CONCENTRATING VELOCITIES AND ACCELERATIONS IN HIERARCHICAL FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical rectified flow models the multimodal velocity distribution of a flow model by learning an acceleration field, replacing a single ODE with a nested pair. Nothing in this construction forces the distribution modeled at one level to be simpler than the one above. Specifically, at $t = 0$, when source and target samples are paired independently, the velocity distribution there is exactly a shifted copy of the data distribution. We address this by applying mini-batch optimal transport couplings within each level, giving two methods at different operating points. Coupling data pairs concentrates the velocity distribution and straightens the outer trajectories. Coupling velocity pairs concentrates the acceleration distribution and straightens the inner trajectories. Neither alone suffices, because the sampling cost is the product of the steps taken at each level: data coupling alone leaves the inner solve expensive, while velocity coupling alone leaves the outer trajectories curved. Applying both makes every level cheap at once, which is what enables generation with small number of neural function evaluations. We prove that data and velocity couplings hierarchically simplify the velocity and acceleration distributions. We also demonstrate that the couplings improve generation quality and efficiency on synthetic data, MNIST, CIFAR-10, CelebA-HQ 256, and ImageNet 256.

1 INTRODUCTION

Flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023) has gained significant attention across computer vision (Esser et al., 2024; Liu et al., 2024), robotics (Zhang & Gienger, 2024), computational biology (Yim et al., 2023; Jing et al., 2023), and time series analysis (Chen et al., 2024; Zhang et al., 2024). This is largely due to its ability to generate high-quality data and due to its simple simulation-free learning of a data distribution. For this, it uses 1) an intermediate state, which is computed by (linearly) interpolating between a sample from a known source distribution and a randomly drawn data point, and 2) the velocity at this intermediate state. This velocity controls a neural ordinary differential equation (ODE), which governs the transformation of the samples from the source distribution to the target data distribution. Note, the distribution of velocities at an intermediate state is multimodal (Zhang et al., 2025).

In classic flow matching, velocities at interpolated states are modeled via a parametric deep net using a mean squared error (MSE) objective. It is known that the MSE objective used in classic flow matching does not permit to capture the multimodal velocity distribution. Instead, training in classic flow matching leads to a velocity model that captures the mean of the velocity distribution. Capturing the mean of the velocity distribution is sufficient for characterizing a multimodal data distribution (Liu et al., 2023). However, it inevitably results in curved forward integration paths, making the sampling process inefficient. Recently, hierarchical flow matching (Zhang et al., 2025) was suggested as an approach to model the multimodal velocity field via coupled ODEs.

To model the multimodal velocity field, hierarchical flow matching essentially applies a flow matching formulation in the velocity space by matching 'acceleration.' It was also suggested to expand the idea further towards an arbitrary hierarchy level. While this enables to model multimodal velocity distributions, and also distributions at arbitrary hierarchy levels, nothing in the construction forces the distribution modeled at one level to be simpler than the distribution modeled at the level above. At $t = 0$ it demonstrably is not: the velocity distribution is then a shifted copy of the data distribution, and the same holds for the acceleration distribution at $t = \tau = 0$. The second level of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the hierarchy therefore inherits in full the multimodality it was introduced to model, so the hierarchy buys expressivity without guaranteeing tractability, potentially limiting benefits.

We hence wonder: *can we gradually simplify the complexity of the distributions across hierarchy levels?* For simplicity, in this paper, we focus on two hierarchy levels. Interestingly, we find mini-batch couplings to provide a compelling way to control the "ground-truth" velocity distribution. Instead of computing intermediate states by interpolating between samples independently drawn from both the known source distribution and the dataset, we draw a mini-batch of samples from both the source distribution and the dataset, and subsequently couple them, e.g., via a procedure like optimal transport. In the limit of exact optimal transport this collapses the velocity distribution to a point mass in any dimension, and applying the same idea in velocity space collapses the acceleration distribution, so that a single inner integration step becomes exact.

Our contributions are: (i) we generalize HRF2 to mini-batch coupled data pairs and prove marginal preservation; (ii) we introduce per-level data and velocity couplings, yielding two methods at different operating points: data coupling (HRF2-D) simplifies the velocity distribution and attains the best generation quality when the step budget is generous, while adding velocity coupling (HRF2-D&V) further simplifies the acceleration distribution and is substantially better at low NFE; (iii) we prove that the couplings hierarchically simplify the modeled distributions, giving an exact decomposition in which the dispersion of one level enters the next additively, together with dimension-free results showing that optimal transport at each level drives that dispersion to zero; and (iv) we show that velocity coupling is not a distillation objective, both through marginal preservation and through a simulation-free construction. Experiments on synthetic data, MNIST, CIFAR-10, CelebA-HQ 256 and ImageNet 256 confirm both operating points. Notably, HRF2-D&V achieves a 78% FID reduction on CelebA-HQ at five neural function evaluations (NFEs) and a 67% FID reduction on ImageNet 256 at 4 NFEs compared with the original HRF2 models.

2 PRELIMINARIES

Rectified Flow (RF). A rectified flow models an unknown target data distribution ρ_1 given a dataset $\mathcal{D} = \{x_1\}$, where we assume data points $x_1 \sim \rho_1$. Given a known source distribution ρ_0 (e.g., standard Gaussian), at inference time, source samples $x_0 \sim \rho_0$ evolve from time $t = 0$ to time $t = 1$ following the ordinary differential equation (ODE)

$$dz_t = v(z_t, t)dt, \text{ with } z_0 \sim \rho_0, \quad t \in [0, 1]. \tag{1}$$

Here, $v(z_t, t)$ is a velocity field that depends on time t and the current intermediate state z_t. This ODE-based sampling enables to capture multimodal data distributions.

At training time, flow matching learns the velocity field $v(z_t, t)$ by minimizing the ℓ_2-loss between the predicted velocity $v(x_t, t)$ and a ground-truth velocity $v_{\text{gt}}(x_t, t)$. To obtain the ground-truth velocity we first define an intermediate state x_t which, in a rectified flow formulation, is obtained by linearly interpolating between a randomly drawn source sample x_0 and a randomly drawn data point x_1, i.e.,

$$x_t = (1 - t)x_0 + tx_1, \quad \text{where } x_0 \sim \rho_0,\ x_1 \sim \mathcal{D}. \tag{2}$$

Interpreting the intermediate state x_t as a location, we obtain the ground-truth velocity $v_{\text{gt}}(x_t, t) = \partial x_t / \partial t = x_1 - x_0$. Combined, training addresses

$$\inf_v \mathbb{E}_{x_0 \sim \rho_0, x_1 \sim \mathcal{D}, t \sim U[0,1]} \left[\|x_1 - x_0 - v(x_t, t)\|_2^2 \right], \tag{3}$$

where the infimum is over all measurable velocity fields. In practice, $v(x_t, t)$ is parameterized by a deep net with trainable parameters θ, i.e., $v(x_t, t) \approx v_\theta(x_t, t)$. The optimization minimizes over θ.

However, for a given t and x_t, different pairs (x_0, x_1) will yield different ground-truth velocities. The ground-truth velocity distribution at a given time t and intermediate state x_t is hence multimodal. However, the ℓ_2-loss averages these velocities, resulting in the 'optimal' velocity field: $v^*(x_t, t) = \mathbb{E}_{\{(x_0, x_1, t):(1-t)x_0 + tx_1 = x_t\}}[v_{\text{gt}}(x_t, t)]$. According to Theorem 3.3 by Liu et al. (2023), using v^* in Eq. (1) ensures that the stochastic process has marginal distributions consistent with the linear interpolation in Eq. (2).

To capture multimodal velocity distributions, hierarchical flow matching (Zhang et al., 2025) was introduced. It explicitly models the multimodal velocity distributions at each time t and intermediate state x_t, enabling a more expressive generative framework.

Hierarchical Rectified Flow (HRF). To model the "ground-truth" velocity distribution more accurately, hierarchical rectified flow extends the classic rectified flow framework by focusing on velocities rather than locations. This approach effectively involves learning acceleration. In a classic rectified flow, the time-dependent location x_t is computed from pairs (x_0, x_1), and the ground-truth velocity $v_{\text{gt}}(x_t, t) = \partial x_t / \partial t$ is used to train a velocity model $v_\theta(x_t, t)$.

In hierarchical rectified flow, a source velocity sample $v_0 \sim \pi_0$ is drawn from a known velocity distribution π_0, while a target velocity sample $v_1(x_t, t) \sim \pi_1(v; x_t, t)$ is defined at each time t and location $x_t = (1-t)x_0 + tx_1$. For rectified flow, $v_1(x_t, t)$ is computed via $x_1 - x_0$, and these samples follow the ground-truth velocity distribution $\pi_1(v; x_t, t)$.

To learn acceleration, a new time axis $\tau \in [0, 1]$ is introduced, and a time-dependent velocity $v_\tau(x_t, t) = (1-\tau)v_0 + \tau v_1(x_t, t)$ is constructed. The ground-truth acceleration is then obtained as $a(x_t, t, v_\tau, \tau) = \partial v_\tau / \partial \tau = v_1(x_t, t) - v_0 = x_1 - x_0 - v_0$. For a fixed (x_t, t), this leads to the ODE:

$$du_\tau(x_t, t) = a(x_t, t, u_\tau, \tau)d\tau, \quad \text{with } u_0 \sim \pi_0. \tag{4}$$

Here, $a(x_t, t, u_\tau, \tau)$ is the expected acceleration vector field: $a(x_t, t, u_\tau, \tau) = \mathbb{E}_{\pi_0, \pi_1(v; x_t, t)}[V_1 - V_0 | V_\tau = u]$. The acceleration vector field is learned by addressing

$$\inf_a \mathbb{E}_{x_0 \sim \rho_0, x_1 \sim \mathcal{D}, t \sim U[0,1], v_0 \sim \pi_0, \tau \sim U[0,1]} \left[\|(x_1 - x_0 - v_0) - a(x_t, t, v_\tau, \tau)\|_2^2 \right]. \tag{5}$$

In practice, the acceleration is parameterized via a deep net $a_\theta(x_t, t, v_\tau, \tau)$, and the model is trained by minimizing this objective over the parameters θ.

During sampling, coupled ODEs are used:

$$\begin{cases} du_\tau(z_t, t) = a(z_t, t, u_\tau, \tau)d\tau, & u_0 \sim \pi_0,\ \tau \in [0, 1], \\ dz_t = u_1(z_t, t)dt, & z_0 \sim \rho_0,\ t \in [0, 1]. \end{cases} \tag{6}$$

These ODEs map $z_0 \sim \rho_0$ to $z_1 \sim \rho_1$. Sampling involves drawing $v_0 \sim \pi_0$ and $x_0 \sim \rho_0$, integrating forward to obtain $v_1(x_0, 0)$, and then performing location updates iteratively until reaching x_1. This procedure can be implemented using the vanilla Euler method and the trained a_θ.

Considering the training objective for acceleration matching (Eq. (5)) and the coupled ODEs for sampling (Eq. (6)), both can be naturally extended to any depth. In this paper, we focus solely on depth-two HRF (HRF2) models.

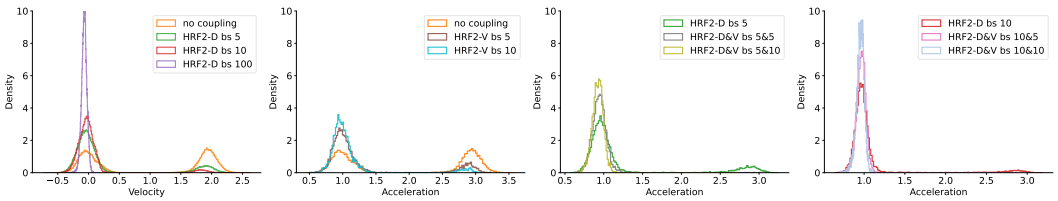
Minibatch Optimal Transport. Optimal Transport (OT) seeks to find an optimal coupling of two distributions that minimizes an expected transport cost (Villani, 2009). Suppose α and β are two distributions in $\mathbb{R}^d$, and $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is some distance function. Then OT aims to find the solution of the following optimization problem: $\inf_{\gamma \in \Gamma} \mathbb{E}_{(x,y) \sim \gamma} c(x, y)$, where Γ is the set of joint distributions with marginals α and β. For empirical distributions this reduces to linear programming and is expensive at scale (Peyré et al., 2019), so mini-batch OT (Fatras et al., 2020; 2021) approximates the coupling on small batches; it has been widely used in generative modeling (Deshpande et al., 2018; 2019; Pooladian et al., 2023; Tong et al., 2024; Cheng & Schwing, 2025).

3 Approach

In Section 3.1, we use 1D data to illustrate how mini-batch couplings in data space and velocity space affect the velocity distribution and the generation of velocity samples. This motivates the development of HRF with mini-batch coupling and extension of its theory. In Section 3.2, we introduce the training of HRF2 with mini-batch coupled data points. In Section 3.3, we combine mini-batch data coupling and velocity coupling, and explain simulation-based and simulation-free constructions for mini-batch velocity coupling. We show that the marginal distributions of the velocity and data are preserved in Section 3.4, and discuss the hierarchical simplification of the distributions in Section 3.5.

3.1 Velocity and Acceleration Distributions

For vanilla HRF2, the source and target distributions are independent, meaning $\gamma(x_0, x_1) = \rho_0(x_0)\rho_1(x_1)$. Consequently, as derived by Zhang et al. (2025), at time $t = 0$, the velocity distribution becomes $\pi_1(v; x_t, t) = \rho_1(x_t + v)$, making it a shifted version of the data distribution. Hence, learning this distribution is as challenging as directly modeling the data distribution.



(a) velocity (b) acceleration w/o DC (c) acceleration DC bs5 (d) acceleration DC bs10

Figure 1: Effect of data coupling (DC) and velocity coupling (VC) on hierarchical distribution simplification. We visualize the empirical distributions at different hierarchy levels under various coupling configurations. (a) Velocity distributions at $(x_t, t) = (-1, 0)$ with data coupling. (b-d) Acceleration distributions at $(x_t, t, v_\tau, \tau) = (-1, 0, -1, 0)$ with data and velocity coupling. Data coupling simplifies the velocity distribution, while velocity coupling further simplifies the acceleration distribution. Increasing the OT batch size leads to progressively stronger simplification of the target distribution, resulting in more unimodal and peaked distributions.

To control the complexity of the velocity distribution, we find couplings in data space and velocity space to be effective. Concretely, for couplings in data space, we sample from a distribution $\gamma(x_0, x_1)$, which doesn't factorize. This can be achieved by coupling samples within each mini-batch, i.e., samples x_0 and data points x_1 are no longer combined via independent sampling into pairs (x_0, x_1). We find that this controls multimodality of the distributions, making them easier to learn and improving overall results. Notably, coupling in data space simplifies both the velocity and acceleration distributions, and coupling in velocity space further simplifies the acceleration distribution, so each coupling eases the matching problem at every subsequent level of the hierarchy. Importantly, the complexity of the learned distribution can be controlled by adjusting the batch size in the mini-batch coupling process.

We illustrate this distribution simplification on a 1D example with a Gaussian source and a two-mode Gaussian-mixture target. Figure 1(a) shows that data coupling (DC) (depth 1) makes the velocity distribution (depth 2) increasingly unimodal as the OT batch size (bs) increases. Velocity coupling (VC) does not change the velocity distribution itself, but simplifies the acceleration distribution at the next hierarchy level as shown in Figure 1(b). Figure 1(c,d) demonstrates that data and velocity coupling are complementary and can be applied jointly. Data coupling alone already leads to a noticeable simplification of the acceleration distribution, and adding velocity coupling results in a much stronger simplification. In the infinite-batch limit, the acceleration distribution at $(x_t, t, v_\tau, \tau) = (-1, 0, -1, 0)$ collapses to a Dirac delta centered at acceleration $a = 1$.

To quantitatively illustrate the hierarchical simplification, we measure how close the empirical distributions under different couplings are to the ideal target using the 1-Wasserstein distance. As shown in Table 1, both couplings independently and significantly reduce the Wasserstein distance. Their combination achieves the most concentrated distribution and gives the lowest distance.

Table 1: Effect of mini-batch coupling on acceleration distributions.

DC \ VC	no VC	VC 5	VC10
no DC	1.02	0.38	0.21
DC 5	0.32	0.13	0.12
DC10	0.15	0.09	**0.08**

These empirical observations motivate the central theoretical question: does HRF2 remain valid when the source and target samples are no longer paired independently?

Formally, HRF was designed with independently sampled x_0 and x_1. In this paper, we first show that the underlying theory can be generalized to an arbitrary joint distribution over x_0 and x_1, i.e., $\gamma(x_0, x_1)$, which has the correct marginal distributions, i.e.,

$$\int \gamma(x_0, x_1) dx_1 = \rho_0(x_0) \quad \text{and} \quad \int \gamma(x_0, x_1) dx_0 = \rho_1(x_1). \tag{7}$$

The following theorem characterizes the distribution of the velocity at a specific space-time location (x_t, t) if an arbitrary joint distribution γ is used instead of a product of two independent distributions.

