimensional settings and given fixed batch sizes, the signal for the quantile function can be noisy, potentially leading to degenerate solutions. To mitigate this, we add a regularization term to the loss that penalizes the expected negative log-determinant of the Jacobian of the quantile.

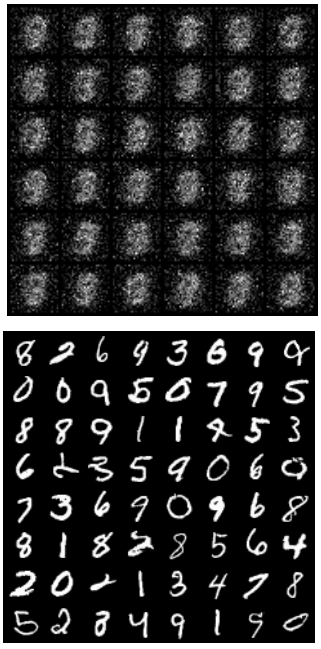
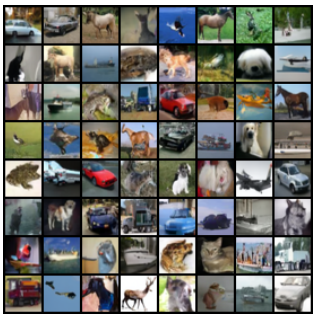


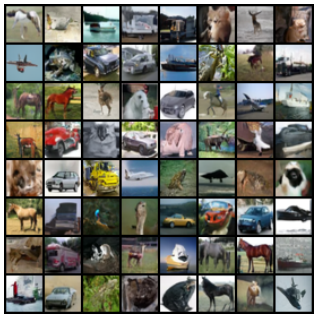
Figure 4: Top: samples from the learned latent. Bottom: generated samples from the learned FM model.

MNIST. The MNIST dataset exhibits strong marginal properties; for instance, pixels near the center are frequently active (non-zero), while pixels at the borders are almost always zero. Our learned quantile function successfully captures these global marginal statistics. As illustrated in Figure 4, the latent distribution learns to concentrate its mass in regions corresponding to active pixels. We also plot mean and standard deviation (Figure 11) as well as empirical and learned quantiles (Figure 12) of our learned latent in the Appendix. While the independence assumption precludes the model from capturing specific spatial correlations (e.g. the shape of a digit), adapting to the correct marginals can provides a improved initialization for the flow model.

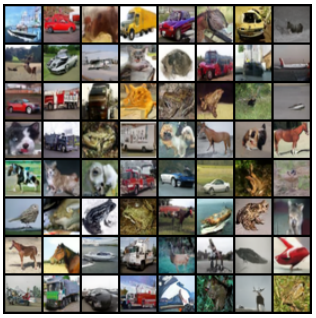
CIFAR-10. To assess the scalability of our approach, we train our model on the CIFAR-10 dataset. The quantile is extremely lightweight compared to the UNet architecture used for the flow model. We reuse the minibatch OT coupling for the latent and freeze the quantile function after a few training epochs. This strategy results in minimal computational overhead compared to the standard Gaussian baseline with minibatch OT coupling. Access to analytic derivatives makes our volume contraction regularization efficient. We evaluate our models for a sufficiently high weight on the quantile loss, we fix it to be $\lambda = 5$. In Figure 5, we report results over different weights β for the regularization parameter. We compared to using a standard gaussian baseline. Our results suggest that for uncorrelated noise, there is a trade-off between the smoothness of the latent and its "closeness" to the data. While out of the scope of this paper, we hypothesize that, for most sampling problems, there is an optimal tradeoff between these properties.



Baseline



$\beta = 3$



$\beta = 4$

	1	2	3	4	Baseline
FID (20 Euler steps)	11.21	10.18	9.97	9.78	10.46
FID (100 Euler steps)	7.22	6.62	6.39	6.26	6.44

Figure 5: CIFAR results for different choices of regularization parameter and for the baseline. The visualized samples were generated using 100 Euler steps.

6 Conclusions

The result of this paper is a “quantile sandbox” for building generative models: a unifying theory and a practical toolkit that turns noise selection into a data-driven design element. Our construction plugs seamlessly into standard objectives including Flow Matching and consistency models, e.g. Inductive Moment Matching. Furthermore our experiments demonstrate that it is possible to learn a freely parametrized, data-dependent latent distribution, beyond the usual smooth transformations of Gaussians. Our work opens several promising directions for future research. Extensions include developing time-dependent quantile functions to optimize the entire path distribution, not just the endpoint as well as designing conditional quantile functions for tasks like class-conditional or text-to-image generation.

REFERENCES

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

L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows*. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel, 2nd edition, 2008. doi: 10.1007/978-3-7643-8722-8.

G. Bartosh, D. Vetrov, and C. A. Naesseth. Neural flow diffusion models: Learnable forward process for improved diffusion modelling, 2025. URL https://arxiv.org/abs/2404.12940.

C. Cattaneo. Sur une forme de l'équation de la chaleur éliminant le paradoxe d'une propagation instantanée. *Comptes Rendus.*, 247, 1958.

R. T. Q. Chen. torchdiffeq, 2018. URL https://github.com/rtqichen/torchdiffeq.

M. Chester. Second sound in solids. *Physical Review*, 131, 1963.

P. Dhariwal and A. Q. Nichol. Diffusion models beat GANs on image synthesis. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=AAWuCvzaVt.

R. Duong, V. Stein, R. Beinert, J. Hertrich, and G. Steidl. Wasserstein gradient flows of MMD functionals with distance kernel and Cauchy problems on quantile functions. *ArXiv:2408.07498*, 2024.

R. Duong, J. Chemseddine, P. Friz, and G. Steidl. Telegrapher's generative model via Kac flows. *arXiv preprint arXiv::2506.20641*, 2025.

C. Durkan, A. Bekasov, I. Murray, and G. Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

R. Flamary, N. Courty, A. Gramfort, M. Z. Alaya, A. Boisbunon, S. Chambon, L. Chapel, A. Corenflos, K. Fatras, N. Fournier, et al. Pot: Python Optimal Transport. *Journal of Machine Learning Research (JMLR)*, 22(1):3571–3578, 2021.

R. Griego and R. Hersh. Theory of random evolutions with applications to partial differential equations. *Transactions of the American Mathematical Society*, 156, 1971.

P. L. Hagemann and S. Neumayer. Stabilizing invertible neural networks using mixture models. *Inverse Problems*, 37(7):085002, 2021.

J. Hertrich, M. Gräf, R. Beinert, and G. Steidl. Wasserstein steepest descent flows of discrepancies with Riesz kernels. *Journal of Mathematical Analysis and Applications*, 531(1):127829, 2024.

M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

A. Janssen. The distance between the Kac process and the Wiener process with applications to generalized telegraph equations. *Journal of Theoretical Probability*, 3(2):349–360, 1990.

M. Kac. A stochastic model related to the telegrapher's equation. *Rocky Mountain Journal of Mathematics*, 4, 1974.

K. Kapusniak, P. Potaptchik, T. Reu, L. Zhang, A. Tong, M. Bronstein, A. J. Bose, and F. Di Giovanni. Metric flow matching for smooth interpolations on the data manifold. *arXiv preprint arXiv:2405.14780*, 2024.

D. Kim, S. Shin, K. Song, W. Kang, and I.-C. Moon. Soft truncation: A universal training technique of score-based diffusion model for high precision score estimation. *ICML*, 2022.

Y. Lipman, R. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. *ICLR*, 2023.

Q. Liu. Rectified flow: A marginal preserving approach to optimal transport. *arXiv preprint arXiv:2209.14577*, 2022.

R. M. Neal. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003. doi: 10.1214/aos/1056562461.

K. Pandey, J. Pathak, Y. Xu, S. Mandt, M. Pritchard, A. Vahdat, and M. Mardani. Heavy-tailed diffusion models, 2024. URL https://arxiv.org/abs/2410.14171.

J. Pidstrigach. Score-based generative models detect manifolds. *NeurIPS*, 2022.

A. Salmona, V. De Bortoli, J. Delon, and A. Desolneux. Can push-forward generative models fit multimodal distributions? *Advances in Neural Information Processing Systems*, 35:0766–10779, 2022.