Theorem 3.1. *The velocity distribution $\pi_1(v; x_t, t)$ at the space-time location (x_t, t) induced by the linear interpolation in Eq. (2) for $(x_0, x_1) \sim \gamma(x_0, x_1)$ is*

$$\pi_1(v; x_t, t) = \frac{\gamma(x_t - tv, x_t + (1-t)v)}{\rho_t(x_t)}, \tag{8}$$

where

$$\rho_t(x_t) = \int \gamma(x_t - tv, x_t + (1-t)v) dv, \tag{9}$$

and $\rho_t(x_t) \neq 0$. The distribution $\pi_1(v; x_t, t)$ is undefined if $\rho_t(x_t) = 0$.

The proof of Theorem 3.1 is deferred to Appendix A.

The same change of variables applied in velocity space gives the acceleration distribution $p(a; x_t, t, v_\tau, \tau) = \kappa_{x_t,t}(v_\tau - \tau a,\, v_\tau + (1-\tau)a)/\rho_\tau(v_\tau)$, where $\kappa_{x_t,t}$ denotes the joint distribution of the velocity pairs $(v_0, v_1(x_t, t))$. Appendix B states this as Corollary B.2 together with the more general version in which v_0, x_0 and x_1 are jointly coupled.

Without data coupling, the velocity distribution at $t = 0$ is $\pi_1(v; x_t, 0) = \rho_1(x_t + v)$, a shifted copy of the data distribution; evaluating the general form at $t = \tau = 0$ under fully independent sampling, $\gamma(v_0, x_0, x_1) = \pi_0(v_0)\rho_0(x_0)\rho_1(x_1)$, gives

$$p(a; x_0, 0, v_0, 0) \;=\; \rho_1(x_0 + v_0 + a), \tag{10}$$

again a shifted copy of the data distribution. The second level therefore inherits exactly the multimodality the first level was unable to avoid, and adding levels without coupling does not make the problem easier at any of them. This is the precise sense in which the hierarchy buys expressivity without buying tractability at $t = \tau = 0$, and it is what the orange curve in Figure 1(b) shows empirically.

Data coupling already improves matters at the second level as shown in Figure 1(b-d). Velocity coupling acts directly on $\kappa_{x_t,t}$ in the acceleration density above and concentrates the acceleration distribution further, as in Figure 1(b-d). Section 3.5 makes both statements precise.

3.2 Coupling in data space

As discussed in Section 3.1 , if source data x_0 and target data x_1 are sampled independently, the multimodality inherent in the data will be preserved in the velocity distribution at $t = 0$. Breaking this independence is hence key to simplifying the velocity distribution.

Following Tong et al. (2024); Pooladian et al. (2023), we apply mini-batch OT on the data used for HRF2 training. Let $\{x_0^{(i)}\}_{i=1}^B \sim \rho_0$ and $\{x_1^{(i)}\}_{i=1}^B \sim \mathcal{D}$. The OT problem can be solved exactly and efficiently on a small batch size using standard solvers, e.g., `POT` (Flamary et al., 2021). The resulting coupling from the algorithm gives us a permutation matrix that pairs $x_0^{(i)}$ with $x_1^{(\sigma(i))}$ for $i \in \{1, \ldots, B\}$. Instead of sampling x_0 and x_1 independently from ρ_0 and the dataset $\mathcal{D}$ during training, we jointly sample pairs (x_0, x_1) from the joint distribution $\gamma(x_0, x_1)$ characterized by the mini-batch OT result. Using these samples, the training objective reads as follows:

$$\min_\theta \mathbb{E}_{(x_0,x_1)\sim\gamma, t\sim U[0,1], v_0\sim\pi_0, \tau\sim U[0,1]} \left[\|(x_1 - x_0 - v_0) - a_\theta(x_t, t, v_\tau, \tau)\|_2^2 \right]. \tag{11}$$

The optimization procedure is detailed in Algorithm 1 in Appendix C.

3.3 Coupling in Velocity Space

Velocity coupling removes the independence between v_0 and $v_1(x_t, t)$ in the same way data coupling removes it between x_0 and x_1: at a fixed space-time location (x_t, t), the pair $(v_0, v_1(x_t, t))$ is drawn from a joint distribution $\kappa_{x_t,t}$ rather than independently. Because the coupling of the velocities depends on the coupling of the data, we apply the two together. This requires joint data samples $(x_0, x_1) \sim \gamma$ alongside coupled velocity samples $(v_0, v_1(x_t, t)) \sim \kappa_{x_t,t}$, and the network is trained with

$$\min_\theta \mathbb{E}_{(x_0,x_1)\sim\gamma, (v_0,v_1)\sim\kappa_{x_t,t}, t\sim U[0,1], \tau\sim U[0,1]} \left[\|(v_1(x_t, t) - v_0) - a_\theta(x_t, t, v_\tau, \tau)\|_2^2 \right]. \tag{12}$$

Algorithm 2 in Appendix C summarizes the optimization procedure. The velocity pairs can be obtained in two ways. In the simulation-based pipeline we first train a base model a_θ using data coupling (Algorithm 1); a pre-trained HRF2 model then maps samples $v_0 \sim \pi_0$ to $v_1(x_t, t)$ by integrating the inner velocity-space ODE, after which mini-batch OT couples the source and generated target velocities (Algorithm 3). The simulation-free construction instead bypasses simulation altogether: starting from mini-batch coupled data pairs (x_0, x_1), it aggregates data-coupled interpolation paths whose interpolated locations fall within an ϵ-ball of x_t, and extracts the corresponding coupled velocity pairs (Algorithm 4). This construction is available in low-dimensional settings and shows that velocity coupling is not inherently a distillation procedure, since it never invokes a teacher model. Appendix C details both constructions and compares them empirically. This construction differs from reflow. Reflow operates in data space, integrating the learned flow to regenerate source-target pairs (z_0, z_1) and retraining on them, whereas velocity coupling operates in velocity space, re-pairing (v_0, v_1) at a fixed (x_t, t) by optimal transport.

3.4 Marginal Preserving Property

The consistency of the velocity distribution with coupled samples drawn from $\kappa_{x_t,t}$ directly follows prior works that use mini-batch coupling and reflow for data generation (Tong et al., 2024; Pooladian et al., 2023). It means that the marginals of the velocity distribution at any (t, τ) are preserved. In addition, we can prove that the generation process according to Eq. (6) with coupled data points $(x_0, x_1) \sim \gamma$ preserves the target data distribution and leads to correct marginals for all times $t \in [0, 1]$.

Theorem 3.2. *The time-differentiable stochastic process $\boldsymbol{Z} = \{Z_t : t \in [0,1]\}$ generated by Eq.* (6) *has the same marginal distribution as the time-differentiable stochastic process $\boldsymbol{X} = \{X_t : t \in [0,1]\}$ generated by the linear interpolation in Eq.* (2) *with the joint distribution γ induced by mini-batch coupling.*

The proof of Theorem 3.2 is deferred to Appendix D.

3.5 Hierarchical Simplification

Having established that coupling at either level leaves the modeled marginals intact, we now make the simplification claim precise. We measure the difficulty of each level by its conditional dispersion,

$$\mathcal{C}_1(\gamma) := \int_0^1 \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(v_1 \mid x_t)\big]\, dt, \qquad \mathcal{C}_2(\gamma, \kappa) := \int_0^1 \int_0^1 \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(a \mid x_t, v_\tau)\big]\, d\tau\, dt, \quad (13)$$

where $v_1 = x_1 - x_0$, $a = v_1 - v_0$, and $\pi_0 = \mathcal{N}(0, I_d)$; assumptions and admissible couplings are stated in Appendix E. By the L^2-projection identity (Lemma E.1), $\mathcal{C}_1$ and $\mathcal{C}_2$ are the irreducible losses of velocity matching (Eq. (3)) and acceleration matching (Eq. (12)). HRF2 trains only the latter; $\mathcal{C}_1$ enters it through Theorem 3.4 and Proposition 3.5.

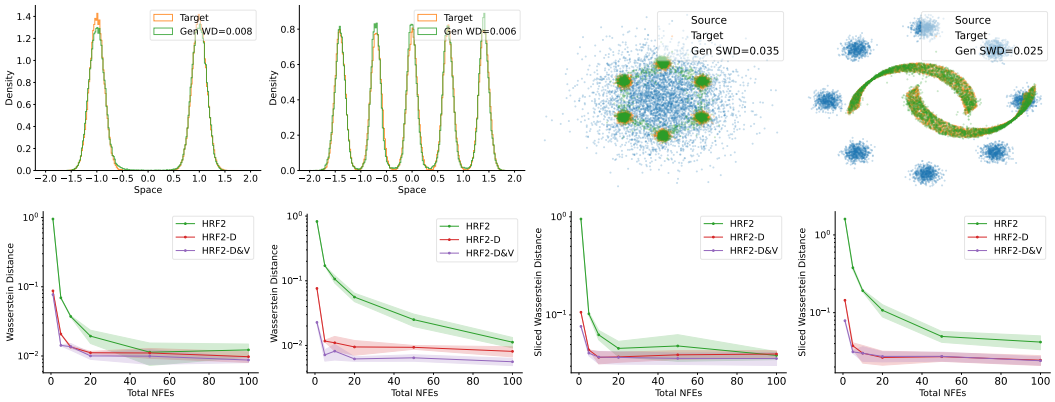
Theorem 3.3. *Let γ^{OT} be the optimal quadratic-cost coupling of ρ_0 and ρ_1. Then for every $t \in [0, 1)$ the velocity distribution $\pi_1(\cdot\,; x_t, t)$ is a Dirac mass for ρ_t-almost every x_t, and $\mathcal{C}_1(\gamma^{\mathrm{OT}}) = 0$.*

At finite batch size, γ_B has the correct marginals and a transport cost that is non-increasing along divisor chains (Proposition E.2), and the transport cost bounds $\mathcal{C}_1$ from above (Lemma E.3). Propositions E.4 and E.6 give convergence $\mathcal{C}_1(\gamma_B) \to 0$ under a Lipschitz hypothesis and a dimension-free gain already at $B = 2$. The two levels are linked exactly, and what concentrating the second level buys is a cheap inner solve.

Theorem 3.4. *For every admissible γ and κ, $\mathcal{C}_2(\gamma, \kappa) = \mathcal{C}_1(\gamma) + \operatorname{tr}\Sigma_{\pi_0} - 2\mathcal{K}(\gamma, \kappa) - \mathcal{D}(\gamma, \kappa)$, where $\mathcal{K}$ averages the trace-covariance between v_0 and v_1 under κ and $\mathcal{D} \geq 0$ is the dispersion explained by observing v_τ.*

Among velocity couplings with fixed marginals, optimal transport maximizes $\mathcal{K}$ (Lemma E.8), which singles it out as the natural coupling at the second level.

Proposition 3.5. *Let $\hat{u}_1 = v_0 + a^\star(x_t, t, v_0, 0)$ be the result of a single Euler step of size one applied to the inner ODE from $\tau = 0$. Then $\mathbb{E}\big[\|\hat{u}_1 - v_1\|^2 \mid x_t\big] = \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(a \mid x_t, v_0) \mid x_t\big]$, the $\tau = 0$ slice of the integrand of $\mathcal{C}_2$. Under independent velocity sampling, $\hat{u}_1 = \bar{v}(x_t, t)$ is the conditional mean velocity and the slice integrates to $\mathcal{C}_1(\gamma)$.*



(a) 1D $\mathcal{N} \to 2\mathcal{N}$ (b) 1D $\mathcal{N} \to 5\mathcal{N}$ (c) 2D $\mathcal{N} \to 6\mathcal{N}$ (d) 2D $8\mathcal{N} \to$ moon

Figure 2: Results on synthetic datasets: (a) 1D $\mathcal{N} \to 2\mathcal{N}$ (b) 1D $\mathcal{N} \to 5\mathcal{N}$ (c) 2D $\mathcal{N} \to 6\mathcal{N}$ (d) 2D $8\mathcal{N} \to$ moon. Top row: HRF2-D&V generated data distributions. Bottom row: performance vs. total NFEs. We use Wasserstein and sliced 2-Wasserstein distances for 1D and 2D data, respectively.

Without velocity coupling, the one-step inner solve thus returns the conditional mean and inherits the entire level-one difficulty: at a total NFE of one, HRF2-D outputs $\mathbb{E}[x_1 \mid x_0]$, the blurred samples of Figure 3(a). Velocity coupling removes this obstruction.

Theorem 3.6. *Let $\kappa^\star_{x_t,t}$ be the optimal quadratic-cost coupling of π_0 and $\pi_1(\cdot\,; x_t, t)$. Then for every admissible data coupling γ and every $\tau \in [0,1)$, the acceleration a is a deterministic function of (x_t, v_τ). Hence $\mathcal{C}_2(\gamma, \kappa^\star) = 0$ and a single inner Euler step is exact.*

Theorem 3.6 holds for every γ, including independent pairing, but it concerns only the inner solve: it makes that solve exact without making it easy. Under independent pairing the velocity distribution at $t = 0$ is a shifted copy of the data distribution (Section 3.1), so the inner map to be learned transports π_0 to the data distribution itself, and the base model that supplies samples of π_1 faces the same problem. Data coupling is what concentrates π_1, so that this map is close to a constant and the velocity samples are accurate; in general, zero dispersion means a deterministic target, not a regular one, since optimal transport maps between multimodal distributions can be steep. Conversely, without velocity coupling the level-two dispersion is bounded below by a Gaussian denoising error of π_1 and is strictly positive whenever $\mathcal{C}_1(\gamma) > 0$ (Corollary E.9): data coupling shrinks it but does not remove it.

Because $\mathcal{C}_1$ and $\mathcal{C}_2$ are irreducible losses, each model's converged acceleration-matching training loss is an empirical upper estimate of its $\mathcal{C}_2$ (Table 9). On ImageNet 256 data coupling leaves it essentially unchanged (from 1.134 to 1.120), consistent with Theorem 3.4, while velocity coupling reduces it by two orders of magnitude (from 1.120 to 0.010), as predicted by Theorem 3.6.

In simulation-based approach, the velocity pairs are built offline. We record locations (x_t, t) by interpolating mini-batch OT coupled data pairs, so they follow the data-level interpolation exactly. At each location, the pre-trained HRF2-D model, used only as a sampler of π_1, generates velocity samples by integrating its inner ODE, and these are coupled with source velocities by mini-batch OT (Appendix E.5). No approximation is made in data space; the only error is in the velocity samples. For the sampler we actually run, this error is controlled by its W_2 size at small outer budgets and by its mean at large ones (Proposition E.12). For HRF2-D&V both are fixed once the pairs are generated and the model is trained, and do not depend on the inference budget (Corollary E.13); for HRF2-D, a one-step inner solve has zero mean error but the full spread of π_1 as distributional error. This is consistent with the crossover in Section 4.2, with HRF2-D&V stronger at small step budgets and HRF2-D at generous ones (Remark E.14).

Theorems 3.3 and 3.6 describe the exact-transport limit, which mini-batch coupling approaches only as the batch size grows. Solving the OT problem exactly at the data level would make the velocity distribution a Dirac mass and the hierarchy unnecessary. So the regime of interest is the finite-batch one, where Section 4 shows HRF2-D improving on OT-CFM. Because the total NFE is the product of the steps taken at each level, data coupling straightens the outer trajectories but leaves the inner solve expensive, whereas velocity coupling simplifies the inner solve but leaves the outer trajectories



(a) HRF2-D, NFE 1 (b) HRF2-D, NFE 500 (c) HRF2-D&V (d) FID vs. NFE

Figure 3: Results on CelebA-HQ 256 (top row) and ImageNet 256 (bottom row). (a) HRF2-D at total NFE 1. (b) HRF2-D at total NFE 500. (c) HRF2-D&V, total NFE 1 for CelebA-HQ and 4 for ImageNet. (d) FID vs. total NFEs, compared with RF, OT-CFM, and HRF2 baselines. While data coupling alone collapses at a single step, incorporating velocity coupling enables coherent generation at very low NFEs. Results on MNIST and CIFAR-10 are provided in Appendix G.2.

curved. Jointly applying the data and velocity couplings reduces the inference cost significantly. Proofs are deferred to Appendix E.

4 Experiments

In this section, we explore how data coupling and velocity coupling influence the velocity distribution and assess whether simplifying it leads to performance improvements. For all experiments, we report the total neural function evaluations (NFEs), which represents the product of the number of integration steps across all HRF levels. All models are trained to convergence and evaluated at the best checkpoint selected on a held-out validation set.

4.1 Synthetic Data

We conduct experiments on four synthetic datasets used by Zhang et al. (2025) to ensure a fair comparison with HRF2. These datasets include two 1D cases ($\mathcal{N} \to 2\mathcal{N}$, $\mathcal{N} \to 5\mathcal{N}$), and two 2D cases ($\mathcal{N} \to 6\mathcal{N}$ and $8\mathcal{N} \to$ moon-shaped data). We use the Wasserstein and sliced 2-Wasserstein distances to evaluate 1D and 2D experiments, respectively. A complete description of the model architecture, parameter settings, and training details is provided in Appendix H.1.

We denote HRF2 with data coupling as HRF2-D and HRF2 with joint data and velocity coupling as HRF2-D&V. The results in Figure 2 demonstrate that joint data and velocity couplings effectively enhance the model performance, particularly when NFEs are low. Additional results on synthetic data are provided in Appendix G.1.

4.2 Image Data

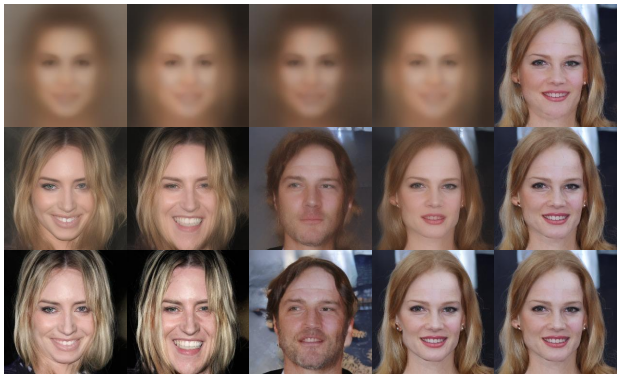
For high-dimensional image data, we present CelebA-HQ 256 (Karras et al., 2018) and ImageNet 256 (Deng et al., 2009) in the main text, and defer MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky, 2009) to Appendix G.2. We use the Fréchet Inception Distance (FID) as the evaluation metric. For both CelebA-HQ and ImageNet we train and sample in the latent space of a pretrained VAE; architecture and training details are provided in Appendix H.2. On ImageNet, FID is computed against the ADM reference statistics over 50K EMA samples with classifier-free guidance, and the total NFE is the product of the outer and inner integration steps.

We compare our method to RF (Liu, 2022), HRF2 (Zhang et al., 2025), OT-CFM (Tong et al., 2024), and ShortCut (Frans et al., 2025) baselines. HRF2 is the base pre-trained model for HRF2-D and HRF2-D&V, and OT-CFM is essentially equivalent to HRF1-D. The couplings hierarchically simplify the modeled distributions and improve generation. Data coupling helps across NFEs while velocity coupling concentrates the gain at low NFE, so HRF2-D&V stays coherent in a few steps whereas HRF2-D collapses at one step (Figures 3 and 4). The two couplings therefore provide two operating points, HRF2-D&V at small budgets and HRF2-D at generous ones. Full FID numbers for all methods, datasets, and total NFEs are reported in Appendix G.2.

Velocity coupling is cheap to add. The velocity pairs are generated once offline and independently across locations, so this step parallelizes across GPUs, and since our model uses a large backbone for the location and a small head for the velocity (Appendix H.2), velocity coupling fine-tunes only the small head with the backbone frozen instead of retraining the model. Sampling is unchanged from HRF2, so it adds no inference cost, and Table 18 lists the model sizes and per-step costs. Optimal-transport coupling also outperforms reflow-style pairing (Table 13).

5 Related Work

Flow Matching: Flow matching (Liu et al., 2023; Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023) learns an ODE velocity field from stochastic processes that interpolate source and target samples. Rectified flow (Liu et al., 2023) uses linear interpolations, while conditional flow matching (Lipman et al., 2023) and stochastic interpolants (Albergo & Vanden-Eijnden, 2023; Albergo et al., 2023) study more general paths. These methods learn the expected velocity field, which matches marginals but can yield curved sampling paths. Extensions include discrete (Gat et al., 2024; Stark et al., 2024) and manifold (Chen & Lipman, 2024) data.



(a) RF (b) OT-CFM (c) HRF2 (d) HRF2-D (e) HRF2-D&V

Figure 4: Generated CelebA-HQ 256 images from the same noise for all models. Rows from top to bottom correspond to total NFEs $= 1, 5, 50$. Only HRF2-D&V produces coherent images in the very-low-NFE regime.

Straightening Flows: Reflow (Liu et al., 2023) iteratively rebuilds source-target pairs from a trained rectified flow and approaches an optimal transport map, but can accumulate model-error bias. Mini-batch OT flow matching (Pooladian et al., 2023; Tong et al., 2024) instead couples source and target samples during training to reduce transport cost and flow variance. Other works learn acceleration (Park et al., 2024) or study conditional settings (Cheng & Schwing, 2025).

Distribution of flow fields: Hierarchical rectified flow (Zhang et al., 2025) models the distribution of random velocity fields via coupled ODEs, letting sampling paths cross. Variational rectified flow (Guo & Schwing, 2025) models flow-field distributions with a variational autoencoder. Building on HRF, we show that data-space coupling simplifies the velocity distribution and improves HRF2 and OT-CFM, while joint data and velocity coupling further improves low-NFE generation.

6 Conclusion

We study how to control the complexity of the multimodal velocity distribution in hierarchical flow matching. Across data from synthetic mixtures to ImageNet 256, mini-batch data coupling consistently improves generation quality over vanilla HRF2 and over flow matching with mini-batch optimal transport, and adding velocity coupling yields two operating points, with joint coupling stronger at low NFE and data coupling stronger at generous budgets. Code will be released for reproducibility.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs and provide some critical ingredients for proving some mathematical claims. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, support qualitative and thematic data analysis, implement methods, and interpret results, and assisting with translation and cleaning and reformating dataset are not applicable to this work. Additionally, we used generative AI tools for editing a research paper to improve readability and identifying relevant literature. We have reviewed all AI-assisted work. LLM-generated wording was checked for correctness against original draft version. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken several steps to make our results reproducible. All theoretical claims state their assumptions and are proven in full in the appendix. The velocity and acceleration distributions are derived in Appendix A and Appendix B, the marginal-preserving property is proven in Appendix D, and the hierarchical simplification results, together with the assumptions and admissible couplings they require, are established in Appendix E. The training procedures for data coupling and for both the simulation-based and simulation-free velocity coupling are given as Algorithms 1 to 4 in Appendix C. Model architectures, hyperparameters, optimization settings, and the per-dataset training and compute details for the synthetic, MNIST, CIFAR-10, CelebA-HQ 256, and ImageNet 256 experiments are provided in Appendix H, and the evaluation protocol, including the FID computation, is described in Section 4.2 and Appendix H. All datasets are publicly available, and the data processing steps, including the pretrained VAE latent spaces, are described in Appendix H.2. We report means and standard deviations over multiple seeds and evaluation runs, and provide additional results in Appendix G. Code will be released to reproduce all results.

REFERENCES

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

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *Proc. ICLR*, 2023.

Luis A Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496, 1996.

Ricky TQ Chen and Yaron Lipman. Flow matching on general geometries. In *Proc. ICLR*, 2024.

Yifan Chen, Mark Goldstein, Mengjian Hua, Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Probabilistic forecasting with stochastic interpolants and Föllmer processes. In *Proc. ICML*, 2024.

Ho Kei Cheng and Alexander Schwing. The Curse of Conditions: Analyzing and Improving Optimal Transport for Conditional Flow-Based Generation. In *Proc. ICCV*, 2025.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, pp. 248–255, 2009.

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative Modeling using the Sliced Wasserstein Distance. In *Proc. CVPR*, 2018.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G Schwing. Max-Sliced Wasserstein Distance and its use for GANs. In *Proc. CVPR*, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. ICML*, 2024.

Kilian Fatras, Younes Zine, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Learning with minibatch wasserstein: asymptotic and gradient properties. In *Proc. AISTATS*, volume 108, pp. 1–20, 2020.

Kilian Fatras, Younes Zine, Szymon Majewski, Rémi Flamary, Rémi Gribonval, and Nicolas Courty. Minibatch optimal transport distances; analysis and applications. *arXiv preprint arXiv:2101.01792*, 2021.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *Proc. ICLR*, 2025.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. In *Proc. NeurIPS*, 2024.

Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in gaussian channels. *IEEE transactions on information theory*, 51(4):1261–1282, 2005.

Pengsheng Guo and Alexander G Schwing. Variational rectified flow matching. In *Proc. ICML*, 2025.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. Alphafold meets flow matching for generating protein ensembles. In *NeurIPS 2023 Generative AI and Biology (GenBio) Workshop*, 2023.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. ICLR*, 2018.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

Tianhong Li, Yonglong Tian, Han Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. In *Proc. NeurIPS*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Matching for Generative Modeling. In *Proc. ICLR*, 2023.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. ICLR*, 2023.

Xingchao Liu, Xiwen Zhang, Jianzhu Ma, and Jian Peng. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In *Proc. ICLR*, 2024.

Anh Nguyen, Viet Van Nguyen, Duc Vu, Trung Tuan Dao, Chi Tran, Toan Tran, and Anh Tuan Tran. Improved training technique for shortcut models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=ovUuNzZZbK`.

Zanlin Ni, Yulin Wang, Renping Zhou, Yizeng Han, Jiayi Guo, Zhiyuan Liu, Yuan Yao, and Gao Huang. Enat: Rethinking spatial-temporal interactions in token-based image synthesis. In *NeurIPS*, 2024.

Zanlin Ni, Yulin Wang, Yeguo Hua, Renping Zhou, Jiayi Guo, Jun Song, Bo Zheng, and Gao Huang. Adagen: Learning adaptive policy for image synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Dogyun Park, Sojin Lee, Sihyeon Kim, Taehoon Lee, Youngjoon Hong, and Hyunwoo J Kim. Constant acceleration flow. In *Proc. NeurIPS*, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, pp. 4195–4205, 2023.

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

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky TQ Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proc. ICML*, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, 2022.

Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to dna sequence design. In *Proc. ICML*, 2024.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *TMLR*, 2024.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Jason Yim, Andrew Campbell, Andrew YK Foong, Michael Gastegger, José Jiménez-Luna, Sarah Lewis, Victor Garcia Satorras, Bastiaan S Veeling, Regina Barzilay, Tommi Jaakkola, and Frank Noe. Fast protein backbone generation with SE(3) flow matching. In *NeurIPS Workshop: Machine Learning in Structural Biology*, 2023.

Fan Zhang and Michael Gienger. Affordance-based robot manipulation with flow matching. *arXiv preprint arXiv:2409.01083*, 2024.

Xi Zhang, Yuan Pu, Yuki Kawamura, Andrew Loza, Yoshua Bengio, Dennis L Shung, and Alexander Tong. Trajectory flow matching with applications to clinical time series modeling. In *Proc. NeurIPS*, 2024.

Yichi Zhang, Yici Yan, Alex Schwing, and Zhizhen Zhao. Towards hierarchical rectified flow. In *Proc. ICLR*, 2025.

Linqi Zhou, Stefano Ermon, and Jiaming Song. Inductive moment matching. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=pwNSUo7yUb`.

APPENDIX: CONCENTRATING VELOCITIES AND ACCELERATIONS IN HIERARCHICAL FLOW MATCHINGS

This appendix is structured as follows: in Appendix A, we provide the proof of Theorem 3.1; in Appendix B, we assess acceleration distributions under couplings; in Appendix C, we provide additional details on algorithms; in Appendix D, we provide the proof of Theorem 3.2; in Appendix E, we prove the theorems and proposition in 3.5; in Appendix F, we provide additional non-asymptotic analysis on 1D Gaussian mixtures;in Appendix G, we provide additional results for both synthetic and image data; in Appendix H we discuss implementation details; in Appendix I, we provide numerical results on using adaptive solvers for the ODEs.

A PROOF OF THEOREM 3.1

Proof of Theorem 3.1: For simplicity, we show the proof for 1D random variables x_0 and x_1 drawn from a joint distribution $\gamma(x_0, x_1)$. The joint distribution of $v(x_t, t)$ and x_t is $\pi_1(v; x_t, t)\rho_t(x_t)$, since $\pi_1(v; x_t, t)$ corresponds to the conditional distribution of the velocity given location x_t. According to the linear interpolation in Eq. (2), we have

$$\begin{bmatrix} v(x_t, t) \\ x_t \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ t & (1-t) \end{bmatrix} \begin{bmatrix} x_1 \\ x_0 \end{bmatrix} = A \begin{bmatrix} x_1 \\ x_0 \end{bmatrix}, \tag{14}$$

where the matrix A has determinant 1. Since $[v(x_t, t), x_t]^T$ is a linear transformation of $[x_1, x_0]^T$, we have the following expression for the joint distribution of $v(x_t, t)$ and x_t:

$$\pi_1(v; x_t, t)\rho_t(x_t) = \frac{1}{\det(A)}\gamma\left(A^{-1}\begin{bmatrix} v(x_t, t) \\ x_t \end{bmatrix}\right) = \gamma(x_t - tv, x_t + (1-t)v). \tag{15}$$

After rearranging, we get $\pi_1(v; x_t, t) = \frac{\gamma(x_t - tv, x_t + (1-t)v)}{\rho_t(x_t)}$. For the higher dimensional case, the relation in Eq. (15) still holds. This completes the proof. ■

B DISTRIBUTION OF ACCELERATION a

In this section, we derive the acceleration distribution induced by linearly interpolating source and target data samples x_0, x_1 and by linearly interpolating source and target velocity samples v_0 and $v(x_t, t)$. Here, (v_0, x_0, x_1) are drawn from an underlying joint distribution γ. In addition, we consider the acceleration distribution under the velocity coupling $\kappa_{x_t,t}$ at (x_t, t).

Theorem B.1. *The acceleration distribution $p(a; x_t, t, v_\tau, \tau)$ is*

$$p(a; x_t, t, v_\tau, \tau) = \frac{\gamma\left(v_\tau - \tau a, x_t - t(v_\tau + (1-\tau)a), x_t + (1-t)(v_\tau + (1-\tau)a)\right)}{p_{t,\tau}(x_t, v_\tau)}, \tag{16}$$

given a location (x_t, t, v_τ, τ) induced by linearly interpolating data and velocity from $(v_0, x_0, x_1) \sim \gamma$ drawn from a joint distribution γ that satisfies

$$\int \gamma(v_0, x_0, x_1)dx_0dx_1 = \pi_0(v_0), \int \gamma(v_0, x_0, x_1)dv_0dx_1 = \rho_0(x_0), \int \gamma(v_0, x_0, x_1)dv_0dx_0 = \rho_1(x_1). \tag{17}$$

Here,

$$p_{t,\tau}(x_t, v_\tau) = \int \gamma\left(v_\tau - \tau a, x_t - t(v_\tau + (1-\tau)a), x_t + (1-t)(v_\tau + (1-\tau)a)\right) da. \tag{18}$$

The distribution $p(a; x_t, t, v_\tau, \tau)$ is undefined if $p_{t,\tau}(x_t, v_\tau) = 0$.

Proof. For simplicity, we show the proof for 1D random variables v_0, x_0 and x_1 drawn from a joint distribution $\gamma(v_0, x_0, x_1)$. The joint distribution of a_τ, v_τ, and x_t is $p(a; x_t, t, v_\tau, \tau)p_{t,\tau}(x_t, v_\tau)$,

since $p(a; x_t, t, v_\tau, \tau)$ corresponds to the conditional distribution of the acceleration given locations x_t and v_τ. According to the linear interpolation in Eq. (2), we have

$$\begin{aligned}\begin{bmatrix} v_\tau \\ a \\ x_t \end{bmatrix} &= \begin{bmatrix} 1-\tau & \tau & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v_0 \\ v(x_t,t) \\ x_t \end{bmatrix} = \begin{bmatrix} 1-\tau & \tau & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 1-t & t \end{bmatrix} \begin{bmatrix} v_0 \\ x_0 \\ x_1 \end{bmatrix} \\ &= \begin{bmatrix} 1-\tau & -\tau & \tau \\ -1 & -1 & 1 \\ 0 & 1-t & t \end{bmatrix} \begin{bmatrix} v_0 \\ x_0 \\ x_1 \end{bmatrix} = A \begin{bmatrix} v_0 \\ x_0 \\ x_1 \end{bmatrix}, \end{aligned} \tag{19}$$

where the matrix A has determinant -1. Since $[v_\tau, a, x_t]^T$ is a linear transformation of $[v_0, x_0, x_1]^T$, we have the following expression for the joint distribution of a and (x_t, v_τ):

$$\begin{aligned} p(a; x_t, t, v_\tau, \tau) p_{t,\tau}(x_t, v_\tau) &= \frac{1}{|\det(A)|} \gamma \left(A^{-1} \begin{bmatrix} v_\tau \\ a \\ x_t \end{bmatrix} \right) \\ &= \gamma(v_\tau - \tau a, -t v_\tau - t(1-\tau)a + x_t, (1-t)v_\tau + (1-t)(1-\tau)a + x_t). \end{aligned} \tag{20}$$

After rearranging we get $p(a; x_t, t, v_\tau, \tau) = \frac{\gamma(v_\tau - \tau a, -t v_\tau - t(1-\tau)a + x_t, (1-t)v_\tau + (1-t)(1-\tau)a + x_t)}{p_{t,\tau}(x_t, v_\tau)}$. For the higher dimensional case, the relation in Eq. (20) still holds. This completes the proof. □

Specializing Theorem B.1 to $t = \tau = 0$ under fully independent sampling, $\gamma(v_0, x_0, x_1) = \pi_0(v_0)\rho_0(x_0)\rho_1(x_1)$, the numerator becomes $\pi_0(v_0)\rho_0(x_0)\rho_1(x_0 + v_0 + a)$, and normalizing over a leaves

$$p(a; x_0, 0, v_0, 0) = \rho_1(x_0 + v_0 + a), \tag{21}$$

which is Eq. (10) in the main text: without coupling, the acceleration distribution is a shifted copy of the data distribution, exactly as the velocity distribution is.