D. Shariatian, U. Simsekli, and A. O. Durmus. Heavy-tailed diffusion with denoising levy probabilistic models. *ICLR 2025*, 2025.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Y. Song and St. Ermon. Generative modeling by estimating gradients of the data distribution. *ArXiv 1907.05600*, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *ICLR*, 2021.

R. C. Tautz and I. Lerche. Application of the three-dimensional telegraph equation to cosmic-ray transport. *Research in Astronomy and Astrophysics*, 2016.

P. Vernotte. Les paradoxes de la theorie continue de l'équation de la chaleur. *Comptes Rendus.*, 246, 1958.

C. Wald and G. Steidl. Flow Matching: Markov kernels, stochastic processes and transport plans. In *Variational and Information Flows in Machine Learning and Optimal Transport, Oberwolfach Seminars. Vol. 56*, pp. 185–254. Birkhäuser, 2025.

T. Zhang, H. Zheng, J. Yao, X. Wang, M. Zhou, Y. Zhang, and Y. Wang. Long-tailed diffusion models with oriented calibration. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NW2s5XXwXU.

L. Zhou, S. Ermon, and J. Song. Inductive moment matching. *ICML 2025*, 2025.

A EXAMPLES OF ONE-DIMENSIONAL FLOWS

We provide three interesting examples, namely the well-established diffusion flow, the recently proposed Kac flow, and the Wasserstein gradient flow of the MMD functional with negative absolute distance kernel towards a uniform measure. Paths of the processes are depicted in Figure 6.

In each case, the absolutely continuous curve μ_t starting in δ_0 (e.g. conditional) and the corresponding velocity field can be given analytically. Note that in the latter two cases, multi-dimensional generalizations of the flows are not trivially given, which further underlines the strength of our 1D approach. Henceforth, if the measures μ_t admit a density function, we will denote it by p_t.

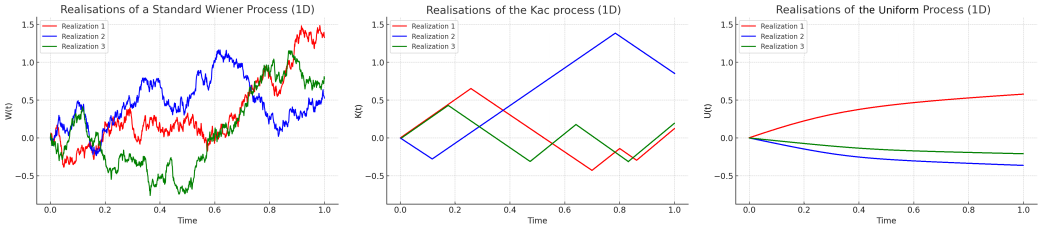


Figure 6: Three realisations of a standard Wiener process (left), the Kac process (middle), and the Uniform process (right), simulated until time $T = 1$.

A.1 WIENER PROCESS AND DIFFUSION EQUATION

First, consider the standard Wiener process (Brownian motion) $(W_t)_t$ starting in 0 whose probability density flow p_t is given by the solution of the diffusion equation

$$\partial_t p_t = \nabla \cdot (p_t \frac{1}{2} \nabla \log p_t) = \frac{1}{2} \Delta p_t, \quad t \in (0, 1], \qquad \lim_{t \downarrow 0} p_t = \delta_0, \tag{14}$$

where the limit for $t \downarrow 0$ is taken in the sense of distributions. The solution is analytically known to be

$$p_t(x) = (2\pi t)^{-\frac{d}{2}} \mathrm{e}^{-\frac{\|x\|^2}{2t}}.$$

Thus, the latent distribution is just the Gaussian $p_1 = \mathcal{N}(0, I_d)$. The velocity field in (14) reads as

$$v_t(x) = -\frac{1}{2} \nabla \log p_t = \frac{x}{2t}. \tag{15}$$

However, its L_2-norm fulfills $\|v_t\|^2_{L_2(\mathbb{R}^1, p_t)} = \frac{d}{4t}$, and is therefore not integrable over time, i.e. $\|v_t\|_{L_2(\mathbb{R}^1, p_t)} \notin L_2(0, 1)$. In practice, instability issues caused by this explosion at times close to the target need to be avoided by e.g. time truncations, see e.g. Kim et al. (2022). For a heuristic analysis also including drift-diffusion flows, we refer to Pidstrigach (2022). Note that in the case of diffusion, there is no significant distinction between the uni- and multivariate setting.

A.2 KAC PROCESS AND DAMPED WAVE EQUATION

The Kac process Kac (1974), also known as persistent random walk, originates from a discrete random walk, which starts in 0 and moves with velocity parameter $c > 0$ in one direction until it reverses its direction with probability $a\Delta_t$, $a > 0$. A continuous-time analogue is given by the Kac process which is defined using the

homogeneous *Poisson point process* N_t with rate a, i.e. i) $N_0 = 0$; ii) the increments of N_t are independent, iii) $N_t - N_s \sim \mathrm{Poi}\big(a(t-s)\big)$ for all $0 \le s < t$. Now the *Kac process starting in* 0 is given by

$$K_t := \mathrm{B}_{\frac{1}{2}}\, c \int_0^t (-1)^{N_s}\, \mathrm{d}s,$$

where $\mathrm{B}_{\frac{1}{2}} \sim \mathrm{Ber}(\frac{1}{2})$ is a Bernoulli random variable[5] taking the values ± 1. Note that in contrast to diffusion processes, the Kac process K_t *persistently* maintains its linear motion between changes of directions (jumps of N_t), see Figure 6.

By the following proposition, the Kac process is related to the damped wave equation, also known as telegrapher's equation, and its probability distribution admits a computable vector field such that the continuity equation is fulfilled. For a proof we refer to Duong et al. (2025).

Proposition 4. *The probability distribution flow of $(K_t)_t$ admits a singular and absolutely continuous part via*

$$\mu_t(x) = \frac{1}{2} e^{-at}\big(\delta_0(x+ct) + \delta_0(x-ct)\big) + \tilde{p}_t(x), \tag{16}$$

with the absolutely continuous part

$$\tilde{p}_t(x) := \frac{1}{2} e^{-at}\Big(\beta ct \frac{I_0'(\beta r_t(x))}{r_t(x)} + \beta I_0(\beta r_t(x))\Big) 1_{[-ct,ct]}(x), \qquad r_t(x) := \sqrt{c^2t^2 - x^2},$$

where $\beta := \frac{a}{c}$, and I_0 denotes the 0*-th* modified Bessel function of first kind*. The distribution* (16) *is the generalized solution of the damped wave equation*

$$\begin{aligned} \partial_{tt} u(t,x) + 2a\, \partial_t u(t,x) &= c^2 \partial_{xx} u(t,x), \\ u(0,x) = \delta_0(x), \qquad \partial_t u(0,x) &= 0. \end{aligned} \tag{17}$$

Further (μ_t, v_t) solves the continuity equation (9) *where the velocity field is analytically given by*

$$v_t(x) := \begin{cases} \frac{x}{t + \frac{r_t(x)}{c} \frac{I_0(\beta r_t(x))}{I_0'(\beta r_t(x))}} & \text{if} \quad x \in (-ct, ct), \\ c & \text{if} \quad x = ct, \\ -c & \text{if} \quad x = -ct, \\ \text{arbitrary} & \text{otherwise}. \end{cases}$$

The Kac velocity field admits the boundedness $\|v_t\|_{L_2(\mathbb{R}^1,\mu_t)} \le c$, and hence, $\|v_t\|_{L_2(\mathbb{R}^1,\mu_t)} \in L_2(0,1)$.

Interestingly, the damped wave equation (17) is closely related to the diffusion equation via Kac' insertion method. It is based on the following theorem, whose proof based on semigroup theory can be found in Griego & Hersh (1971), see also Janssen (1990); Kac (1974).