Theorem B.1 is stated for a general form of coupling among three random variables V_0, X_0, and X_1. In practice, we focus on hierarchically coupling the data and the velocity, for which we have the following corollary,

Corollary B.2. *The acceleration distribution $p(a; x_t, t, v_\tau, \tau)$ is*

$$p(a; x_t, t, v_\tau, \tau) = \frac{\kappa_{x_t,t}(v_\tau - \tau a, v_\tau + (1-\tau)a)}{\rho_\tau(v_\tau)}, \tag{22}$$

given location (x_t, t, v_τ, τ) induced by linearly interpolating between v_0 and v_1 from $(v_0, v_1) \sim \kappa_{x_t,t}$ drawn from a joint distribution κ that satisfies

$$\int \kappa_{x_t,t}(v_0, v_1) dv_1 = \pi_0(v_0) \quad \textit{and} \quad \int \kappa_{x_t,t}(v_0, v_1) dv_0 = \pi_1(v_1; x_t, t). \tag{23}$$

Here, $\rho_\tau(v_\tau) = \int \kappa_{x_t,t}(v_\tau - \tau a, v_\tau + (1-\tau)a) da$. The distribution $p(a; x_t, t, v_\tau, \tau)$ is undefined if $\rho_\tau(v_\tau) = 0$.

Proof. The proof strategy is similar to the proof of Theorem 3.1, replacing x_t with v_τ, $v(x_t, t)$ with $a(x_t, t, v_\tau, \tau)$, and γ with $\kappa_{x_t,t}$. □

Combining Corollary B.2 with the results of Appendix E, we see that data coupling yields simpler target velocity distributions, and that velocity coupling further concentrates the acceleration distribution; Theorem 3.6 shows that under exact optimal transport in velocity space the distribution in Eq. (22) collapses to a point mass for every $\tau \in [0, 1)$.

C Details on Algorithms

In this section, we provide additional details on HRF2-D and HRF-D&V (see Algorithm 1 and Algorithm 2) and simulation-based and simulation-free velocity coupling methods (see Algorithm 3 and Algorithm 4).

Algorithm 1: HRF2 with Data Coupling

Input : The source distributions ρ_0 and π_0, the dataset $\mathcal{D}$, and the batch size B.

while *stopping conditions not satisfied* **do**
- Sample $\{x_0^{(i)}\}_{i=1}^B \sim \rho_0$, $\{x_1^{(i)}\}_{j=1}^B \sim \mathcal{D}$, and $\{v_0^{(i)}\}_{k=1}^B \sim \pi_0$;
- Sample $\{t^{(i)}\}_{i=1}^B \sim U[0,1]$ and $\{\tau^{(i)}\}_{i=1}^B \sim U[0,1]$;
- Use optimal transport to construct a set of coupled source and target pairs $\{(x_0^{(i)}, x_1^{(\sigma(i))})\}_{i=1}^B$;
- Compute loss following Eq. (11);
- Perform gradient update on θ;

end

Output : θ

Algorithm 2: HRF2 with D&V Coupling

Input : Coupled source and data samples with joint distribution γ

while *stopping conditions not satisfied* **do**
- Sample $(x_0, x_1) \sim \gamma$, and $v_0 \sim \pi_0$;
- Sample $t \sim U[0,1]$ and $\tau \sim U[0,1]$;
- Compute coupled v_0 and $v_1(x_t, t)$ using Algorithm 3 or Algorithm 4;
- Compute loss using Eq. (12);
- Perform gradient update on θ;

end

Output : θ

Algorithm 3: Simulation-based Velocity Coupling via Mini-Batch OT

Input : Location (x_t, t), source distribution π_0, batch size B, and a pre-trained HRF2 model a_θ.

Sample $\{v_0^{(i)}\}_{i=1}^B \sim \pi_0$;

Generate $v_1^{(i)}(x_t, t)$ from $v_0^{(i)}$ via numerically solving Eq. (4) with a pre-trained a_θ for $i = 1, \ldots, B$;

Apply OT to $\{v_0^{(i)}\}_{i=1}^B$ and $\{v_1^{(i)}(x_t, t)\}_{i=1}^B$;

Output : Coupled samples $\{(v_0^{(i)}, v_1^{(\sigma(i))}(x_t, t))\}_{i=1}^B$.

Algorithm 4: Simulation-Free Velocity Coupling via Mini-Batch OT

Input : Location (x_t, t), source distribution π_0, batch size B, radius ϵ, and a data-coupled path pool $\mathcal{D}_\gamma = \{(x_0^{(j)}, x_1^{(j)})\}_{j=1}^M$.

Compute $x_t^{1:M} = (1-t)x_0^{1:M} + t x_1^{1:M}$;

Find local paths $\mathcal{N}_\epsilon(x_t, t) = \{j : \|x_t^{(j)} - x_t\|_2 \le \epsilon\}$;

Compute target velocities $\tilde{v}_1^{(j)} = x_1^{(j)} - x_0^{(j)}$ for $j \in \mathcal{N}_\epsilon(x_t, t)$;

Sample $\{v_0^{(i)}\}_{i=1}^B \sim \pi_0$ and $\{\tilde{v}_1^{(i)}(x_t, t)\}_{i=1}^B$ from $\{\tilde{v}_1^{(j)} : j \in \mathcal{N}_\epsilon(x_t, t)\}$;

Apply OT to $\{v_0^{(i)}\}_{i=1}^B$ and $\{\tilde{v}_1^{(i)}(x_t, t)\}_{i=1}^B$;

Output : Coupled samples $\{(v_0^{(i)}, \tilde{v}_1^{(\sigma(i))}(x_t, t))\}_{i=1}^B$.

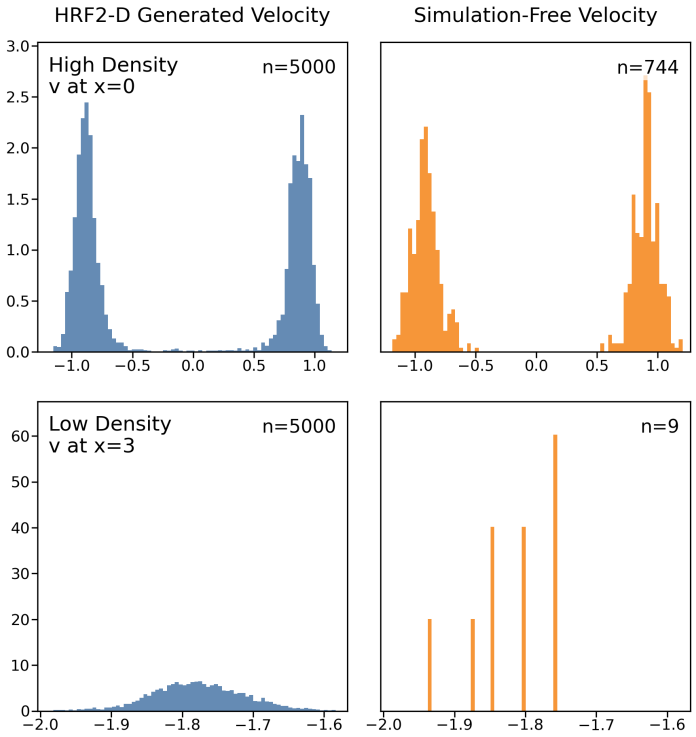


Figure 5: Data-coupled 1D $N \to 2N$ experiment comparing HRF2-D generated velocity and simulation-free velocity at $t = 0$. The matching threshold is $\epsilon = 0.01$.

In Section 3.3, we describe velocity coupling by first obtaining target velocity samples $v_1(x_t, t) \sim \pi_1(v; x_t, t)$ at a fixed space-time location (x_t, t) and then coupling them with source velocity samples $v_0 \sim \pi_0$ using mini-batch OT. In the main experiments, these target velocity samples are generated by numerically integrating the inner HRF2 ODE using a pre-trained HRF2-D model. Here, we focus on an alternative *simulation-free* construction, which obtains local target velocity samples directly from data-coupled interpolation paths, without using a pre-trained velocity generator.

Simulation-based velocity coupling. The simulation-based construction used in the main paper relies on a pre-trained HRF2-D acceleration model $a_\theta(x_t, t, v_\tau, \tau)$. For a fixed (x_t, t), we sample $\{v_0^{(i)}\}_{i=1}^{B_v} \sim \pi_0$ and integrate the velocity-space ODE

$$du_\tau(x_t, t) = a_\theta(x_t, t, u_\tau, \tau)d\tau, \qquad u_0 \sim \pi_0, \tag{24}$$

to obtain generated samples $\{v_1^{(i)}(x_t, t)\}_{i=1}^{B_v}$ from the learned conditional velocity distribution $\pi_1(v; x_t, t)$. We then solve a mini-batch OT problem between $\{v_0^{(i)}\}_{i=1}^{B_v}$ and $\{v_1^{(i)}(x_t, t)\}_{i=1}^{B_v}$ to obtain coupled velocity pairs

$$\{(v_0^{(i)}, v_1^{(\sigma(i))}(x_t, t))\}_{i=1}^{B_v}. \tag{25}$$

These pairs define the empirical coupling $\kappa_{x_t,t}(v_0, v_1)$ used in the HRF2-D&V objective.

Simulation-free velocity coupling. The simulation-free construction replaces the generated velocity samples with local empirical samples from data-coupled straight paths. Let

$$\mathcal{D}_\gamma = \{(x_0^{(j)}, x_1^{(j)})\}_{j=1}^{M}, \qquad (x_0^{(j)}, x_1^{(j)}) \sim \gamma, \tag{26}$$

where γ is the mini-batch OT data coupling used for HRF2-D. For a query location (x_t, t), we evaluate all candidate paths at the same interpolation time:

$$x_t^{(j)} = (1 - t)x_0^{(j)} + tx_1^{(j)}. \tag{27}$$

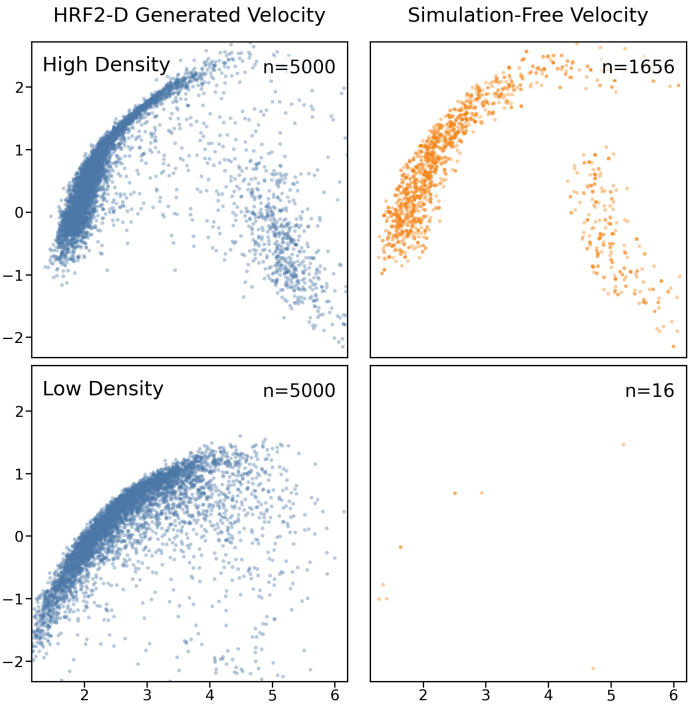


Figure 6: Data-coupled 2D moon experiment comparing HRF2-D generated velocity and simulation-free velocity at $t = 0$. The matching threshold is $\epsilon = 0.3$.

We then collect the paths whose interpolated locations fall inside an ϵ-ball around the query point:

$$\mathcal{N}_\epsilon(x_t, t) = \left\{ j \in \{1, \dots, M\} : \left\| x_t^{(j)} - x_t \right\|_2 \le \epsilon \right\}. \tag{28}$$

For each matched path, the corresponding rectified-flow velocity is available in closed form:

$$v^{(j)} = x_1^{(j)} - x_0^{(j)}, \qquad j \in \mathcal{N}_\epsilon(x_t, t). \tag{29}$$

Thus, the local empirical target velocity distribution is

$$\widehat{\pi}_{1,\epsilon}(v; x_t, t) = \frac{1}{|\mathcal{N}_\epsilon(x_t, t)|} \sum_{j \in \mathcal{N}_\epsilon(x_t, t)} \delta_{x_1^{(j)} - x_0^{(j)}}(v). \tag{30}$$

We then draw a mini-batch of velocities from this empirical distribution and couple them with $\{v_0^{(i)}\}_{i=1}^{B_v} \sim \pi_0$ using the same mini-batch OT step as in the simulation-based method. The resulting coupled pairs are used in the same acceleration matching objective:

$$\min_\theta \mathbb{E}_{(x_0, x_1) \sim \gamma,\, (v_0, v_1) \sim \widehat{\kappa}_{x_t, t},\, t \sim \mathcal{U}[0,1],\, \tau \sim \mathcal{U}[0,1]} \left[\left\| (v_1(x_t, t) - v_0) - a_\theta(x_t, t, v_\tau, \tau) \right\|_2^2 \right], \tag{31}$$

where $\widehat{\kappa}_{x_t, t}$ denotes the empirical OT coupling between velocity samples from a standard Gaussian and simulation-free local target velocities.

Why simulation-free coupling is limited to low dimensional settings. The simulation-free estimator is a local conditional estimator for $\pi_1(v; x_t, t)$. Its reliability depends on the number of matched paths $m_\epsilon(x_t, t) = |\mathcal{N}_\epsilon(x_t, t)|$. For a candidate pool of size M, the expected number of local matches scales approximately as $\mathbb{E}[m_\epsilon(x_t, t)] \approx M \rho_t^\gamma(x_t) \operatorname{Vol}(B_\epsilon^d)$, where ρ_t^γ is the marginal density induced by the data coupling γ at time t and $\operatorname{Vol}(B_\epsilon^d)$ is the volume of a d-dimensional ϵ-ball. This creates a bias-variance trade-off. A small ϵ gives a more faithful estimate of the conditional distribution at the fixed location (x_t, t) but can produce too few samples, especially in low-density regions. A large ϵ provides more samples but mixes velocities from different locations and therefore biases the conditional velocity distribution. Since $\operatorname{Vol}(B_\epsilon^d)$ decays rapidly in high dimensions for

small ϵ, the number of required paths grows quickly with dimension. For this reason, the simulation-free construction is practical in low-dimensional synthetic settings, but not suitable for image-scale or latent image-scale experiments. In those settings, the simulation-based construction is preferred, because the pre-trained HRF2-D model can generate many velocity samples at the exact query location (x_t, t) without relying on local nearest-neighbor matching in data space.

Difference from distillation. The simulation-based construction uses a pre-trained HRF2-D model only as a sampler for the conditional velocity distribution $\pi_1(v; x_t, t)$. It does not regress to the pre-trained model's output vector field, nor does it train a student model to imitate a teacher trajectory. After the velocity samples are obtained, the method constructs a new mini-batch OT coupling in velocity space and trains the HRF2-D&V model by acceleration matching. The simulation-free construction further clarifies this distinction: the same velocity-coupling objective can be formed directly from data-coupled paths without any pre-trained velocity generator. Therefore, velocity coupling is a way to modify the hierarchy-level coupling distribution, rather than a distillation procedure.

Low-dimensional results. We next empirically study when the simulation-free construction provides reliable velocity samples. All experiments evaluate local velocity distributions at $t = 0$. For a query location x_t, the simulation-free method collects data-coupled straight paths whose interpolated locations fall inside a fixed ϵ-ball around x_t, and then uses their path velocities $x_1 - x_0$ as local target velocity samples.

Our first experiment is a controlled 1D $N \to 2N$ setting where the true velocity distribution is analytically available. This experiment is designed to quantify whether simulation-free velocity samples can recover the true local velocity distribution. Figure 7 compares the analytic ground truth, simulation-free velocity samples, and HRF2 simulated velocity samples. The main observation is that, at the high-density location $x_t = 0$, simulation-free matching obtains $n = 4078$ local samples and closely matches the analytic velocity distribution. This is also confirmed quantitatively in Table 2, where the Wasserstein distance between simulation-free velocity and the analytic ground truth is 0.1125 at $x_t = 0$. In contrast, at the low-density tail location $x_t = 3$, the same threshold gives only $n = 47$ local samples, and the empirical velocity distribution becomes unreliable. Across the tail regions, the Wasserstein distance is much larger, e.g., 1.1319 at $x_t = 3$. Thus, this controlled experiment shows that simulation-free velocity can approximate the true velocity distribution in high-density regions, while its accuracy degrades in regions with insufficient local path coverage.

Table 2: Wasserstein distance between simulation-free velocity and the no-OT analytic ground truth on the 1D $N \to 2N$ diagnostic.

x_t	$\rho_t(x_t)$	WD (sim-free, GT)
-3.0	0.0044	1.1474
-2.0	0.0539	1.1537
-1.0	0.2419	0.9000
0.0	0.3989	0.1125
1.0	0.2419	0.9326
2.0	0.0539	1.1264
3.0	0.0044	1.1319

The next two experiments evaluate the data-coupled setting used by HRF2-D&V, where no closed-form ground truth is available. As shown in Figures 5 and 6, simulation-free velocity works well in high-density regions, where many local paths are available ($n = 744$ in 1D and $n = 1656$ in 2D). However, it becomes unreliable in low-density regions, where only a few paths are matched ($n = 9$ in 1D and $n = 16$ in 2D). This local sample-starvation issue becomes much worse in high dimensions. Therefore, simulation-free velocity coupling is useful as a low-dimensional experiment, but simulation-based velocity coupling remains the practical choice for high-dimensional settings.