Theorem 5. *For any initial function $f_0 \in H^2(\mathbb{R}^d)$, $d \ge 1$, let $w_c(t,x)$ be the solution of the* undamped *wave equation with velocity $c > 0$ given by*

$$\begin{aligned} \partial_{tt} w(t,x) = c^2 \Delta w(t,x), \quad x \in \mathbb{R}^d,\ t > 0, \\ w(0,x) = f_0(x), \qquad \partial_t w(0,x) = 0. \end{aligned}$$

Then, the functions defined by

$$h(t,x) := \mathbb{E}\left[w_1\left(\sigma W_t, x\right)\right], \quad \textit{resp.} \quad u(t,x) := \mathbb{E}\big[w_c(c^{-1} S_t, x)\big]$$

[5] More precisely, $\mathrm{B}_{\frac{1}{2}}$ is *two-point* distributed with values $\{-1, 1\}$.

solve the diffusion equation

$$\partial_t h(t,x) = \frac{\sigma^2}{2}\Delta h(t,x), \quad x \in \mathbb{R}^d,\ t>0,$$
$$h(0,x) = f_0(x),$$

resp. the multi-dimensional damped wave equation

$$\partial_{tt} u(t,x) + 2a\,\partial_t u(t,x) = c^2 \Delta u(t,x), \quad x \in \mathbb{R}^d,\ t>0,$$
$$u(0,x) = f_0(x), \qquad \partial_t u(0,x) = 0. \tag{18}$$

As a consequence, it is not hard to show the following corollary, see Duong et al. (2025).

Corollary 6. *For any $t \geq 0$, the solution $u^{a,c}(t,\cdot)$ of the damped wave equation* (18) *converges to the solution $h(t,\cdot)$ of the diffusion equation for $a, c \to \infty$ with fixed $\sigma^2 = \frac{c^2}{a}$.*

In other words, diffusion can be seen as *"an infinitely a-damped wave with infinite propagation speed c"*. Note that the diffusion-related concept of particles traveling with infinite speed violates Einstein's laws of relativity and has therefore found resistance in the physics community Cattaneo (1958); Chester (1963); Vernotte (1958); Tautz & Lerche (2016).

We also like to stress that in multiple dimensions, the damped wave equation (17) is *no longer* mass-conserving as in 1D Tautz & Lerche (2016), and hence eludes a characterization via stochastic processes.

A.3 Uniform process and MMD gradient flow

Wasserstein gradient flows are special absolutely continuous measure flows whose velocity fields are negative Wasserstein (sub-)gradients of functionals $\mathcal{F}_\nu$ on $\mathcal{P}_2(\mathbb{R}^d)$ with the unique minimizer ν. The gradient descent flow should reach this minimizer as $t \to \infty$. In this context, the MMD functional with the non-smooth negative distance kernel

$$\mathcal{F}_\nu(\mu) = \mathrm{MMD}^2(\mu,\nu) \coloneqq -\frac{1}{2}\int_{\mathbb{R}^2} |x-y| \,\mathrm{d}\,(\mu(x)-\nu(x))\,\mathrm{d}\,(\mu(y)-\nu(y)) \tag{19}$$

stands out for its flexible flow behavior between distributions of different support Hertrich et al. (2024). In 1D, its Wasserstein gradient flow μ_t can be equivalently described by the flow of its quantile functions Q_{μ_t} with respect to an associated functional on $L_2(0,1)$. Note that the MMD functional (19) loses its convexity (along generalized geodesics) in multiple dimensions Hertrich et al. (2024), and the general existence of their Wasserstein gradient flows is unclear in the multivariate case. This yields another reason to work in 1D, where we have have the following proposition.

Proposition 7. *The Wasserstein gradient flow μ_t of the MMD functional* (19) *starting in $\mu_0 = \delta_0$ towards the uniform distribution $\nu = \mathcal{U}[-b,b]$ with fixed $b>0$ reads as*

$$\mu_t = \left(1-\exp(-\tfrac{t}{b})\right)\mathcal{U}[-b,b],, \qquad t>0, \tag{20}$$

with corresponding velocity field

$$v_t(x) = \frac{x}{b\left(\exp\left(\frac{t}{b}\right)-1\right)}, \quad x \in \mathrm{supp}(\mu_t). \tag{21}$$

It holds $\|v_t\|^2_{L_2(\mathbb{R}^1,\mu_t)} = \frac{2b}{3}\exp(-\frac{2t}{b})$, and hence, $\|v_t\|_{L_2(\mathbb{R}^1,\mu_t)} \in L_2(0,1)$. A corresponding (stochastic) process $(U_t)_t$ is given by $U_t \coloneqq b\left(1-\exp\left(-\frac{t}{b}\right)\right)U$, where $U \sim \mathcal{U}[-1,1]$, such that $\mathrm{Law}(U_t) = \mu_t$.

We prove the proposition more general for $\nu = \mathcal{U}[a,b]$ and a flow starting in $x_0 \in [a,b]$, i.e. we show

$$\mu_t = \mathcal{U}\left[a + (x_0 - a)\exp\left(-r(t)\right), b - (b - x_0)\exp\left(-r(t)\right)\right], \qquad t > 0 \tag{22}$$

with $r(t) \coloneqq \frac{2t}{b-a}$ and

$$v_t(x) = \frac{2}{b-a}\left(\frac{x - x_0}{\exp(r(t)) - 1}\right). \tag{23}$$

To this end, we need the relation between measures in $\mathcal{P}_2(\mathbb{R})$ and cumulative distribution functions, see (10). For $\nu = \mathcal{U}[a,b]$, we have that

$$r_\nu(x) = \begin{cases} 0, & \text{if } x < a, \\ \frac{x-a}{b-a}, & \text{if } a \le x \le b, \\ 1, & \text{if } x > b \end{cases}$$

and $Q_\nu(s) = a(1-s) + bs$. In Hertrich et al. (2024) it was shown that the functional $F_\nu \colon L_2(0,1) \to \mathbb{R}$ defined by

$$F_\nu(u) \coloneqq \int_0^1 \left((1-2s)\big(u(s) + Q_\nu(s)\big) + \int_0^1 |u(s) - Q_\nu(t)|\,\mathrm{d}t\right)\mathrm{d}s \tag{24}$$

fulfills $\mathcal{F}_\nu(\mu) = F_\nu(Q_\mu)$ for all $\mu \in \mathcal{P}_2(\mathbb{R})$. Moreover, we have the following equivalent characterization of Wasserstein gradient flows of $\mathcal{F}_\nu$, which can be found in (Duong et al., 2024, Theorem 4.5).

Theorem 8. *Let $\mathcal{F}_\nu$ and F_ν be defined by* (19) *and* (24)*, respectively. Then the Cauchy problem*

$$\begin{cases} \partial_t g(t) \in -\partial F_\nu(g(t)), & t \in (0,\infty), \\ g(0) = Q_{\mu_0}, \end{cases}$$

has a unique strong solution g, and the associated curve $\gamma_t \coloneqq (g(t))_\# \Lambda_{(0,1)}$ is the unique Wasserstein gradient flow of $\mathcal{F}_\nu$ with $\gamma(0+) = (Q_{\mu_0})_\# \Lambda_{(0,1)}$. More precisely, there exists a velocity field v_t^* such that (γ_t, v_t^*) satisfies the continuity equation (9)*, and it holds the relations*

$$v_t^* \circ g(t) \in -\partial F_\nu(g(t)) \quad \textit{and} \quad v_t^* \in -\partial\, \mathcal{F}_\nu(\gamma_t). \tag{25}$$

Lastly note that here, the subdifferential $\partial F_\nu(u)$ is explicitly given by the singleton

$$-\partial F_\nu(u) = -\nabla F_\nu(u) = 2(\cdot - r_\nu \circ u) \quad \text{for all } u \in L_2(0,1),$$

see (Duong et al., 2024, Lemma 4.3).

Proof of Proposition 7. We want to apply Theorem 8 to (μ_t, v_t) in (22) and (23). The uniform distribution in (22) has the quantile function

$$Q_{\mu_t}(s) = \big(1 - \exp\left(-r(t)\right)\big)\big(a + (b-a)s\big) + x_0 \exp\left(-r(t)\right), \qquad s \in (0,1).$$

For all $t > 0$ and all $s \in (0,1)$, we have $Q_{\mu_t}(s) \in [a,b]$ since $x_0 \in [a,b]$, and thus

$$\begin{aligned} -\nabla F_\nu(Q_{\mu_t})(s) &= 2s - 2r_\nu(Q_{\mu_t}(s)) \\ &= 2s - 2\frac{\big(1 - \exp\left(-r(t)\right)\big)\big(a + (b-a)s\big) + x_0\exp\left(-r(t)\right) - a}{b-a} \\ &= 2\left(s - \frac{x_0 - a}{b-a}\right)\exp\left(-r(t)\right). \end{aligned}$$

On the other hand, it holds

$$\partial_t Q_{\mu_t}(s) = -2\frac{x_0-a}{b-a}\exp\left(-r(t)\right) - \frac{(-2)(b-a)s}{b-a}\exp\left(-r(t)\right) = 2\left(s - \frac{x_0-a}{b-a}\right)\exp\left(-r(t)\right).$$

By Theorem 8, (μ_t) is the unique Wasserstein gradient flow of $\mathcal{F}_\nu$ starting in δ_0.

Furthermore, there exists a velocity field v_t^* satisfying the continuity equation (9) and the relations (25). For $s \in (0,1)$ and $t > 0$, let $y \coloneqq g_s(t) = a + (x_0 - a)\exp\left(-r(t)\right) + (b-a)\left(1 - \exp\left(-r(t)\right)\right)s$. Then, we have $s = \frac{y-a-(x_0-a)\exp(-r(t))}{(b-a)(1-\exp(-r(t)))}$, and thus by (25),

$$\begin{aligned}
v_t^*(y) = v_t^*(Q_{\mu_t}(s)) &= 2\left(s - \frac{x_0-a}{b-a}\right)\exp\left(-r(t)\right) \\
&= 2\left(\frac{y-a-(x_0-a)\exp\left(-r(t)\right)}{(b-a)\left(1-\exp\left(-r(t)\right)\right)} - \frac{x_0-a}{b-a}\right)\exp\left(-r(t)\right) \\
&= \frac{2}{b-a}\left(\frac{y-a-(x_0-a)}{1-\exp\left(-r(t)\right)}\right)\exp\left(-r(t)\right) \\
&= \frac{2}{b-a}\left(\frac{y-x_0}{\exp\left(r(t)\right)-1}\right)
\end{aligned}$$

for all $y \in g_s(t)(0,1) = [a + (x_0-a)\exp(-r(t)), b - (b-x_0)\exp\left(-r(t)\right)]$. Lastly, let us compute the action. For $t > 0$ we have

$$\begin{aligned}
\|v_t\|^2_{L^2(\mu_t)} &= \int_{a+(x_0-a)\exp\left(-\frac{2t}{b-a}\right)}^{b-(b-x_0)\exp\left(-\frac{2t}{b-a}\right)} \frac{4(x-x_0)^2}{(b-a)^2\left(\exp\left(\frac{2t}{b-a}\right)-1\right)^2}\frac{1}{(b-a)\left(1-\exp\left(-\frac{2t}{b-a}\right)\right)}\,\mathrm{d}x \\
&= \frac{4}{(b-a)^3\left(\exp\left(\frac{2t}{b-a}\right)-1\right)^2\left(1-\exp\left(-\frac{2t}{b-a}\right)\right)}\int_{a+(x_0-a)\exp\left(-\frac{2t}{b-a}\right)}^{b-(b-x_0)\exp\left(-\frac{2t}{b-a}\right)} (x-x_0)^2\,\mathrm{d}x \\
&= \frac{4}{(b-a)^2\exp\left(-\frac{2t}{b-a}\right)\left(\exp\left(\frac{2t}{b-a}\right)-1\right)^3}\left[\frac{(x-x_0)^3}{3}\right]_{a+(x_0-a)\exp\left(-\frac{2t}{b-a}\right)}^{b-(b-x_0)\exp\left(-\frac{2t}{b-a}\right)} \\
&= \frac{4\left(1-\exp\left(-\frac{2t}{b-a}\right)\right)^3}{3(b-a)^2\exp\left(-\frac{2t}{b-a}\right)\left(\exp\left(\frac{2t}{b-a}\right)-1\right)^3}\left[(b-x_0)^3-(a-x_0)^3\right] \\
&= \frac{4\left[(b-x_0)^3-(a-x_0)^3\right]}{3(b-a)^2}\exp\left(-\frac{4t}{b-a}\right).
\end{aligned}$$

and the proof is finished. □

Note that the fact that v_t^* is uniquely determined on $\operatorname{supp}\mu_t = \overline{g_t(0,1)}$, correlates with the fact that the gradient $v_t^* \circ g(t) = -\nabla F_\nu(g(t))$ is a *singleton*. Outside of $\operatorname{supp}\mu_t$, the velocity field may be arbitrarily extended, which yields a velocity $\tilde{v}_t \in -\partial\mathcal{F}_\nu(\mu_t)$ in a *non-singleton* subdifferential. The velocity v_t^* may be *uniquely* chosen from the tangent space $T_{\mu_t}\mathcal{P}_2(\mathbb{R})$, or equivalently, by choosing it to have minimal norm, i.e. $v_t^* \equiv 0$ outside of $\operatorname{supp}\mu_t$.

B FLOW MATCHING AS SPECIAL MEAN REVERTING PROCESSES

B.1 THE GAUSSIAN CASE

Let us shortly verify that our componentwise approach using the mean-reverting process (4), i.e.

$$\mathbf{X}_t \;\coloneqq\; f(t)\,\mathbf{X_0} \;+\; \mathbf{Y}_{g(t)},$$

leads to the usual flow matching objective. where we choose the scheduling functions $f(t) \coloneqq 1-t$, $g(t) \coloneqq t^2$, the target random variable $\mathbf{X}_0 \sim \mu_0$, and a standard Wiener process $\mathbf{Y}_t$ in $\mathbb{R}^d$ (independent of $\mathbf{X}_0$): First, it holds $\mathbf{Y}_{t^2} \sim \mathcal{N}(0, t^2 I_d)$, hence $\mathbf{Y}_{t^2} \overset{d}{=} t\,\mathbf{Z}$ with $\mathbf{Z} \sim \mathcal{N}(0, I_d)$, so that

$$\mathbf{X}_t \overset{d}{=} (1-t)\mathbf{X_0} + t\,\mathbf{Z}.$$

Furthermore, by (15) the 1D components of $\mathbf{Y}_t$ admit the velocity field $v_t^i(x^i) = \frac{x^i}{2t}$, $x^i \in \mathbb{R}$, and by Proposition 2 the multi-dimensional process $\mathbf{Y}_t$ admits the velocity field $v_{\mathbf{Y}}(t,x) = (\frac{x^1}{2t}, ..., \frac{x^d}{2t}) = \frac{x}{2t}$, $x = (x^1, ..., x^d) \in \mathbb{R}^d$. By the calculation (6), the conditional velocity field corresponding to $\mathbf{X}_t$ starting in $x_0 \in \mathbb{R}^d$ reads as

$$\begin{aligned} v_{\mathbf{X}}(t, x \mid x_0) &= \dot{f}(t)\,x_0 \;+\; \dot{g}(t)\,v_{\mathbf{Y}}\big(g(t),\, x - f(t)x_0 \mid 0\big) \\ &= -x_0 \;+\; 2t\,v_{\mathbf{Y}}\big(t^2,\, x - (1-t)x_0 \mid 0\big) \\ &= -x_0 \;+\; \frac{x-(1-t)x_0}{t}. \end{aligned}$$

Now, if $x \sim P_{\mathbf{X}_t}(\cdot \mid x_0)$, i.e. $x = (1-t)x_0 + tz$ with $z \sim \mathcal{N}(0, I_d)$, then it follows

$$v_{\mathbf{X}}(t, x \mid x_0) = -x_0 \;+\; \frac{(1-t)x_0 + tz - (1-t)x_0}{t} = z - x_0, \tag{26}$$

which is the usual constant-in-time conditional FM velocity along the straight-line trajectories between $x_0 \sim \mu_0$ and $z \sim \mathcal{N}(0, I_d)$.