To further verify the practicality of simulation-free velocity coupling in low dimensions, we compare simulation-based and simulation-free HRF2-D&V using the best Wasserstein distance over the tested time settings and integration allocations for each total NFE. As shown in Table 3, simulation-free velocity coupling achieves performance comparable to simulation-based velocity coupling on both low-dimensional datasets once more than one function evaluation is used. This supports that simulation-free velocity coupling is feasible in low-dimensional settings, although, as discussed above, it does not scale well to high-dimensional data due to local sample starvation.

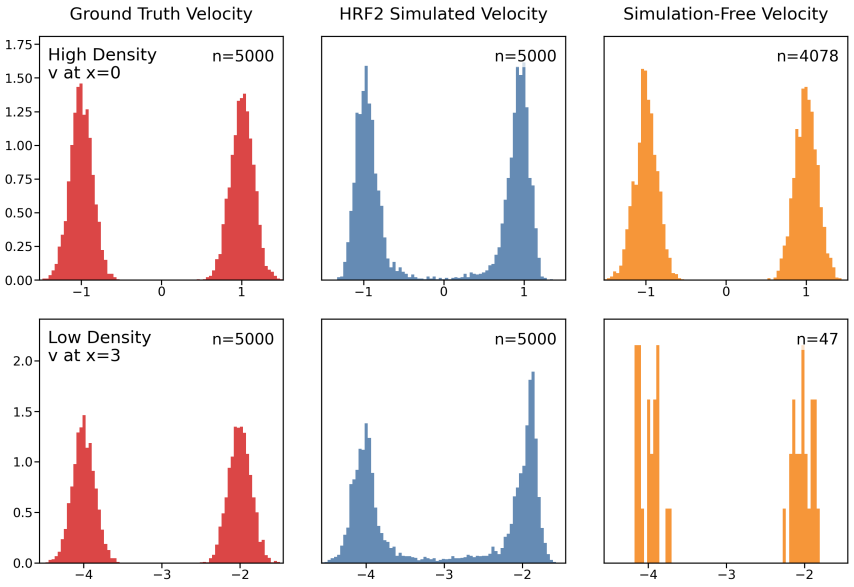


Figure 7: No-OT ground truth, HRF2 simulated velocity, and simulation-free velocity on the 1D $N \to 2N$ dataset at $t = 0$. The high-density point is $x_t = 0$ and the low-density point is $x_t = 3$. The simulation-free pool uses mini-batch OT data coupling with batch size 10 and matching threshold $\epsilon = 0.01$.

Table 3: Low-dimensional generation results comparing simulation-based and simulation-free velocity coupling. For each total NFE, we report the best Wasserstein distance over the test time settings and integration allocations. Lower is better.

Total NFE	$N \to 2N$ Sim.	$N \to 2N$ Sim.-Free	Moon Sim.	Moon Sim.-Free
5	0.0134	0.0099	0.0219	0.0480
10	0.0122	0.0128	0.0267	0.0336
50	0.0103	0.0193	0.0258	0.0576
100	0.0117	0.0113	0.0203	0.0338

D Proof of Theorem 3.2

Proof of Theorem 3.2: We consider the characteristic function of $Z_{t+\Delta t} = Z_t + V\Delta t$ for $t \in [0, 1]$ and $\Delta t \in [0, 1-t]$, assuming that Z_t has the same distribution as X_t. If the characteristic functions of $Z_{t+\Delta t}$ and $X_{t+\Delta t}$ agree, then $Z_{t+\Delta t}$ and $X_{t+\Delta t}$ have the same distribution.

To show this, we evaluate the characteristic function of $Z_{t+\Delta t}$,

$$\begin{aligned}
\mathbb{E}\left[e^{\imath\langle k, Z_{t+\Delta t}\rangle}\right] &= \mathbb{E}_{\rho_t, \pi_1}\left[e^{\imath\langle k, X_t + V\Delta t\rangle}\right] \\
&= \int\int e^{\imath\langle k, x_t + v\Delta t\rangle}\pi_1(v; x_t, t)\rho_t(x_t)dvdx_t \\
&\overset{a}{=} \int\int e^{\imath\langle k, x_t + v\Delta t\rangle}\frac{\gamma(x_t - vt, x_t + (1-t)v)}{\rho_t(x_t)}\rho_t(x_t)dvdx_t \\
&= \int\int e^{\imath\langle k, (x_t + v\Delta t)\rangle}\gamma(x_t - tv, x_t + (1-t)v)dvdx_t \\
&\overset{b}{=} \int\int e^{\imath\langle k, (1-t-\Delta t)x_0 + (t+\Delta t)x_1\rangle}\gamma(x_0, x_1)dx_0dx_1 \\
&= \mathbb{E}_{\rho_{t+\Delta t}}\left[e^{\imath\langle k, X_{t+\Delta t}\rangle}\right].
\end{aligned} \tag{32}$$

We use the notation $\langle \cdot, \cdot \rangle$ to denote the inner product. Equality a is valid due to Theorem 3.1. Equality b holds because $x_0 = x_t - tv$ and $x_1 = x_t + (1-t)v$ due to the linear interpolation. Therefore, we find that $Z_{t+\Delta t}$ and $X_{t+\Delta t}$ follow the same distribution. In addition, since Z_0 and X_0 follow the same distribution ρ_0, we can conclude that Z_t and X_t follow the same marginal distribution at t for $t \in [0, 1]$. This completes the proof. ■

E HIERARCHICAL SIMPLIFICATION OF THE VELOCITY AND ACCELERATION DISTRIBUTIONS

This section proves the results stated in Section 3.5. Throughout, ρ_0, ρ_1 and π_0 have finite second moments and ρ_0 and π_0 are absolutely continuous; π_0 has mean m_0 and covariance Σ_{π_0} (for $\pi_0 = \mathcal{N}(0, I_d)$, $m_0 = 0$ and $\operatorname{tr} \Sigma_{\pi_0} = d$). Stronger hypotheses are stated where they are used. We write $v_1 = x_1 - x_0$ for the velocity, $a = v_1 - v_0$ for the acceleration, $\bar{v}(x_t, t) = \mathbb{E}[v_1 \mid x_t]$ for the conditional mean velocity, $\mathcal{T}(\gamma) = \mathbb{E}_\gamma \|x_1 - x_0\|^2$ for the transport cost, and recall

$$\mathcal{C}_1(\gamma) = \int_0^1 \mathbb{E}\big[\operatorname{tr} \operatorname{Cov}(v_1 \mid x_t)\big]\, dt, \qquad \mathcal{C}_2(\gamma, \kappa) = \int_0^1 \int_0^1 \mathbb{E}\big[\operatorname{tr} \operatorname{Cov}(a \mid x_t, v_\tau)\big]\, d\tau\, dt. \tag{33}$$

Conditioning on x_t and v_τ is understood at fixed t and τ. A data coupling γ is admissible if $\gamma \in \Gamma(\rho_0, \rho_1)$, and a velocity coupling κ is admissible if, for ρ_t-almost every x_t, $\kappa_{x_t,t}$ has marginals π_0 and $\pi_1(\cdot\,; x_t, t)$ and depends measurably on (x_t, t).

E.1 THE TWO FUNCTIONALS ARE THE OPTIMAL TRAINING LOSSES

Lemma E.1. *Let Y be square integrable and $\mathcal{G}$ a sub-σ-algebra. Then $\inf_f \mathbb{E}\|Y - f\|^2 = \mathbb{E}[\operatorname{tr} \operatorname{Cov}(Y \mid \mathcal{G})]$, where the infimum is over $\mathcal{G}$-measurable f and is attained at $f = \mathbb{E}[Y \mid \mathcal{G}]$.*

Proof. Write $Y - f = (Y - \mathbb{E}[Y \mid \mathcal{G}]) + (\mathbb{E}[Y \mid \mathcal{G}] - f)$. The cross term vanishes because $\mathbb{E}[Y \mid \mathcal{G}] - f$ is $\mathcal{G}$-measurable and $\mathbb{E}\big[Y - \mathbb{E}[Y \mid \mathcal{G}] \mid \mathcal{G}\big] = 0$, so $\mathbb{E}\|Y - f\|^2 = \mathbb{E}\|Y - \mathbb{E}[Y \mid \mathcal{G}]\|^2 + \mathbb{E}\|\mathbb{E}[Y \mid \mathcal{G}] - f\|^2$. The tower rule turns the first term into $\mathbb{E}[\operatorname{tr} \operatorname{Cov}(Y \mid \mathcal{G})]$. □

Applying Lemma E.1 with $Y = v_1$, $\mathcal{G} = \sigma(x_t, t)$ and with $Y = a$, $\mathcal{G} = \sigma(x_t, t, v_\tau, \tau)$ shows that $\mathcal{C}_1(\gamma)$ is the optimal value of the velocity-matching objective in Eq. (3) and $\mathcal{C}_2(\gamma, \kappa)$ is the optimal value of the acceleration-matching objective in Eq. (12). HRF2 trains only the latter; $\mathcal{C}_1$ enters it through Theorem 3.4 and Proposition 3.5.

E.2 DATA COUPLING SIMPLIFIES THE VELOCITY DISTRIBUTION

Proof of Theorem 3.3. Since ρ_0 is absolutely continuous, Brenier's theorem gives $\gamma^{\mathrm{OT}} = (\mathrm{id}, T)_{\#}\rho_0$ with $T = \nabla\phi$ for a convex ϕ, defined ρ_0-almost everywhere; no assumption on ρ_1 beyond a finite second moment is needed. Let E be a Borel set of full ρ_0-measure on which ϕ is differentiable. For $t \in [0, 1)$, the interpolation map

$$S_t(x) = (1-t)x + tT(x) = \nabla\Big(\tfrac{1-t}{2}\|x\|^2 + t\phi(x)\Big), \qquad x \in E, \tag{34}$$

satisfies $\langle S_t(x) - S_t(y), x - y\rangle \geq (1-t)\|x - y\|^2 > 0$ for $x \neq y$ in E, by monotonicity of $\nabla\phi$. Hence S_t is injective on E. As an injective Borel map on a Borel set, it has a Borel image and a Borel inverse (Lusin–Souslin), so $x_0 = S_t^{-1}(x_t)$ is $\sigma(x_t)$-measurable, and so is $v_1 = T(x_0) - x_0$. Therefore $\operatorname{Cov}(v_1 \mid x_t) = 0$ and $\pi_1(\cdot\,; x_t, t)$ is a Dirac mass for ρ_t-almost every x_t. Integrating over $t \in [0, 1)$ gives $\mathcal{C}_1(\gamma^{\mathrm{OT}}) = 0$. □

Proposition E.2. *For every batch size B, the mini-batch coupling γ_B is admissible, and its expected transport cost satisfies $\mathcal{T}(\gamma_{mB}) \leq \mathcal{T}(\gamma_B)$ for every positive integer m.*

Proof. Admissibility. Draw a batch $\{x_0^{(i)}\}_{i=1}^B$ i.i.d. from ρ_0 and $\{x_1^{(i)}\}_{i=1}^B$ i.i.d. from ρ_1, let σ be the optimal assignment, and let I be uniform on $\{1, \ldots, B\}$ and independent of the batch. Then γ_B is the law of $(x_0^{(I)}, x_1^{(\sigma(I))})$. The first coordinate is uniform over an i.i.d. sample, so its law is

ρ_0. Conditionally on the batch, σ is a bijection, so $\sigma(I)$ is uniform and the second coordinate has conditional law $\frac{1}{B}\sum_j \delta_{x_1^{(j)}}$; taking expectations gives ρ_1.

Monotonicity. We have $\mathcal{T}(\gamma_B) = \mathbb{E}\big[\frac{1}{B}\min_\sigma \sum_i \|x_0^{(i)} - x_1^{(\sigma(i))}\|^2\big] = \mathbb{E}\big[W_2^2(\hat{\rho}_0^B, \hat{\rho}_1^B)\big]$. Partition a batch of size mB into m blocks of size B. Any product of within-block permutations is feasible for the size-mB assignment problem, so the optimal cost on the large batch is at most the sum of the m optimal within-block costs. The blocks are i.i.d. batches of size B, so taking expectations and dividing by mB gives $\mathcal{T}(\gamma_{mB}) \le \mathcal{T}(\gamma_B)$. □

Theorem 3.3 controls $\mathcal{C}_1$ only in the exact-transport limit, whereas Proposition E.2 concerns $\mathcal{T}$. The two functionals are related exactly.

Lemma E.3. *For every $\gamma \in \Gamma(\rho_0, \rho_1)$,*

$$\mathcal{C}_1(\gamma) = \mathcal{T}(\gamma) - \int_0^1 \mathbb{E}\|\bar{v}(x_t, t)\|^2\, dt \;\le\; \mathcal{T}(\gamma) - \|m_{\rho_1} - m_{\rho_0}\|^2, \tag{35}$$

where m_{ρ_i} is the mean of ρ_i. The subtracted term on the right is determined by the marginals alone.

Proof. Fix t. The law of total variance applied to v_1 gives $\mathbb{E}\|v_1\|^2 = \mathbb{E}\|\bar{v}(x_t,t)\|^2 + \mathbb{E}[\operatorname{tr}\operatorname{Cov}(v_1 \mid x_t)]$. The left-hand side equals $\mathcal{T}(\gamma)$ and does not depend on t, so integrating over $t \in [0,1]$ gives the identity. For the inequality, Jensen's inequality gives $\mathbb{E}\|\bar{v}(x_t,t)\|^2 \ge \|\mathbb{E}\bar{v}(x_t,t)\|^2 = \|\mathbb{E}v_1\|^2 = \|m_{\rho_1} - m_{\rho_0}\|^2$ for every t. □

Combined with Proposition E.2, $\mathcal{T}(\gamma_B) - \|m_{\rho_1} - m_{\rho_0}\|^2$ is a non-increasing upper bound on $\mathcal{C}_1(\gamma_B)$ along divisor chains. This does not establish monotonicity of $\mathcal{C}_1(\gamma_B)$ itself, and the bound does not tend to zero: $\mathcal{T}(\gamma_B) \to W_2^2(\rho_0, \rho_1)$, while the subtracted term is in general strictly smaller. The Jensen step is lossy precisely at the optimal-transport end, where v_1 is a function of x_t and the subtracted term in the identity equals $\mathcal{T}(\gamma^{\mathrm{OT}})$ in full. Proposition E.4 closes the gap under a regularity hypothesis.

Proposition E.4. *Let $v^\star(x_t, t) = T(S_t^{-1}x_t) - S_t^{-1}x_t$ be the optimal-transport velocity field, with T and S_t as in the proof of Theorem 3.3, and suppose $v^\star(\cdot, t)$ is L-Lipschitz on the support of $\rho_t^{\gamma^{\mathrm{OT}}}$, uniformly in $t \in [0,1)$. Then for every $\gamma \in \Gamma(\rho_0, \rho_1)$,*

$$\mathcal{C}_1(\gamma) \le 2(1+L)^2\, W_2^2(\gamma, \gamma^{\mathrm{OT}}), \tag{36}$$

the Wasserstein distance being taken between measures on $\mathbb{R}^d \times \mathbb{R}^d$. In particular $\mathcal{C}_1(\gamma_B) \to 0$ as $B \to \infty$.

Proof. By Kirszbraun's theorem, $v^\star(\cdot, t)$ extends to an L-Lipschitz map on $\mathbb{R}^d$; we use the same symbol for the extension, which is needed because x_t under γ need not lie in the support of $\rho_t^{\gamma^{\mathrm{OT}}}$. By Lemma E.1, $\mathcal{C}_1(\gamma)$ is an infimum over measurable velocity fields, so it is bounded by the value at the competitor $v^\star$: $\mathcal{C}_1(\gamma) \le \int_0^1 \mathbb{E}_\gamma\|v_1 - v^\star(x_t,t)\|^2\, dt$. Couple γ and γ^{OT} optimally, writing (x_0, x_1) and $(x_0^\star, x_1^\star)$ for the two pairs and $\Delta_i = x_i - x_i^\star$, so that $\mathbb{E}[\|\Delta_0\|^2 + \|\Delta_1\|^2] = W_2^2(\gamma, \gamma^{\mathrm{OT}})$. By Theorem 3.3, $x_1^\star - x_0^\star = v^\star(x_t^\star, t)$ almost surely, and $x_t - x_t^\star = (1-t)\Delta_0 + t\Delta_1$. Hence

$$\|v_1 - v^\star(x_t,t)\| \le \|\Delta_0\| + \|\Delta_1\| + L\|x_t - x_t^\star\| \le (1+L)\big(\|\Delta_0\| + \|\Delta_1\|\big). \tag{37}$$

Squaring, applying $(\alpha+\beta)^2 \le 2(\alpha^2+\beta^2)$ and integrating gives Eq. (36).