B.2 THE UNIFORM CASE

Now consider any component of the mean-reverting process (4) with $f(t), g(t)$ to be chosen, X_0 being a component of $\mathbf{X}_0 \sim \mu_0$, and Y_t given by the MMD gradient flow (20), i.e. $Y_t \coloneqq b\left(1 - \exp\left(-\frac{t}{b}\right)\right) U$, where $U \sim \mathcal{U}[-1,1]$. Let v_Y be the corresponding velocity field from (21). Then, we have

$$\begin{aligned} v_X(t, x|x_0) &= \dot{f}(t)\,x_0 \;+\; \dot{g}(t)\,v_Y\big(g(t),\, |x - f(t)x_0|\big)\,\frac{x - f(t)x_0}{|x - f(t)x_0|} \\ &= \dot{f}(t)\,x_0 \;+\; \dot{g}(t)\,\frac{x - f(t)x_0}{b\left(\exp\left(\frac{g(t)}{b}\right) - 1\right)}. \end{aligned}$$

Now, along the trajectory $x \sim P_{X_t}(\cdot \mid x_0)$, i.e.

$$x \;=\; f(t)\,x_0 + b\left(1 - \exp\left(-\frac{g(t)}{b}\right)\right) u \;=:\; \alpha_t\,x_0 + \sigma_t\,u, \tag{27}$$

with $u \sim \mathcal{U}(-1,1)$, the velocity calculates as

$$\begin{aligned} v_X(t,x \mid x_0) &= \dot{f}(t)\, x_0 \;+\; \dot{g}(t)\, \frac{b\left(1-\exp\left(-\frac{g(t)}{b}\right)\right) u}{b\left(\exp\left(\frac{g(t)}{b}\right)-1\right)} \\ &= \dot{f}(t)\, x_0 \;+\; \dot{g}(t) \exp\left(\frac{-g(t)}{b}\right) u \\ &= \dot{\alpha}_t\, x_0 + \dot{\sigma}_t\, u, \end{aligned} \tag{28}$$

where $\alpha_t \coloneqq f(t)$ and $\sigma_t \coloneqq b\left(1-\exp\left(-\frac{g(t)}{b}\right)\right)$. Hence, in order to minimize the CFM loss, we only need to sample $t \sim \mathcal{U}[0,1]$, $x_0 \sim X_0$, and $u \sim \mathcal{U}(-1,1)$. Note the similarity between the MMD path (27) and the FM/diffusion path (7); by choosing $b = 1$, $f(t) \coloneqq 1-t$ and $g(t) \coloneqq -\log(1-t)$ it follows $\alpha(t) = 1-t$, $\sigma(t) = t$, and we obtain in (28) the FM-velocity along the trajectory (26), where the Gaussian noise $z \sim \mathcal{N}(0,1)$ is just replaced by a uniform noise $u \sim \mathcal{U}(-1,1)$.

C IMM WITH QUANTILE INTERPOLANTS

In this section, we want to demonstrate how the IMM framework proposed in Zhou et al. (2025) can be realized by our quantile approach. Note that in the following – for notational simplicity – we consider the one-dimensional case $X_0, Z_t \in \mathbb{R}$ where we can employ quantile functions. By combining the 1D components into a multivariate model $\mathbf{X}_0 = (X_0^1, ..., X_0^d)$, $\mathbf{Z}_t = (Z_t^1, ..., Z_t^d)$, the results of this chapter trivially extend to $\mathbb{R}^d$.

Recall our definition of the *quantile process*

$$Z_t = f(t)X_0 + Q_{g(t)}(U), \quad U \sim \mathcal{U}(0,1),\, t \in [0,1]. \tag{29}$$

and the *quantile interpolants*

$$I_{s,t}(x,y) = f(s)x + Q_{g(s)}\big(R_{g(t)}(y - f(t)x)\big), \quad s,t \in [0,1]. \tag{30}$$

Note that by the assumptions (5) it holds $Z_0 = X_0$ and $Z_1 = Q_1(U)$.

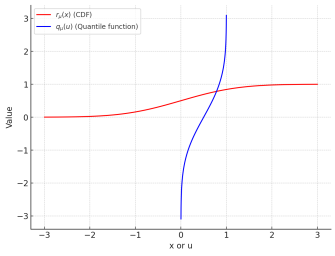


Figure 7: The CDF R_μ and quantile function Q_μ of a standard normal distribution μ.

By the following remark, our quantile interpolants generalize the interpolants used in Denoising Diffusion Implicit Models (DDIM).

Remark 9 (Relation to DDIM)**.** *The interpolants used in Denoising Diffusion Implicit Models (DDIMs) Song et al. (2020) are given by*

$$\mathrm{DDIM}_{s,t}(x,y) \coloneqq \left(\alpha_s - \frac{\sigma_s}{\sigma_t}\alpha_t\right)x + \frac{\sigma_s}{\sigma_t}y. \tag{31}$$

Now let $f(t) \coloneqq 1-t$, $g(t) \coloneqq t^2$ and let Q_t be the quantile of the law of a standard Brownian motion W_t. First we obtain

$$Q_{g(t)}(p) = Q_{t^2}(p) = Q_{\mathcal{N}(0,t^2)}(p) = t\sqrt{2}\,\mathrm{erf}^{-1}(2p-1) = t\,Q_{\mathcal{N}(0,1)}(p), \quad p \in (0,1),$$

with the error function erf. *Hence,* (29) *exactly becomes (not only in distribution)*

$$Z_t = (1-t)Y_0 + t\,Q_{\mathcal{N}(0,1)}(U) = (1-t)Y_0 + tZ,$$

where $Z \coloneqq Q_{\mathcal{N}(0,1)}(U) \sim \mathcal{N}(0,1)$, i.e. the components of (7) *with the choice $\alpha_t = 1-t$, $\sigma_t = t$. Furthermore, since $R_{t^2}(z) = R_{\mathcal{N}(0,t^2)}(z) = \frac{1}{2}(1+\mathrm{erf}\left(\frac{z}{t\sqrt{2}}\right))$, the quantile interpolant* (12) *reads as*

$$\begin{aligned} I_{s,t}(x,y) &= (1-s)x + s\sqrt{2}\,\mathrm{erf}^{-1}\left(\mathrm{erf}\left(\frac{y-(1-t)x}{t\sqrt{2}}\right)\right) = (1-s)x + \frac{s}{t}(y-(1-t)x) \\ &= ((1-s) - \frac{s}{t}(1-t))x + \frac{s}{t}y. \end{aligned}$$

which is exactly $\mathrm{DDIM}_{s,t}(x,y)$ *in* (31) *with $\alpha_t = f(t)$ and $\sigma_t^2 = g(t)$.* $\diamond$

Exactly as the DDIM interpolants, our quantile interpolants (30) satisfy the following crucial interpolation properties.

Proposition 10 (a.k.a Proposition 3)**.** *For all $x, y \in \mathbb{R}$ and all $s, r, t \in [0,1]$, it holds*

$$I_{0,t}(x,y) = x, \quad I_{t,t}(x,y) = y, \tag{32}$$

and

$$I_{s,r}(x, I_{r,t}(x,y)) = I_{s,t}(x,y).$$

Furthermore, inserting the quantile process (11) *yields*

$$I_{s,t}(Z_0, Z_t) = Z_s. \tag{33}$$

Proof. By assumptions it holds

$$I_{0,t}(x,y) = f(0)x + Q_{g(0)}\big(R_{g(t)}(y - f(t)x)\big) = x,$$

and

$$I_{t,t}(x,y) = f(t)x + Q_{g(t)}\big(R_{g(t)}(y - f(t)x)\big) = y.$$

Furthermore, it holds the interpolation/consistency property

$$\begin{aligned} I_{s,r}(x, I_{r,t}(x,y)) &= f(s)x + Q_{g(s)}\big(R_{g(r)}(I_{r,t}(x,y) - f(r)x)\big) \\ &= f(s)x + Q_{g(s)}\big(\cancel{R_{g(r)}}(\cancel{f(r)x} + \cancel{Q_{g(r)}}\big(R_{g(t)}(y - f(t)x)\big) - \cancel{f(r)x})\big) \\ &= f(s)x + Q_{g(s)}\big(R_{g(t)}(y - f(t)x)\big) \\ &= I_{s,t}(x,y) \end{aligned}$$

for all $x, y \in \mathbb{R}$. Also note that inserting the random variables Z_0, Z_t yields

$$\begin{aligned} I_{s,t}(Z_0, Z_t) &= f(s)Z_0 + Q_{g(s)}\big(R_{g(t)}\big(Z_t - f(t)Z_0\big)\big) \\ &= f(s)Z_0 + Q_{g(s)}\big(U\big) \\ &= Z_s. \end{aligned}$$

This finishes the proof. □

Proposition 10 represents the key observation which allows us to utilize our quantile process (29) in the IMM framework the same way as Zhou et al. (2025) employ the DDIM interpolants (31):

For this, let us now recall the basic idea of inductive moment matching and the corresponding loss functions. Let us distinguish between real numbers written in small letters ($x_0, u, z_t \in \mathbb{R}$) and random variables written with capital letters ($X_0, U, Z_t, \ldots$). We assume that the probability distributions have densities:

$\mathrm{Law}(X_0)$	$\mathrm{Law}(Z_t)$	$\mathrm{Law}(Z_s \mid X_0 = x_0, Z_t = z_t)$	$\mathrm{Law}(Z_t \mid X_0 = x_0, U = u)$	$\mathrm{Law}(X_0 \mid Z_t = z_t)$
$\rho_0(x_0)$	$\rho_t(z_t)$	$\rho_{s\|0,t}(z_s \mid x_0, z_t)$	$\rho_{t\|0,1}(z_t \mid x_0, u)$	$\rho_{0\|t}(x_0 \mid z_t)$

Note that by (33) we have $\rho_{s|0,t}(z_s|x_0, z_t) = \mathrm{Law}(I_{s,t}(x_0, z_t))(z_s) = \delta(z_s - I_{s,t}(x_0, z_t))$, hence sampling from $\rho_{s|0,t}(z_s|x_0, z_t)$ is just applying $I_{s,t}(x_0, z_t)$. Similarly, sampling from $\rho_{t|0,1}(z_t|x_0, u)$ is just evaluating $I_{t,1}(x_0, Q_1(u))$.

The following proposition follows directly from Proposition 10 as in Zhou et al. (2025). It is essential for deriving the appropriate loss functions.

Proposition 11. *For all* $0 \le s \le r \le t \le 1$*, the quantile interpolant* (30) *is self-consistent, i.e.*

$$\rho_{s|0,t}(z_s|x_0, z_t) = \int_{\mathbb{R}} \rho_{s|0,r}(z_s|x_0, z_r)\, \rho_{r|0,t}(z_r|x_0, z_t)\, \mathrm{d}z_r,$$

and the quantile process (29) *is marginal preserving, i.e.*

$$\rho_s(z_s) = \mathbb{E}_{z_t \sim \rho_t, x_0 \sim \rho_{0|t}(\cdot|z_t)} \left[\rho_{s|0,t}(z_s|x_0, z_t)\right].$$

Learning. The conditional probability $\rho_{0|t}(\cdot|z_t)$ is now approximated by a network p^θ_{s,t,z_t} where the parameter s describes the dependence on ρ_s such that

$$\rho_s \approx \mathbb{E}_{z_t \sim \rho_t, x_0 \sim p^\theta_{s,t,z_t}} \left[\rho_{s|0,t}(\cdot|x_0, z_t)\right] =: p^\theta(s, t). \tag{34}$$

Then it is proposed in (Zhou et al., 2025, Eq. (7)) to minimize the so-called *naïve objective*

$$\mathcal{L}_{\mathrm{naive}}(\theta) := \mathbb{E}_{s,t}\big[D(\rho_s, p^\theta(s, t)\big], \tag{35}$$

with an appropriate metric D, e.g. MMD. The procedure is now as follows: starting in a sample x_0 from X_0, we can sample z_s, z_t from Z_s, Z_t by (29), respectively; then given z_t we sample $\tilde{x}_0$ from p^θ_{s,t,z_t}, and finally we can evaluate $\tilde{z}_s = I(\tilde{x}_0, z_t)$ from (33), which is then compared with z_s.

Inference. The following iterative multi-step sampling can be applied: for chosen decreasing $t_k \in (0, 1]$, $k = 0, \ldots, T$ with $t_0 = 1$, starting with $x_0^{(0)} \sim p^\theta_{0,1,z_1}$, we compute

$$z_{t_k} = I_{t_k, t_{k-1}}\left(x_0^{(k-1)}, z_{t_{k-1}}\right), \quad x_0^{(k)} \sim p^\theta_{0,t_k,z_{t_k}}, \quad k = 1, \ldots, T.$$

Although for marginal-preserving interpolants, a minimizer of $\mathcal{L}_{\mathrm{naive}}$ exists with minimum 0, the authors of Zhou et al. (2025) object that directly optimizing (35) faces practical difficulties when t is far away from s. Instead, they propose to apply the following "inductive bootstrapping" technique:

Bootstrapping. Instead of minimizing (35), we consider the *general objective*

$$\mathcal{L}_{\mathrm{general}}(\theta) := \mathbb{E}_{s,t}\left[w(s, t)\mathrm{MMD}^2(p^{\theta_{n-1}}(s, r), p^{\theta_n}(s, t))\right], \tag{36}$$

with a weighting function $w(s, t)$ to be chosen. The kernel of the squared MMD distance can be chosen as e.g. the (time-dependent) Laplace kernel. Importantly, the value r is chosen to be a function $r = r_{s,t} \in [s, t]$ being "close to t" and fulfilling a suitable monotonicity property.

Let us assume the simplest case $r_{s,t} \coloneqq \max\{s, t-\varepsilon\}$ with a small fixed $\varepsilon > 0$ and hereby demonstrate the bootstrapping technique: Fix $s \in [0,1]$. Then, it holds for all $t \in [s, s+\varepsilon]$ that $r_{s,s} = s$. By the definition (34) and property (32), it holds (independently of θ) that $p^\theta(s,s)(z_s) = \rho_s(z_s)$. Hence, minimizing (36) in the first step $n = 1$ yields

$$0 = \mathrm{MMD}^2(p^{\theta_0}(s,s), p^{\theta_1}(s,t_1)) = \mathrm{MMD}^2(\rho_s, p^{\theta_1}(s,t_1)) \quad \text{for all } t_1 \in [s, s+\varepsilon].$$

In the second step $n = 2$, it holds for all $t_2 \in [s, s+2\varepsilon]$ that $r_{s,t_2} \in [s, s+\varepsilon]$. Hence, minimizing (36) in the second step yields, together with the first step,

$$0 = \mathrm{MMD}^2(p^{\theta_1}(s, r_{s,t_2}),\ p^{\theta_2}(s,t_2)) = \mathrm{MMD}^2(\rho_s, p^{\theta_2}(s,t_2)) \quad \text{for all } t_2 \in [s, s+2\varepsilon].$$

Thus, for the number of steps $n \to \infty$, it holds $0 = \mathrm{MMD}^2(\rho_s, p^{\theta_n}(s,t_n))$ even for the entire interval $t_n \in [s,1]$. Hence, minimizing the general objective (36) with a large number of steps eventually minimizes the naïve objective (35), see (Zhou et al., 2025, Theorem 1) for more details.

D Learn The Noise

D.1 Counterexample: Marginal Product

For the Wasserstein distance, the minimizer is not necessarily the product measure with the correct marginals. For the measure

$$\mu = \tfrac{1}{2}\delta_{(1,1)} + \tfrac{1}{2}\delta_{(-1,-1)} \in \mathcal{P}_2(\mathbb{R}^2),$$

the product measure with the correct marginals is given by

$$\mu_{\text{marg}} = \left(\tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_1\right) \times \left(\tfrac{1}{2}\delta_{-1} + \tfrac{1}{2}\delta_1\right).$$

Then the Wasserstein distance is

$$W_2^2(\mu, \mu_{\text{marg}}) = 2,$$

but the product measure

$$\nu_\alpha = \left(\tfrac{1}{2}\delta_{-\alpha} + \tfrac{1}{2}\delta_\alpha\right) \times \left(\tfrac{1}{2}\delta_{-\alpha} + \tfrac{1}{2}\delta_\alpha\right)$$

is, for $\alpha = 0.5$, closer in the Wasserstein distance:

$$W_2^2(\mu, \nu_\alpha) = 2(1 - \alpha + \alpha^2) = 1.5.$$

D.2 Details on the architecture of the learned quantiles Q_ϕ

We implement the quantile transport with rational quadratic splines (RQS) Durkan et al. (2019). For each coordinate i the map takes the form

$$Q_\phi^i(u) = H_\phi^i\big(\mathrm{logit}(u)\big), \qquad u \in (0,1),$$

where $H_\phi^i : \mathbb{R} \to \mathbb{R}$ is a monotone spline with K bins and linear tails. A lightweight conditioner network outputs bin widths, bin heights, and knot slopes; we pass these raw values through softplus, add a small constant to the slopes, and normalise widths/heights so they sum to one. The positive lower bound on each slope ensures that H_ϕ^i is strictly increasing, hence Q_ϕ^i is strictly increasing on $(0,1)$. The derivative exists almost everywhere and satisfies

$$(Q_\phi^i)'(u) = H_\phi^{i\,'}\big(\mathrm{logit}(u)\big)\,\frac{1}{u(1-u)} > 0 \quad \text{for a.e. } u \in (0,1).$$

D.3 Toy Target Distributions

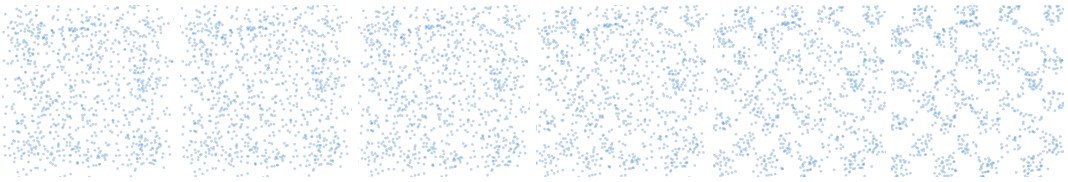


Figure 8: A generated sample path from the learned quantile latent to the checkerboard. The adapted latent (left) is already close to the target distribution.