For the limit, the empirical measures $\hat{\rho}_i^B$ converge to ρ_i weakly with convergence of second moments almost surely, and the optimal plan between ρ_0 and ρ_1 is unique because ρ_0 is absolutely continuous, so the empirical optimal plans $\hat{\gamma}_B$ converge to γ^{OT} in W_2 almost surely. Since $W_2^2(\hat{\gamma}_B, \gamma^{\mathrm{OT}})$ is dominated by twice the sum of the second moments of the two plans, which are uniformly integrable, it also converges in expectation. Finally $\gamma_B = \mathbb{E}[\hat{\gamma}_B]$ and $W_2^2(\cdot, \nu)$ is convex, so $W_2^2(\gamma_B, \gamma^{\mathrm{OT}}) \le \mathbb{E}[W_2^2(\hat{\gamma}_B, \gamma^{\mathrm{OT}})] \to 0$. □

Remark E.5 (On the Lipschitz hypothesis). If $\lambda I \preceq \nabla^2\phi \preceq \Lambda I$, then, since $v^\star(\cdot,t) = (T - \mathrm{id}) \circ S_t^{-1}$ and $\nabla S_t \succeq \min(1,\lambda) I$, one may take $L = \max(|\Lambda - 1|, |1 - \lambda|)/\min(1,\lambda)$. Global two-sided Hessian bounds of this kind require more than densities bounded above and below: they follow

from the regularity theory of Caffarelli (1996) for Hölder-continuous densities on uniformly convex domains. More importantly, for well-separated multimodal targets the optimal-transport map is steep between modes, and it is discontinuous when the support of ρ_1 is disconnected, so L is large or infinite in the regime that motivates this work. The rate implied by Proposition E.4 is also governed by empirical W_2 rates and degrades with dimension. Proposition E.6 is the dimension-free complement.

Proposition E.6. *Let $X_0, X_0' \sim \rho_0$ and $X_1, X_1' \sim \rho_1$ be independent. Then*

$$\mathcal{T}(\gamma_1) - \mathcal{T}(\gamma_2) = \tfrac{1}{2}\,\mathbb{E}\big|\langle X_0 - X_0',\, X_1 - X_1'\rangle\big|, \tag{38}$$

which is strictly positive whenever ρ_1 is not a point mass.

Proof. On a batch of two, write $C_{\mathrm{id}} = \|x_0 - x_1\|^2 + \|x_0' - x_1'\|^2$ and $C_{\mathrm{sw}} = \|x_0 - x_1'\|^2 + \|x_0' - x_1\|^2$ for the two feasible assignments. Expanding the squares, the four norm terms cancel and

$$C_{\mathrm{id}} - C_{\mathrm{sw}} = -2\langle x_0 - x_0',\, x_1 - x_1'\rangle. \tag{39}$$

By exchangeability $\mathbb{E}C_{\mathrm{id}} = \mathbb{E}C_{\mathrm{sw}}$, so $\mathcal{T}(\gamma_1) = \frac{1}{2}\mathbb{E}\big[\frac{C_{\mathrm{id}}+C_{\mathrm{sw}}}{2}\big]$, whereas mini-batch OT selects the cheaper assignment and $\mathcal{T}(\gamma_2) = \frac{1}{2}\mathbb{E}[\min(C_{\mathrm{id}}, C_{\mathrm{sw}})]$. The identity follows from $\frac{\alpha+\beta}{2} - \min(\alpha,\beta) = \frac{|\alpha-\beta|}{2}$.

For strictness, $X_0 - X_0'$ is a difference of independent absolutely continuous vectors and therefore has a density, so for any fixed $b \neq 0$ the inner product $\langle X_0 - X_0', b\rangle$ is almost surely non-zero. Since ρ_1 is not a point mass, $X_1 - X_1' \neq 0$ with positive probability, and conditioning on (X_1, X_1') gives a strictly positive expectation. Absolute continuity of ρ_0 cannot be replaced by non-degeneracy alone: if ρ_0 and ρ_1 are supported on orthogonal lines, the inner product vanishes identically. □

For $\rho_0 = \mathcal{N}(0, I_d)$ the gain has a closed form: conditionally on the targets, $\langle X_0 - X_0', X_1 - X_1'\rangle \sim \mathcal{N}(0, 2\|X_1 - X_1'\|^2)$, and $\mathbb{E}|Z| = s\sqrt{2/\pi}$ for $Z \sim \mathcal{N}(0, s^2)$, so the gain equals $\pi^{-1/2}\,\mathbb{E}\|X_1 - X_1'\|$, which grows like $\sqrt{d}$. Since $\mathcal{T}(\gamma_1) = d + \operatorname{tr}\Sigma_{\rho_1} + \|m_{\rho_1}\|^2$, the relative gain decays like $d^{-1/2}$ rather than exponentially. Through Lemma E.3, this transport-cost gain is also a reduction of the upper bound on $\mathcal{C}_1$. The gain at batch size two is a lower bound on the improvement at larger batch sizes, not a characterization of it.

Theorem F.1 in Appendix F is a one-dimensional, non-asymptotic counterpart of Theorem 3.3. Because our analysis controls dispersion rather than mode count, we record the link between the two.

Proposition E.7. *Suppose $\pi_1(\cdot\,; x_t, t) = p\,\mu_1 + (1-p)\,\mu_2$ with $p \in (0,1)$ and with component means separated by Δ. Then $\operatorname{tr}\operatorname{Cov}(v_1 \mid x_t) \geq p(1-p)\Delta^2$. Consequently, if the conditional dispersion is below $p(1-p)\Delta^2$, the velocity distribution admits no decomposition into two components with weights at least p whose means are Δ-separated.*

Proof. Let $J \in \{1, 2\}$ be the mixture index. The law of total variance gives $\operatorname{tr}\operatorname{Cov}(v_1) \geq \operatorname{tr}\operatorname{Cov}(\mathbb{E}[v_1 \mid J])$, and the right-hand side is the variance of a two-point distribution taking the two component means with probabilities p and $1-p$, namely $p(1-p)\|m_{\mu_1} - m_{\mu_2}\|^2$. □

E.3 An exact decomposition across the two levels

In the statement of Theorem 3.4, $\mathcal{K}(\gamma, \kappa) = \int_0^1 \mathbb{E}[c_{x_t,t}]\,dt$ with $c_{x_t,t} = \mathbb{E}_{\kappa_{x_t,t}}\langle v_0 - m_0,\, v_1 - \bar{v}(x_t, t)\rangle$, and $\mathcal{D}(\gamma, \kappa) = \int_0^1\int_0^1 \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(\mathbb{E}[a \mid x_t, v_\tau] \mid x_t)\big]\,d\tau\,dt \geq 0$.

Proof of Theorem 3.4. Fix t and x_t. Because the first marginal of $\kappa_{x_t,t}$ is π_0 for every x_t, $\mathbb{E}[v_0 \mid x_t] = m_0$ and $\operatorname{Cov}(v_0 \mid x_t) = \Sigma_{\pi_0}$. Bilinearity of covariance applied to $a = v_1 - v_0$ gives

$$\operatorname{tr}\operatorname{Cov}(a \mid x_t) = \operatorname{tr}\operatorname{Cov}(v_1 \mid x_t) + \operatorname{tr}\Sigma_{\pi_0} - 2c_{x_t,t}, \tag{40}$$

the two cross terms being transposes of one another. The right-hand side does not depend on τ. The law of total variance with $\sigma(x_t) \subseteq \sigma(x_t, v_\tau)$ gives

$$\mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(a \mid x_t, v_\tau)\big] = \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(a \mid x_t)\big] - \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(\mathbb{E}[a \mid x_t, v_\tau] \mid x_t)\big]. \tag{41}$$

Integrating over τ and t and substituting Eq. (40) gives the decomposition; $\mathcal{D} \geq 0$ because it is an expectation of the trace of a covariance. □

The level-one dispersion enters with coefficient one, and $\operatorname{tr}\Sigma_{\pi_0}$ is the price of introducing a new source distribution at the second level. Both $\mathcal{K}$ and $\mathcal{D}$ depend on γ as well as on κ, so the identity does not by itself imply that reducing $\mathcal{C}_1$ reduces $\mathcal{C}_2$. The next lemma identifies the velocity coupling that is optimal for the decomposition.

Lemma E.8. *Fix an admissible γ. Among admissible velocity couplings, the family $\kappa^\star = \{\kappa^\star_{x_t,t}\}$ of optimal quadratic-cost couplings of π_0 and $\pi_1(\cdot\,; x_t, t)$ maximizes $c_{x_t,t}$ for every (x_t, t), and hence maximizes $\mathcal{K}(\gamma,\cdot)$ and minimizes $\mathcal{C}_2(\gamma,\cdot) + \mathcal{D}(\gamma,\cdot)$.*

Proof. Fix (x_t, t). For any coupling of π_0 and $\pi_1(\cdot\,; x_t, t)$, $\mathbb{E}\|v_1 - v_0\|^2 = \mathbb{E}\|v_0\|^2 + \mathbb{E}\|v_1\|^2 - 2\,\mathbb{E}\langle v_0, v_1\rangle$, and the first two terms depend only on the marginals. Hence the optimal coupling maximizes $\mathbb{E}\langle v_0, v_1\rangle$. Since $\mathbb{E}v_0 = m_0$ and $\mathbb{E}v_1 = \bar v(x_t, t)$, $c_{x_t,t} = \mathbb{E}\langle v_0, v_1\rangle - \langle m_0, \bar v(x_t, t)\rangle$, and the subtracted term also depends only on the marginals, so $\kappa^\star$ maximizes $c_{x_t,t}$. Integrating gives the claim for $\mathcal{K}$, and $\mathcal{C}_2 + \mathcal{D} = \mathcal{C}_1(\gamma) + \operatorname{tr}\Sigma_{\pi_0} - 2\mathcal{K}$ by Theorem 3.4, where $\mathcal{C}_1(\gamma)$ does not depend on κ. Admissibility of $\kappa^\star$ is shown in the proof of Theorem 3.6. □

E.4 Velocity coupling simplifies the acceleration distribution

Recall that $a^\star(x_t, t, v_\tau, \tau) = \mathbb{E}[a \mid x_t, v_\tau]$ at fixed (t, τ).

Proof of Proposition 3.5. At $\tau = 0$ we have $v_\tau = v_0$, so $a^\star(x_t, t, v_0, 0) = \mathbb{E}[a \mid x_t, v_0]$ and $\hat u_1 - v_1 = v_0 + \mathbb{E}[a \mid x_t, v_0] - v_1 = \mathbb{E}[a \mid x_t, v_0] - a$. Taking conditional expectations given x_t and using the tower rule,

$$\mathbb{E}\big[\|\hat u_1 - v_1\|^2 \mid x_t\big] = \mathbb{E}\Big[\mathbb{E}\big[\|a - \mathbb{E}[a \mid x_t, v_0]\|^2 \mid x_t, v_0\big]\,\Big|\,x_t\Big] = \mathbb{E}\big[\operatorname{tr}\operatorname{Cov}(a \mid x_t, v_0) \mid x_t\big]. \quad (42)$$

Under independent velocity sampling, v_0 and v_1 are conditionally independent given x_t, so $\mathbb{E}[a \mid x_t, v_0] = \bar v(x_t, t) - v_0$ and $\hat u_1 = \bar v(x_t, t)$. Moreover $\operatorname{Cov}(a \mid x_t, v_0) = \operatorname{Cov}(v_1 \mid x_t, v_0) = \operatorname{Cov}(v_1 \mid x_t)$, whose expectation integrates over t to $\mathcal{C}_1(\gamma)$. □

At a total NFE of one, the outer step then gives $x_0 + \bar v(x_0, 0) = \mathbb{E}[x_1 \mid x_0]$ under γ; under independent coupling this is the data mean for every x_0.

Proof of Theorem 3.6. Admissibility of $\kappa^\star$. For each (x_t, t), the optimal coupling of π_0 and $\pi_1(\cdot\,; x_t, t)$ is unique because π_0 is absolutely continuous. Since $(x_t, t) \mapsto \pi_1(\cdot\,; x_t, t)$ is a measurable kernel, a measurable selection of optimal plans exists (Villani, 2009, Cor. 5.22), and by uniqueness it coincides with $\kappa^\star$ almost everywhere; hence $\kappa^\star$ is admissible.

Collapse. Fix (x_t, t). Brenier's theorem with π_0 as the source, which requires nothing of the target beyond a finite second moment, gives $v_1 = T_{x_t,t}(v_0)$ with $T_{x_t,t} = \nabla\psi$ for a convex ψ. For $\tau \in [0, 1)$,

$$v_\tau = (1-\tau)v_0 + \tau T_{x_t,t}(v_0) = \nabla\Big(\tfrac{1-\tau}{2}\|v\|^2 + \tau\psi(v)\Big)\Big|_{v=v_0}, \quad (43)$$

which is injective on a set of full π_0-measure by the argument in the proof of Theorem 3.3, with a Borel inverse. Therefore v_0 is recoverable from v_τ, and $a = T_{x_t,t}(v_0) - v_0$ is a measurable function of (x_t, v_τ), so $\operatorname{Cov}(a \mid x_t, v_\tau) = 0$. Integrating over $\tau \in [0, 1)$ and t gives $\mathcal{C}_2(\gamma, \kappa^\star) = 0$. At $\tau = 0$ the slice vanishes, so $\hat u_1 = v_1$ almost surely by Proposition 3.5: a single inner Euler step is exact. The argument covers the degenerate case $\pi_1(\cdot\,; x_t, t) = \delta_c$ produced by strong data coupling, for which $T_{x_t,t} \equiv c$ and $a = c - v_0$. □

The proof of Theorem 3.6 uses only that π_0 is absolutely continuous. For Corollary E.9 we take $\pi_0 = \mathcal{N}(m_0, \Sigma_{\pi_0})$ with $\Sigma_{\pi_0} \succ 0$, and define $\operatorname{mmse}_{x,t}(\sigma)$ as the minimum mean-squared error of estimating $v_1 \sim \pi_1(\cdot\,; x, t)$ from $v_1 + \sigma\xi$, with $\xi \sim \mathcal{N}(0, \Sigma_{\pi_0})$ independent of v_1; in the main text $\pi_0 = \mathcal{N}(0, I_d)$ and $\xi \sim \mathcal{N}(0, I_d)$. Write $\kappa^\perp$ for independent velocity pairing.

Corollary E.9. *Let $\pi_0 = \mathcal{N}(m_0, \Sigma_{\pi_0})$ with $\Sigma_{\pi_0} \succ 0$. For every admissible data coupling γ,*

$$\mathcal{C}_2(\gamma, \kappa^\perp) \;\geq\; \int_0^1\!\!\int_0^1 \mathbb{E}\big[\operatorname{mmse}_{x_t,t}(\sigma_\tau)\big]\, d\tau\, dt, \qquad \sigma_\tau = \frac{1-\tau}{\tau}, \quad (44)$$

and the right-hand side is strictly positive whenever $\mathcal{C}_1(\gamma) > 0$. In particular, if ρ_0 has an everywhere positive density and ρ_1 is not a point mass, then $\mathcal{C}_2(\gamma^\perp, \kappa^\perp) > 0$; $\mathcal{C}_2(\gamma_B, \kappa^\perp) > 0$ whenever $\mathcal{C}_1(\gamma_B) > 0$; and $\mathcal{C}_2(\gamma, \kappa^\star) = 0$ for every γ. Joint data and velocity coupling therefore yields a strictly more concentrated acceleration distribution than no coupling and than data coupling alone.

Proof. The equality $\mathcal{C}_2(\gamma, \kappa^\star) = 0$ is Theorem 3.6. For the lower bound, fix t and x_t and work under $\kappa^\perp$, so that $v_0 \sim \pi_0$ and $v_1 \sim \pi_1(\cdot\,; x_t, t)$ are independent given x_t.

Reduction to denoising. For $\tau \in (0,1)$, the relation $v_\tau = (1-\tau)v_0 + \tau v_1$ gives $a = (v_1 - v_\tau)/(1-\tau)$, so $\mathrm{Cov}(a \mid x_t, v_\tau) = (1-\tau)^{-2}\,\mathrm{Cov}(v_1 \mid x_t, v_\tau)$. Moreover v_τ is an invertible affine function of $Y_\tau := v_1 + \sigma_\tau(v_0 - m_0)$ with $\sigma_\tau = (1-\tau)/\tau$, and $v_0 - m_0 \sim \mathcal{N}(0, \Sigma_{\pi_0})$ is independent of v_1 given x_t. Hence

$$\mathbb{E}\big[\,\mathrm{tr}\,\mathrm{Cov}(a \mid x_t, v_\tau)\,\big|\,x_t\big] = \frac{\mathrm{mmse}_{x_t,t}(\sigma_\tau)}{(1-\tau)^2} \;\geq\; \mathrm{mmse}_{x_t,t}(\sigma_\tau). \tag{45}$$

Taking expectations over x_t and integrating over τ and t gives Eq. (44).

Monotonicity and limit. For $\sigma < \sigma'$, the observation $v_1 + \sigma'\xi$ has the same law as $v_1 + \sigma\xi + \sqrt{\sigma'^2 - \sigma^2}\,\xi'$, with ξ' an independent copy of ξ. It is therefore a degraded version of $v_1 + \sigma\xi$, and $\sigma \mapsto \mathrm{mmse}_{x,t}(\sigma)$ is nondecreasing. As $\sigma \to \infty$ the observation becomes uninformative and $\mathrm{mmse}_{x,t}(\sigma) \to \mathrm{tr}\,\mathrm{Cov}\,\pi_1(\cdot\,; x, t)$ (Guo et al., 2005). Since $\sigma_\tau \uparrow \infty$ as $\tau \downarrow 0$, the map $\tau \mapsto \mathbb{E}[\mathrm{mmse}_{x_t,t}(\sigma_\tau)]$ is nonincreasing and, by monotone convergence, tends to $\mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(v_1 \mid x_t)]$ as $\tau \downarrow 0$.

Positivity. If $\mathcal{C}_1(\gamma) > 0$, then $\mathbb{E}[\mathrm{tr}\,\mathrm{Cov}(v_1 \mid x_t)] > 0$ for a set of t of positive measure. For each such t, the integrand in τ is positive on an interval $(0, \tau_0(t))$, so the right-hand side of Eq. (44) is positive.

The two particular cases. The case γ_B is the hypothesis $\mathcal{C}_1(\gamma_B) > 0$. For $\gamma^\perp$, suppose in addition that ρ_0 has an everywhere positive density, as for the Gaussian source used in all experiments, and that ρ_1 is not a point mass. For $t \in (0,1)$, the posterior of x_1 given x_t is proportional to $\rho_1(x_1)\,\rho_0\big((x_t - t x_1)/(1-t)\big)$, whose support is that of ρ_1, so $\mathrm{Cov}(x_1 \mid x_t) \neq 0$. Since $v_1 = \big(x_1 - x_t\big)/(1-t)$, $\mathrm{Cov}(v_1 \mid x_t) = (1-t)^{-2}\mathrm{Cov}(x_1 \mid x_t) \neq 0$ for every x_t and every $t \in (0,1)$, hence $\mathcal{C}_1(\gamma^\perp) > 0$. □

Under independent velocity pairing, the $\tau = 0$ slice of $\mathcal{C}_2$ therefore equals $\mathcal{C}_1(\gamma)$, and Corollary E.9 shows that the slices near $\tau = 0$ remain bounded below by a denoising error of π_1. Data coupling reduces both quantities by concentrating the velocity distribution but leaves them positive at finite batch size, consistent with the empirical observation that HRF2-D degrades sharply when the total NFE is reduced to one, whereas HRF2-D&V does not.

E.5 Bias induced by simulated target velocity samples

Theorem 3.6 assumes that the velocity coupling is built from the exact conditional law $\pi_1(\cdot\,; x_t, t)$. In high dimensions we instead draw target velocities from a pre-trained HRF2-D model, so the model used at inference produces velocities from an approximation $\hat{\pi}_1$. We bound the resulting error for the sampler that is actually run: the outer level is integrated with N Euler steps, and at every outer step a fresh velocity is produced by the inner solver.

Setup. Let $h = 1/N$ and $t_k = kh$. At outer step k a fresh $u_0 \sim \pi_0$ is drawn and the inner ODE is solved; write $\hat{\pi}_1(\cdot\,; x, t)$ for the conditional law of its output. The sampler is

$$\hat{Z}_0 \sim \rho_0, \qquad \hat{Z}_{k+1} = \hat{Z}_k + h\,\hat{V}_k, \qquad \hat{V}_k \mid \mathcal{F}_k \sim \hat{\pi}_1(\cdot\,; \hat{Z}_k, t_k), \tag{46}$$

where $\mathcal{F}_k$ is generated by the history up to step k. The ideal sampler (Z_k) is defined by the same recursion with π_1 in place of $\hat{\pi}_1$. We write $m(x,t) = \mathbb{E}_{\pi_1(\cdot;x,t)}[v]$, which equals $\bar{v}(x,t)$, and $\hat{m}(x,t) = \mathbb{E}_{\hat{\pi}_1(\cdot;x,t)}[v]$.

Assumption E.10. **(a)** $\varepsilon := \sup_{x,k} W_2\big(\hat{\pi}_1(\cdot\,; x, t_k),\, \pi_1(\cdot\,; x, t_k)\big) < \infty$, and we set $b := \sup_{x,k} \|\hat{m}(x, t_k) - m(x, t_k)\|$. **(b)** The kernel $\pi_1(\cdot\,; x, t)$ is defined for every $x \in \mathbb{R}^d$ and is W_2-Lipschitz on the grid: $W_2\big(\pi_1(\cdot\,; x, t_k), \pi_1(\cdot\,; y, t_k)\big) \leq L\|x - y\|$.

By Jensen's inequality $b \le \varepsilon$, and since $\|m(x,t) - m(y,t)\| \le W_1 \le W_2$, the mean field $m(\cdot, t_k)$ is L_m-Lipschitz with $L_m \le L$. Assumption E.10(b) contains the Lipschitz-mean-field hypothesis as a special case: under exact optimal transport $\pi_1(\cdot\,; x, t)$ is a Dirac mass (Theorem 3.3) and the two coincide. Under independent coupling at $t = 0$, $\pi_1(\cdot\,; x, 0) = \rho_1(x + \cdot)$ is a translate, so $L = 1$ exactly.

Lemma E.11 (Discrete marginal exactness)**.** *For every N, the ideal sampler satisfies $\mathrm{Law}(Z_k) = \rho_{t_k}$ for $k = 0, \dots, N$.*

Proof. The proof of Theorem 3.2 shows that if $Z_t \sim \rho_t$ and $V \mid Z_t \sim \pi_1(\cdot\,; Z_t, t)$, then $Z_t + V\Delta t \sim \rho_{t+\Delta t}$ for every $\Delta t \in [0, 1-t]$. Apply this with $\Delta t = h$ and induct on k. □

Proposition E.12. *Under Assumption E.10, with $c = 1 + 2L_m$,*

$$W_2\big(\mathrm{Law}(\hat{Z}_N), \rho_1\big) \;\le\; \min\left\{ \varepsilon\,\frac{(1 + L/N)^N - 1}{L}\,,\; \sqrt{\frac{e^c - 1}{c}}\,\sqrt{b^2 + \frac{\varepsilon^2 e^{2L}}{N}} \right\}. \tag{47}$$

Proof. Coupling. Construct $(Z_k, \hat{Z}_k)$ on a common probability space with $Z_0 = \hat{Z}_0$. At step k, conditionally on $\mathcal{F}_k$, draw $(V_k, \hat{V}_k)$ from an optimal W_2 coupling of $\pi_1(\cdot\,; Z_k, t_k)$ and $\hat{\pi}_1(\cdot\,; \hat{Z}_k, t_k)$; a measurable choice exists by Villani (2009, Cor. 5.22). Each coordinate has the correct conditional law, so $\hat{Z}_N$ has the law of the sampler Eq. (46), and $\mathrm{Law}(Z_N) = \rho_1$ by Lemma E.11. Hence $W_2(\mathrm{Law}(\hat{Z}_N), \rho_1) \le e_N := \|\hat{Z}_N - Z_N\|_{L^2}$. Write $\Delta_k = \hat{Z}_k - Z_k$.

Velocity mismatch. By the triangle inequality for W_2 and Assumption E.10,

$$\mathbb{E}\big[\|\hat{V}_k - V_k\|^2 \mid \mathcal{F}_k\big] = W_2^2\big(\pi_1(\cdot\,; Z_k, t_k), \hat{\pi}_1(\cdot\,; \hat{Z}_k, t_k)\big) \le \big(L\|\Delta_k\| + \varepsilon\big)^2, \tag{48}$$

so by Minkowski's inequality $\|\hat{V}_k - V_k\|_{L^2} \le L e_k + \varepsilon$.

First bound. Since $\Delta_{k+1} = \Delta_k + h(\hat{V}_k - V_k)$, $e_{k+1} \le (1 + hL)e_k + h\varepsilon$ with $e_0 = 0$, and summing the geometric series gives $e_N \le \varepsilon\big((1 + hL)^N - 1\big)/L$. Because $(1 + L/N)^N \le e^L$, the same argument gives $e_k \le \varepsilon(e^L - 1)/L$ for every k, and therefore $Le_k + \varepsilon \le \varepsilon e^L$.

Second bound. Expanding the square,

$$e_{k+1}^2 = e_k^2 + 2h\,\mathbb{E}\langle \Delta_k, \hat{V}_k - V_k\rangle + h^2\|\hat{V}_k - V_k\|_{L^2}^2. \tag{49}$$

Conditioning on $\mathcal{F}_k$, $\mathbb{E}[\hat{V}_k - V_k \mid \mathcal{F}_k] = \hat{m}(\hat{Z}_k, t_k) - m(Z_k, t_k)$, whose norm is at most $b + L_m\|\Delta_k\|$. Using $\mathbb{E}\|\Delta_k\| \le e_k$, $2be_k \le b^2 + e_k^2$, and the bound on the velocity mismatch from the first part,

$$e_{k+1}^2 \le e_k^2\,(1 + hc) + h\big(b^2 + h\varepsilon^2 e^{2L}\big). \tag{50}$$

Iterating from $e_0 = 0$,

$$e_N^2 \le \big(b^2 + h\varepsilon^2 e^{2L}\big)\,\frac{(1 + hc)^N - 1}{c} \le \big(b^2 + h\varepsilon^2 e^{2L}\big)\,\frac{e^c - 1}{c}. \tag{51}$$

Taking square roots and the minimum of the two bounds gives Eq. (47). □

At $N = 1$ the first bound equals ε. As $N \to \infty$ the full distributional error ε enters only at rate $N^{-1/2}$, while the mean error b persists: zero-mean fluctuations of the velocity sampler average out across outer steps, whereas a biased mean accumulates.

Corollary E.13 (Sources of the velocity error)**.** *Consider models whose acceleration fields are learned exactly.*

(i) HRF2-D with M inner steps. *Write $\varepsilon_{\mathrm{D}}(M)$ and $b_{\mathrm{D}}(M)$ for the constants of Assumption E.10 for its velocity sampler. For $M = 1$,*

$$b_{\mathrm{D}}(1) = 0, \qquad \varepsilon_{\mathrm{D}}(1) = \sup_{x,k} \sqrt{\mathrm{tr}\,\mathrm{Cov}\,\pi_1(\cdot\,; x, t_k)}. \tag{52}$$

(ii) HRF2-D&V *whose velocity pairs are constructed as above, with an HRF2-D base model using $\widetilde{M}$ inner steps and trained under the same data coupling γ that defines the recorded locations, and let $\tilde{\pi}_1 = \pi_1^{\mathrm{D},\widetilde{M}}$ be the law of its velocity samples. For M inner steps at inference, let $\eta(M) := \sup_{x,k} W_2\big(\hat{\pi}_1^{(M)}(\cdot\,; x, t_k), \tilde{\pi}_1(\cdot\,; x, t_k)\big)$ collect the learning error of HRF2-D&V (including the finite number J of recorded locations) and its inner discretization error. Then*

$$\varepsilon_{\mathrm{DV}}(M) \le \varepsilon_{\mathrm{D}}(\widetilde{M}) + \eta(M), \qquad b_{\mathrm{DV}}(M) \le b_{\mathrm{D}}(\widetilde{M}) + \eta(M). \tag{53}$$

*If the acceleration field of HRF2-D&V is learned exactly, then: **(a)** under the optimal velocity coupling $\kappa^\star$, $\eta(M) = 0$ for every $M \ge 1$; **(b)** under the mini-batch velocity coupling κ_{B_v} with finite B_v,*

$$\eta(1)^2 \;\le\; \sup_{x,k} \mathbb{E}_{\kappa_{B_v}}\big[\,\mathrm{tr}\,\mathrm{Cov}(a \mid x_t = x, v_0)\big] \;\le\; \sup_{x,k} \mathrm{tr}\,\mathrm{Cov}\,\tilde{\pi}_1(\cdot\,; x, t_k), \tag{54}$$

with $a = \tilde{v}_1 - v_0$, and $\eta(M) \to 0$ as $M \to \infty$ whenever the inner acceleration field is Lipschitz in v_τ.

Proof. (i) By Proposition 3.5, under independent velocity sampling the one-step output is $\hat{u}_1 = \bar{v}(x_t, t) = m(x_t, t)$, so the velocity law is $\delta_{m(x_t,t)}$. Its mean is $m(x_t, t)$, and its W_2 distance to $\pi_1(\cdot\,; x_t, t)$ is $\big(\mathbb{E}_{\pi_1}\|v - m\|^2\big)^{1/2} = \big(\,\mathrm{tr}\,\mathrm{Cov}\,\pi_1(\cdot\,; x_t, t)\big)^{1/2}$.

(ii) Eq. (53) follows from the triangle inequality for W_2, and for the means from $\|\mathbb{E}_\mu v - \mathbb{E}_\nu v\| \le W_2(\mu, \nu)$. For (a), Theorem 3.6 requires only that π_0 be absolutely continuous and that the target have a finite second moment. Applied with target $\tilde{\pi}_1$, the exactly learned inner ODE under $\kappa^\star$ is solved exactly by a single Euler step, and hence by any number of steps, so the output law is $\tilde{\pi}_1$ for every $M \ge 1$. For (b), κ_{B_v} is admissible, with marginals π_0 and $\tilde{\pi}_1(\cdot\,; x_t, t)$, by the argument in the proof of Proposition E.2. Proposition 3.5 holds for any admissible coupling and any target law; applied with $\tilde{\pi}_1$ and κ_{B_v}, it gives $\mathbb{E}[\|\hat{u}_1 - \tilde{v}_1\|^2 \mid x_t] = \mathbb{E}_{\kappa_{B_v}}[\mathrm{tr}\,\mathrm{Cov}(a \mid x_t, v_0) \mid x_t]$. Since $\hat{u}_1$ and $\tilde{v}_1$ are both functions of the coupled pair, this is an upper bound on $W_2^2\big(\mathrm{Law}(\hat{u}_1 \mid x_t), \tilde{\pi}_1(\cdot\,; x_t, t)\big)$, which gives the first inequality in Eq. (54). The second follows from $\mathrm{Cov}(a \mid x_t, v_0) = \mathrm{Cov}(\tilde{v}_1 \mid x_t, v_0)$ and the law of total variance. Finally, the exactly learned field is the rectified-flow field of the coupling κ_{B_v}, so its flow maps π_0 to $\tilde{\pi}_1$ for any coupling (Liu et al., 2023, Thm. 3.3), and Euler's method converges to this flow when the field is Lipschitz, so $\eta(M) \to 0$. □

Remark E.14 (One-step samples and the crossover). By Corollary E.13(i), HRF2-D at a total NFE of one outputs $\hat{Z}_1 = \mathbb{E}[x_1 \mid x_0 = Z_0]$, the conditional mean of the data given the source sample under γ, which is the blur visible in Figure 3(a). At small outer budgets the first bound in Eq. (47) is the relevant one, and the comparison is between $\varepsilon_{\mathrm{D}}(M)$ and $\varepsilon_{\mathrm{D}}(\widetilde{M}) + \eta(M)$; at $M = 1$ the former is the entire spread of the velocity distribution, while under velocity coupling $\eta(1)$ is bounded by the one-step dispersion of the coupled velocities (Eq. (54)). At large outer budgets only the mean error survives, and HRF2-D&V carries a floor proportional to $b_{\mathrm{D}}(\widetilde{M}) + \eta(M)$. The first term is fixed offline, when the velocity pairs are generated, and does not depend on the inference budget; the second is HRF2-D&V's own inner-solve error. These statements are upper bounds, so they are consistent with the crossover observed in Section 4.2.

In the exact-transport limit of the velocity coupling, Corollary E.13 predicts that HRF2-D&V with $(N, 1)$ outer and inner steps matches HRF2-D with $(N, \widetilde{M})$, so that velocity coupling amortizes the base model's inner integration into a single step. At the finite B_v used in practice this holds only up to $\eta(1)$, which is bounded by Eq. (54) and closes as M grows. Comparing HRF2-D&V at (N, M) for small M with HRF2-D at $(N, \widetilde{M})$ therefore measures the finite-batch effect directly.

F Mini-batch OT analysis on 1D Gaussian Mixtures

Here we use 1D distributions to illustrate how mini-batch OT in the data space simplifies the velocity distributions. We have the following result:

Theorem F.1. *Let ρ_0 be a standard Gaussian distribution on $\mathbb{R}$, and let $\rho_1 = \frac{1}{2}\mathcal{N}(-1, \sigma^2) + \frac{1}{2}\mathcal{N}(1, \sigma^2)$, where $\sigma \ll 1$. Let the batch size be $B = 2k$ for some positive integer k. Let $\delta_1 = \delta_1(k)$ and $\delta_2 = \delta_2(k)$ be non-negative functions of k satisfying $\delta_i \to 0$ and $\delta_i^2 k \to \infty$ as $k \to \infty$ for $i \in \{1, 2\}$; for instance $\delta_1 = \delta_2 = k^{-1/4}$. Then there exist positive constants c_1, c_2 independent of k, such that with probability at least $1 - 2\exp(-c_1\delta_1^2 k) - 2\exp(-c_2\delta_2^2 k)$, the velocity distribution $\pi_1(v; x, t)$ induced by mini-batch optimal transport matching is unimodal for all $x \notin [l, u]$,where the endpoints $l = l(k, \delta_1, \delta_2, t)$ and $u = u(k, \delta_1, \delta_2, t)$ depend on k, δ_1, δ_2, and the interpolation time t. Moreover, the probability that $x \in [l, u]$ is given by $\int_l^u \rho_t(x)dx$. As $k \to \infty$, the failure probability tends to zero, the interval $[l, u]$ collapses to a single point, namely the median $F^{-1}(1/2)$ of ρ_0, and the velocity distribution $\pi_1(v; x, t)$ is unimodal for almost every x.*

The proof is given below and relies on order statistics; the theorem is extended to a mixture of K Gaussians in Corollary F.5. Appendix B analyzes the corresponding acceleration distributions. Note that the two requirements on δ_i pull in opposite directions and must both be met: $\delta_i \to 0$ is what makes the exceptional interval collapse, while $\delta_i^2 k \to \infty$ is what makes the concentration bounds non-trivial. Choosing δ_i that tend to zero slowly, such as $k^{-1/4}$, satisfies both.