We use three standard challenging low-dimensional distributions: Neal's funnel, a 3×3 Gaussian mixture, and a checkerboard.

Funnel. For the toy illustration in Figure 2, we work with the dataset known as Neals Funnel (Neal, 2003). The distribution of Neal's funnel is defined as follows:

$$p(x_1, x_2) \;=\; \mathcal{N}\big(x_1; 0, 3\big)\, \mathcal{N}\big(x_2; 0, \exp(x_1/2)\big).$$

Grid Gaussian Mixture. We give more details about the mixture of Gaussian we consider in our experiment. It is designed in a grid pattern in $[-1, 1]^2$, as follows:

$$\sum_{i=1}^{9} w_i \cdot \mathcal{N}(\mu_i, \sigma^2 I_2)\,,$$

where $(w_i)_{i=1}^9 = (0.01,\ 0.1,\ 0.3,\ 0.2,\ 0.02,\ 0.15,\ 0.02,\ 0.15,\ 0.05)$, $\mu_i = (\mu_1, \mu_2)$ with $\mu_1 = (i \bmod 3) - 1$, $\mu_2 = \left\lfloor \frac{i}{3} \right\rfloor - 1$, and $\sigma = 0.025$.

Checkerboard. Fix $\ell < h$ and domain $\Omega = [\ell, h]^2$. Define the support

$$\mathcal{S} = \big\{(x, y) \in \Omega :\ \lfloor x \rfloor + \lfloor y \rfloor \text{ is even}\big\}.$$

The checkerboard distribution is uniform on $\mathcal{S}$ and zero elsewhere:

$$p_{\text{Checker}}(x, y) = \begin{cases} \dfrac{1}{\text{area}(\mathcal{S})}, & (x, y) \in \mathcal{S}, \\ 0, & \text{otherwise.} \end{cases}$$

For integer ℓ, h with even side length (e.g. $\ell = -4, h = 4$), exactly half of Ω is active, hence

$$p_{\text{Checker}}(x, y) = \frac{2}{(h - \ell)^2}\, \mathbf{1}_{\mathcal{S}}(x, y).$$

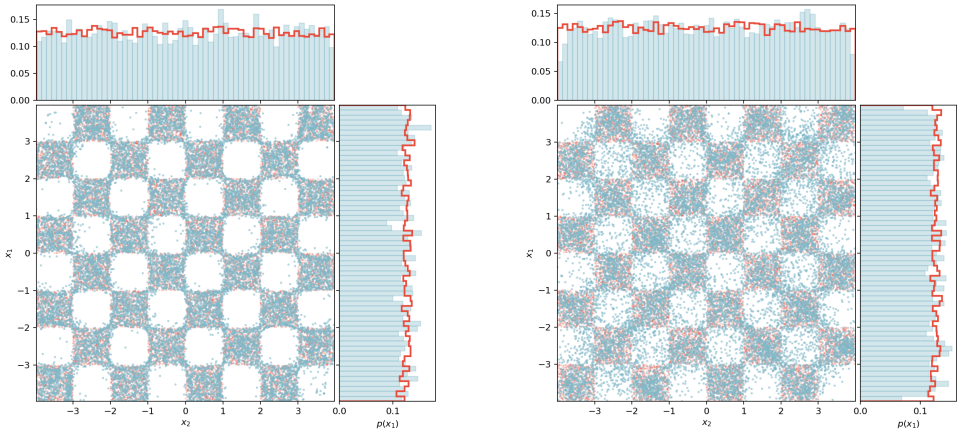


Figure 9: Comparing the generated samples due to our learned latent (left) to those due to a Gaussian latent (right), after 20k steps of training with the same network architecture. Our model converges much faster.

D.4 Minibatch optimal transport

Since the learned latent distribution is close to the data distribution, we can exploit this improved matching via an optimal transport coupling. For training, the minibatch OT is computed empirically as follows: draw a minibatch $\{\mathbf{x}_0^{(i)}\}_{i=1}^B \sim \mu_0$ and $\{\mathbf{u}^{(j)}\}_{j=1}^B \sim \mathcal{U}([0,1]^d)$, set $\mathbf{y}^{(j)} = \mathbf{Q}_\phi(\mathbf{u}^{(j)})$, and define the empirical measures

$$\hat{\mu}_0^B = \tfrac{1}{B}\sum_{i=1}^B \delta_{\mathbf{x}_0^{(i)}}, \qquad \hat{\nu}_\phi^B = \tfrac{1}{B}\sum_{j=1}^B \delta_{\mathbf{y}^{(j)}}.$$

The minibatch objective is

$$\widehat{\mathcal{E}}_{\mathsf{Q}}(\phi) \;=\; D\big(\hat{\mu}_0^B, \hat{\nu}_\phi^B\big),$$

and gradients backpropagate through $\mathbf{y}^{(j)} = \mathbf{Q}_\phi(\mathbf{u}^{(j)})$.

Furthermore, we use the linear path $\mathbf{x}_t^{(j)} = (1-t_j)\mathbf{x}_0^{(j)} + t_j\,\mathbf{y}^{(T(j))}$, $j = 1,\dots,B$, with $t_j \sim \mathcal{U}(0,1)$, the target velocity $\mathbf{y}^{(\pi(j))} - \mathbf{x}_0^{(j)}$, and we optimize the empirical versions

$$\widehat{\mathcal{E}}_{\mathsf{FM}}(\theta;\phi) \;=\; \frac{1}{B}\sum_{j=1}^B \big\| v_\theta\big(\mathbf{x}_t^{(j)}, t_j\big) - \big(\mathbf{y}_\phi^{(T(j))} - \mathbf{x}_0^{(j)}\big)\big\|_2^2, \quad \widehat{\mathcal{L}}_{\mathsf{joint}} = \widehat{\mathcal{E}}_{\mathsf{FM}} + \lambda_Q\,\widehat{\mathcal{E}}_Q.$$

E Implementation Details

Algorithm 1 Joint learning of 1D quantiles and FM velocity

Require: Dataset $\mathcal{D}$, batch size B, weight λ_Q, iterations K
Require: Quantile model $\mathbf{Q}_\phi$, velocity model v_θ
for $k = 1$ to K **do**
  Sample $\{\mathbf{x}_i\}_{i=1}^B \sim \mathcal{D}$, $\{\mathbf{u}_i\}_{i=1}^B \sim \mathcal{U}([0,1]^d)$, $\{t_i\}_{i=1}^B \sim \mathcal{U}(0,1)$
  $C_{ij} \leftarrow \|\mathbf{x}_i - \mathbf{Q}_\phi(\mathbf{u}_j)\|_2^2$
  $T \leftarrow \arg\min_T \sum_{i=1}^B C_{i,T(i)}$
  $L_Q \leftarrow \frac{1}{B}\sum_{i=1}^B \|\mathbf{x}_i - y_{T(i)}\|_2^2$
  $\mathbf{z}_i \leftarrow (1-t_i)\mathbf{x}_i + t_i\,\tilde{\mathbf{y}}_i$
  $L_{\mathrm{FM}} \leftarrow \frac{1}{B}\sum_{i=1}^B \|v_\theta(\mathbf{z}_i, t_i) - \mathbf{y}_{\mathbf{T(i)}} + \mathbf{x}_i\|_2^2$
  $L \leftarrow L_{\mathrm{FM}} + \lambda_Q L_Q$
  Update (θ, ϕ) by a gradient step on L
end for
return (θ, ϕ)

We support baseline flow matching, optional quantile pretraining, and joint quantile+velocity optimisation. Pretraining fits the RQS transport before optionally freezing it; joint training updates both modules simultaneously. Once the quantile learning rate decays to zero we freeze its weights and continue optimising the velocity field only.

The coupling plans are calculated using the Python Optimal Transport package (Flamary et al., 2021). For inference simulate the corresponding ODEs using the torchiffeq (Chen, 2018) package. For all models we only used the batch size 128 and learning rate $2e-4$ for the velocities. Quantile transports are parameterised by stacked rational-quadratic splines following Durkan et al. (2019).

E.1 Synthetic Examples

All models include a sinusoidal time embedding and SiLU activation functions.

Funnel. For the funnel distribution, we pretrain our quantiles and use the frozen quantiles during flow matching. We trained our quantile for 20,000 steps and to compensate we trained our velocity for only 150,000 steps. For the RQS we choose the parameters number of bins 64, bound 25, layers 3.

Grid Gaussian Mixture and Checker. The quantiles were trained for the first 20,000 steps, after which the learning rate was linearly decayed to 0 by step 25,000. For both datasets, we trained the velocity model with 4 layers and a hidden width of 256 for 100,000 steps. For the RQS we choose the parameters number of bins 32, bound 5, layers 3.

E.2 Image Experiments

For both image datasets, we adapt the U-Net from (Dhariwal & Nichol, 2021) to parametrize our velocity field.

MNIST. For the MNIST dataset we use the U-Net with base width 64, channel multipliers $(1, 2, 4)$, two residual blocks per resolution, attention at 7×7, 1 attention head, and dropout 0.1. We clip the gradient

norm to 1 and use exponential moving averaging with a decay of 0.99. The quantiles were trained for the first 20,000 steps, after which the learning rate was linearly decayed to 0 by step 25,000.

CIFAR. Here we use the U-Net with base width 128, channel multipliers $(1, 2, 2, 2)$, two residual blocks per resolution, attention at 16×16, four attention heads , and dropout 0.1. We clip the gradient norm to 1 and use exponential moving averaging with a decay of 0.9999. To evaluate our results, we use the Fréchet inception distance (FID) (Heusel et al., 2017). The quantiles were trained for the first 20,000 steps, after which the learning rate was linearly decayed to 0 by step 25,000.

CIFAR-10 inputs are normalized to $[-1, 1]$ with random horizontal flips.

Table 1: Overview of the default parameters. Adam with LR $2{\times}10^{-4}$.

Setting	Funnel	GMM / Checker	MNIST	CIFAR-10
Dimensionality d	2	2	784 (1×28×28)	3072 (3×32×32)
MLP layers	3	3	–	–
MLP width	64	128	–	–
Positional embeddings	none	sinusoidal	–	–
UNet channels	–	–	64	128
UNet mult	–	–	$(1, 2, 4)$	$(1, 2, 2, 2)$
UNet attention depth	–	–	7	16
Quantile objective	W_2^2	W_2^2	W_2^2	W_2^2
RQS (layers / bound / bins)	2 / 25/ 64	3 / 5/ 32	3 / 5/ 16	4 / 5/ 16
KL weight λ_{KL}	0	0	0.1	$1, 2, 3, 4$
Minibatch OT	off	on	on	on
Batch size	128	128	128	128
Train steps (k)	150/200	100	100	400
EMA (model)	0.999	0.99	0.99	0.9999 (warmup 5k)
Quantile loss weight	50.0	50	5.0	5.0
Quantile LR / batch / steps	1e−4/128/50k	2e−4/128/25k	1e−4/128/25k	2e−4/128/25k
Quantile EMA (used)	0.99 (on)	–	–	–

F Further Experimental Results

F.1 Synthetic Examples

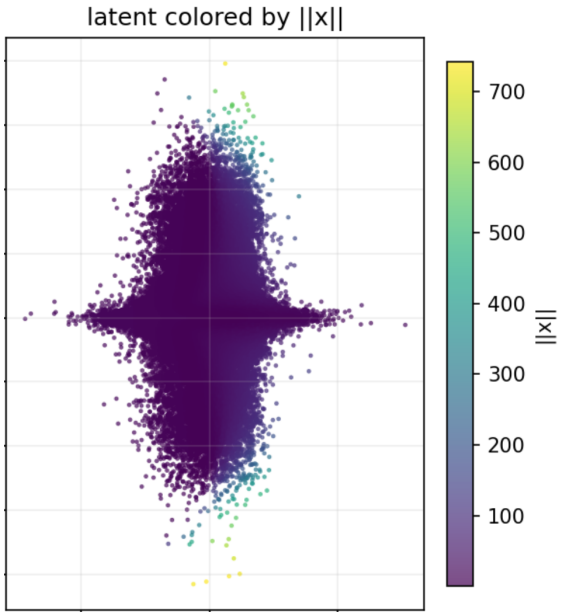


Figure 10: We plot 1M samples from our learned funnel latent and color them by the norm of the associated endpoint after solving the flow ODE.

F.2 Images

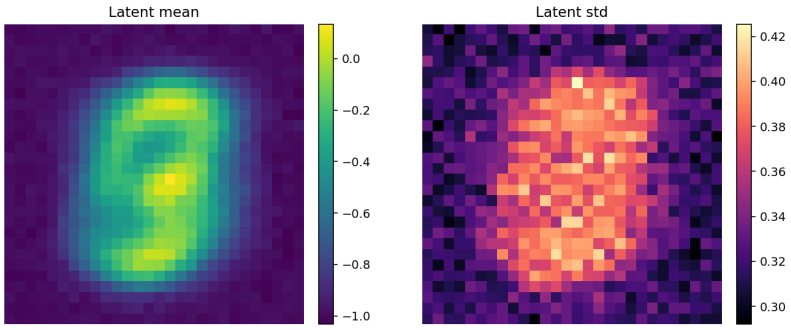


Figure 11: Mean and standard deviation of our learned MNIST latent.

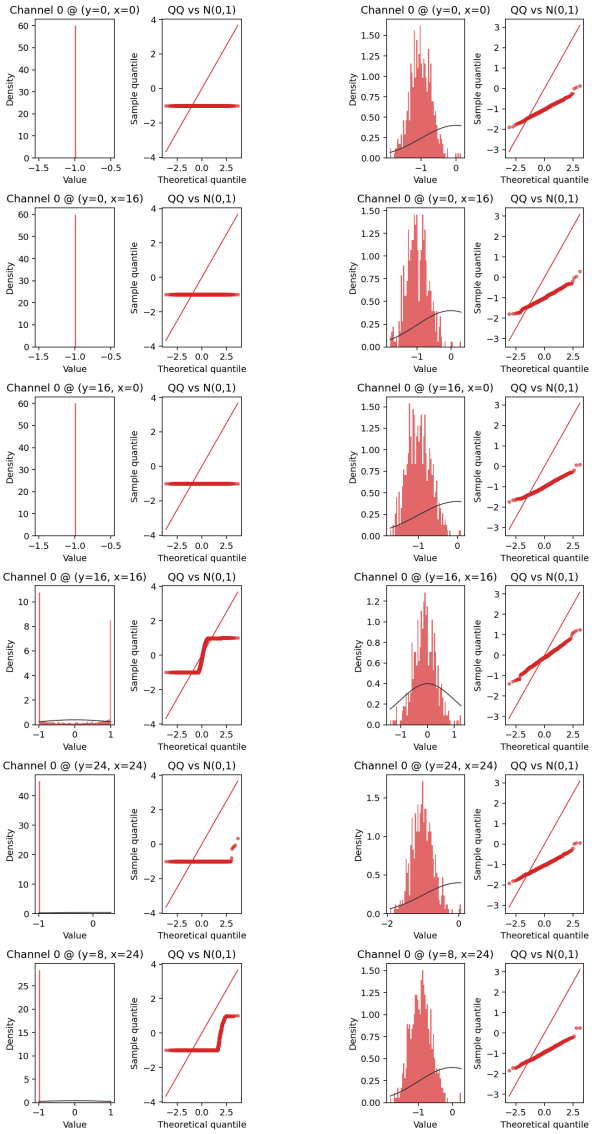


Figure 12: Comparison of the empirical (left) and learned (right) histograms and quantiles for the MNIST dataset at given pixel locations.