Theorem F.1 is the one-dimensional, non-asymptotic counterpart of Theorem 3.3, which gives the corresponding statement in arbitrary dimension without any assumption on ρ_1.

Proof of Theorem F.1:

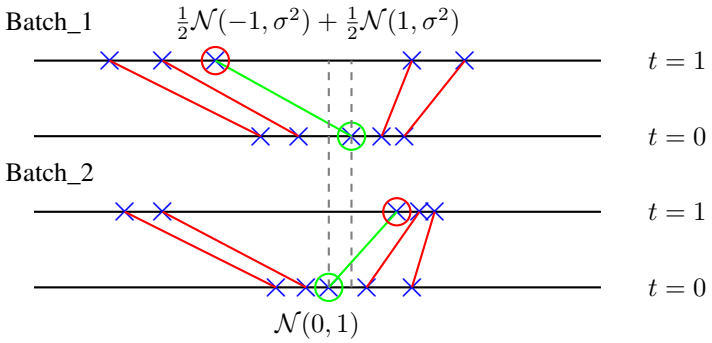


Figure 8: OT mapping of points between $t = 1$ and $t = 0$ for two mini-batches in red line. The green-circled samples are $x_0^{(J_1)}$ and $x_0^{(J_2+1)}$. The velocity distribution is $\pi(v; x, 0)$ is multimodal for $x \in \left[x_0^{(J_2+1)}, x_0^{(J_1)}\right]$.

Let $\{x_0^i\}_{i=1}^B$ and $\{x_1^i\}_{i=1}^B$ be a mini-batch of data points drawn independently from the source distribution ρ_0 and the target ρ_1 respectively. We assume that the target mixture distribution is well-separated such that for each batch the largest sample drawn from the left mode $\mathcal{N}(-1, \sigma^2)$ is smaller than the smallest sample drawn from the right mode $\mathcal{N}(1, \sigma^2)$ with high probability.

Let $\{x_0^{(i)}\}_{i=1}^B$ and $\{x_1^{(i)}\}_{i=1}^B$ be the respective order statistics. That is $x_0^{(1)} \le x_0^{(2)} \le \cdots \le x_0^{(B)}$ and $x_1^{(1)} \le x_1^{(2)} \le \cdots \le x_1^{(B)}$. Thus the optimal transport maps $x_0^{(i)}$ to $x_1^{(i)}$ for all $i \in [B]$. We define a jump as $J = \min\{i \in [B] : x_1^{(i)} \sim \mathcal{N}(-1, \sigma^2), x_1^{(i+1)} \sim \mathcal{N}(1, \sigma^2)\}$.

We first consider velocity v at $t = 0$. Note that for each mini-batch, $v^{(b)}(x_0^{(i)}, 0) = x_1^{(i)} - x_0^{(i)}$ is always unimodal after an OT match, as it is deterministic. However, multimodality takes place if $\{v^{(b)}(x_0, 0)\}_{b\in[M]}$ points to different GMM modes in a neighborhood of x_0. In other words, multimodality of $\pi(v; x_0, 0)$ is in fact caused by randomness among different batches. For example in two different batches, whenever $x_0^{(J_1)} > x_0^{(J_2+1)}$, $\pi(v; x_0, 0)$ is multimodal in this region, see Figure 8 for an intuitive graphical illustration. Thus, we aim to control $x_0^{(J)}$ to restrict the region of x_0 that have multimodal velocity distributions. To control $x_0^{(J)}$, we start with providing a concentration bound on J in the following lemma.

Lemma F.2. *For $\delta_1 > 0$ there exists $c_1 > 0$ such that with probability at least $1 - 2\exp(-c_1\delta_1^2 k)$,*

$$|J - k| \le \delta_1 k. \tag{55}$$

Proof. By definition, J is essentially the number of samples drawn from the first Gaussian mode, which follows a $\mathrm{Binomial}(B, 1/2)$. Thus, $\mathbb{E}J = k$. By Chernoff's inequality (Vershynin, 2018), we have for $\delta_1 \in (0, 1]$,

$$\mathbb{P}(|X - \mu| \ge \delta_1\mu) \le 2\exp(-c_1\mu\delta_1^2),$$

where X is binomial and $\mu = \mathbb{E}X$. Thus, with probability at least $1 - 2\exp(-c_1\delta_1^2 k)$, we have $|J - k| \le \delta_1 k$. □

The above lemma suggests that we lower bound $x_0^{((1-\delta_1)k)}$ and upper bound $x_0^{((1+\delta_1)k)}$ respectively, where we utilize the following lemma.

Lemma F.3. *Let $X_1, \dots, X_N$ be i.i.d. random variables with CDF F. Let $X_{(1)}, \dots, X_{(N)}$ be the order statistics. Then $\forall r \in [N]$,*

$$F(X_{(r)}) \sim \mathrm{Beta}(r, N + 1 - r). \tag{56}$$

Proof. Let F and f be the CDF and PDF of X. We first show the CDF of the r-th order statistic $X_{(r)}$.

$$F_{(r)}(x) = \mathbb{P}(\text{at least } r \ X_i \le x) = \sum_{j=r}^{N} \binom{N}{j} F(x)^j [1 - F(x)]^{N-j}.$$

Hence the PDF is

$$f_{(r)}(x) = \frac{d}{dx} F_{(r)}(x) = \frac{N!}{(r-1)!(N-r)!} F(x)^{r-1}[1 - F(x)]^{N-r} f(x).$$

Let $U = F(X_{(r)})$, so that $U \in [0, 1]$. Using change of variable, $u = F(x)$, the PDF f_U is

$$\begin{aligned} f_U(u) &= f_{(r)}(x) \left| \frac{dx}{du} \right|_{x=F^{-1}(u)} \\ &\propto F(x)^{r-1}[1 - F(x)]^{N-r} f(x) \frac{1}{f(x)} \\ &\propto u^{r-1}(1 - u)^{N-r}, \end{aligned}$$

which is a $\mathrm{Beta}(r, N + 1 - r)$ distribution. □

Lemma F.3 shows that $\mathbb{P}(X_{(r)} \ge a) = \mathbb{P}(F(X_{(r)}) \ge F(a))$, where the later tail probability could be bounded using concentration of the Beta distribution. We state it in the following lemma, which is a direct result of (Skorski, 2023).

Lemma F.4. *Let F be the CDF of a continuous random variable x. For some $\delta_1 \in (0, 1)$ and $\delta_2 > 0$, let $l = F^{-1}(\mathbb{E}(F(x^{((1-\delta_1)k)})) - \delta_2)$ and $u = F^{-1}(\mathbb{E}(F(x^{((1+\delta_1)k)})) + \delta_2)$. Then there exist positive constants C_1, C_2 such that with probability at least $1 - \exp(-C_1\delta_2^2 k)$ and $1 - \exp(-C_2\delta_2^2 k)$: $x^{((1-\delta_1)k)} > l$ and $x^{((1+\delta_1)k)} < u$ respectively. Further, $\lim_{k\to\infty} l = \lim_{k\to\infty} u$ if δ_1, δ_2 are chosen such that $\delta_i \to 0$ and $\delta_i^2 k \to \infty$ when $k \to \infty$ for $i \in \{1, 2\}$.*

Proof. The proof is an application of Theorem 1 in (Skorski, 2023). We assume $(1 \pm \delta_1)k$ are integers. Now $U_1 = F(x_{((1-\delta_1))k}) \sim \mathrm{Beta}((1 - \delta_1)k, (1 + \delta_1)k + 1)$. Thus

$$\mathbb{P}(U_1 < \mathbb{E}U_1 - \delta_2) \le \exp(-\delta_2^2/2\theta),$$

where $\theta = \frac{(1-\delta_1)k[(1+\delta_1)k+1]}{(2k+1)^2(2k+2)} \simeq C_1/k$ for some absolute constant $C_1 > 0$. Denoting $l = F^{-1}(\mathbb{E}(F(x^{((1-\delta_1)k)})) - \delta_2)$, we have with probability at least $1 - \exp(-C_1\delta_2^2 k)$, $x^{((1-\delta_1)k)} > l$. A similar proof leads to the upper bound u.

If δ_1, δ_2 are chosen such that $\delta_i \to 0$ and $\delta_i^2 k \to \infty$ when $k \to \infty$, then $(1 \pm \delta_1)k \sim k$ and, since $\mathbb{E}(F(x^{(r)})) = r/(B + 1)$ by Lemma F.3, both $\mathbb{E}(F(x^{((1-\delta_1)k)}))$ and $\mathbb{E}(F(x^{((1+\delta_1)k)}))$ converge to $1/2$. Thus $l = u = F^{-1}(1/2)$ in the limit, i.e. the median of ρ_0, which coincides with $\mathbb{E}x$ here by symmetry of the Gaussian. □

Combining Lemma F.2 and Lemma F.4, the multimodal region of $\pi_1(v; x_0, 0)$ is bounded by $[l, u]$ with probability at least $1 - 2\exp(-c_1\delta_1^2 k) - 2\exp(-c_2\delta_2^2 k)$ for some absolute constant c_1, c_2. The probability that $x_0 \in [l, u]$ is $\int_l^u \rho_0(x)dx$. By further choosing δ_1, δ_2 such that $\delta_i \to 0$ and $\delta_i^2 k \to \infty$ when $k \to \infty$, $l = u$ in the limit, eliminating all multimodality almost surely.

Since there is a bijection between (x_0, x_1) and (x_t, x_1) for all $t \in (0, 1)$, OT matching has the same effect on (x_t, x_1) as on (x_0, x_1). The joint distribution of (x_t, x_1) is uniquely determined by (x_0, x_1). Thus the proof on (x_0, x_1) can be used for (x_t, x_1), completing the proof of Theorem F.1.

A direct extension of Theorem F.1 is as follows.

Corollary F.5. *Let the target ρ_1 be a uniform mixture of K well-separated one-dimensional Gaussian components,*

$$\rho_1 = \frac{1}{K}\sum_{i=1}^{K} \mathcal{N}(\mu_i, \sigma^2),$$

where the means $\{\mu_i\}_{i=1}^K$ are pairwise well separated and $\sigma \ll 1$. Under the same settings as in Theorem F.1, let the batch size be $B = 2k$, and let $\delta_1 = \delta_1(k)$ and $\delta_2 = \delta_2(k)$ be non-negative functions of k satisfying $\delta_i \to 0$ and $\delta_i^2 k \to \infty$ when $k \to \infty$ for $i \in \{1, 2\}$. Then there exist positive constants c_1, c_2, independent of k, such that with probability at least $1 - 2(K-1)\exp(-c_1\delta_1^2 k) - 2(K-1)\exp(-c_2\delta_2^2 k)$, the velocity distribution $\pi_1(v; x, t)$ induced by mini-batch optimal transport matching is unimodal for $x \notin \bigcup_{i=1}^{K-1}[l_i, u_i]$, where the interval end points $\{l_i\}_{i=1}^{K-1}$ and $\{u_i\}_{i=1}^{K-1}$ depend on k, δ_1, δ_2, and t.

In addition, the probability that x lies in the exceptional region is given by

$$\mathbb{P}\left(x \in \bigcup_{i=1}^{K-1}[l_i, u_i]\right) = \sum_{i=1}^{K-1}\int_{l_i}^{u_i} \rho_t(x)dx.$$

As $k \to \infty$, each interval $[l_i, u_i]$ collapses to a single point (i.e., $l_i = u_i$ for all $i \in \{1, \ldots, K-1\}$, and the velocity distribution $\pi(v; x, t)$ are unimodal almost surely.

Proof. In view of Theorem F.1, we prove the Corollary by induction. We first view $\mathcal{N}(\mu_1, \sigma^2)$ as one mode and the rest jointly as another. Note that this grouping is unbalanced, with weights $(1/K, (K-1)/K)$ rather than $(1/2, 1/2)$, and that the right group is not a single Gaussian. Theorem F.1 applies to this setting once stated for general weights $(p, 1-p)$: the jump index becomes $J \sim \mathrm{Binomial}(B, p)$ with $\mathbb{E}J = pB$, the Chernoff bound of Lemma F.2 is unchanged, Lemma F.4 requires only that F be continuous, and the limiting collapse point becomes $F^{-1}(p)$. Applying Theorem F.1 in this form, we have with probability at least $1 - 2\exp(-c_1\delta_1^2 k) - 2\exp(-c_2\delta_2^2 k)$ the multimodal region of v is bounded by (l_1, u_1). We then consider the left two modes $\mathcal{N}(\mu_1, \sigma^2)$ and $\mathcal{N}(\mu_2, \sigma^2)$ jointly as one mode, and the rest jointly as another. Again applying Theorem F.1 we have with probability at least $1 - 2\exp(-c_1\delta_1^2 k) - 2\exp(-c_2\delta_2^2 k)$ the multimodal region of v is bounded by (l_2, u_2). Thus by induction,

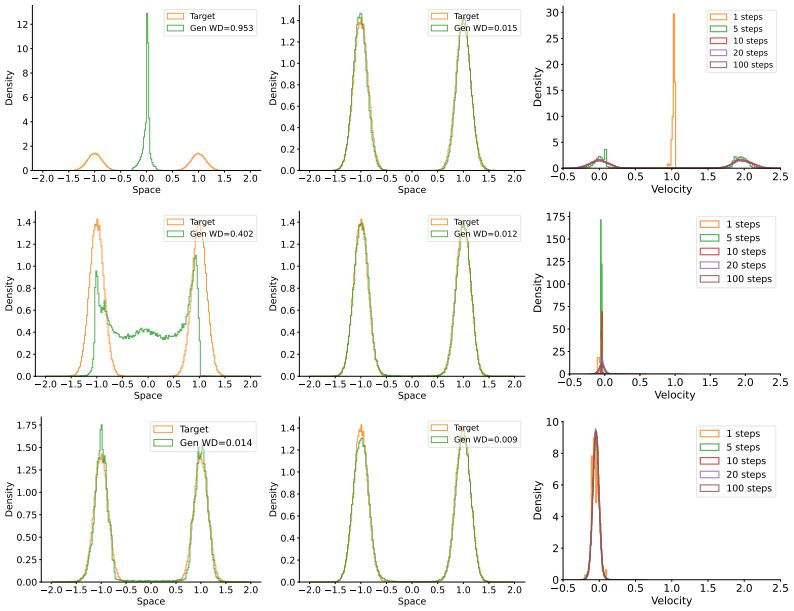
$$\begin{aligned}\mathbb{P}\left(v \text{ is multimodal for some } x_t \notin \bigcup_{i=1}^{K-1}[l_i, u_i]\right) &\le \sum_{i=1}^{K-1}\mathbb{P}\left(\text{concentration fails at boundary } i\right) \\ &\le (K-1)[2\exp(-c_1\delta_1^2 k) + 2\exp(-c_2\delta_2^2 k)].\end{aligned}$$

In the above derivation, the first inequality uses the union bound over the $K-1$ boundaries, and the second uses the result of Theorem F.1. Note that multimodality occurs *inside* the exceptional region, so the union bound is taken over the per-boundary concentration failure events rather than over its complement. Similar to the proof of Lemma F.4, $l_i = u_i$ when $k \to \infty$. □

G Additional Experimental Results

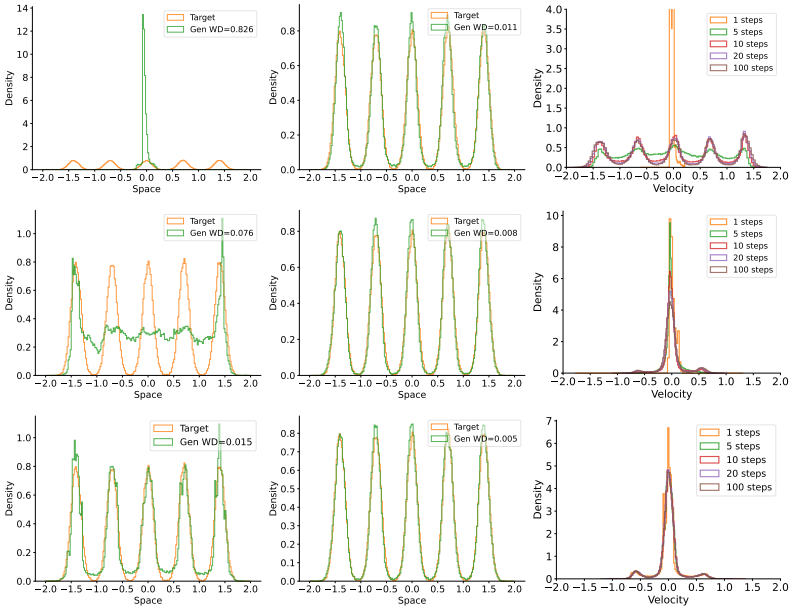
G.1 Synthetic Data Results

We present more results on synthetic data: Figure 9 for 1D $\mathcal{N} \to 2\mathcal{N}$ data, Figure 10 for 1D $\mathcal{N} \to 5\mathcal{N}$ data, Figure 11 for $8\mathcal{N} \to$ moon data, and Figure 12 for 2D $\mathcal{N} \to 6\mathcal{N}$ data. As shown in Figure 11,



(a) Data distribution total NFEs 1 (b) Data distribution total NFEs 100 (c) Velocity distribution

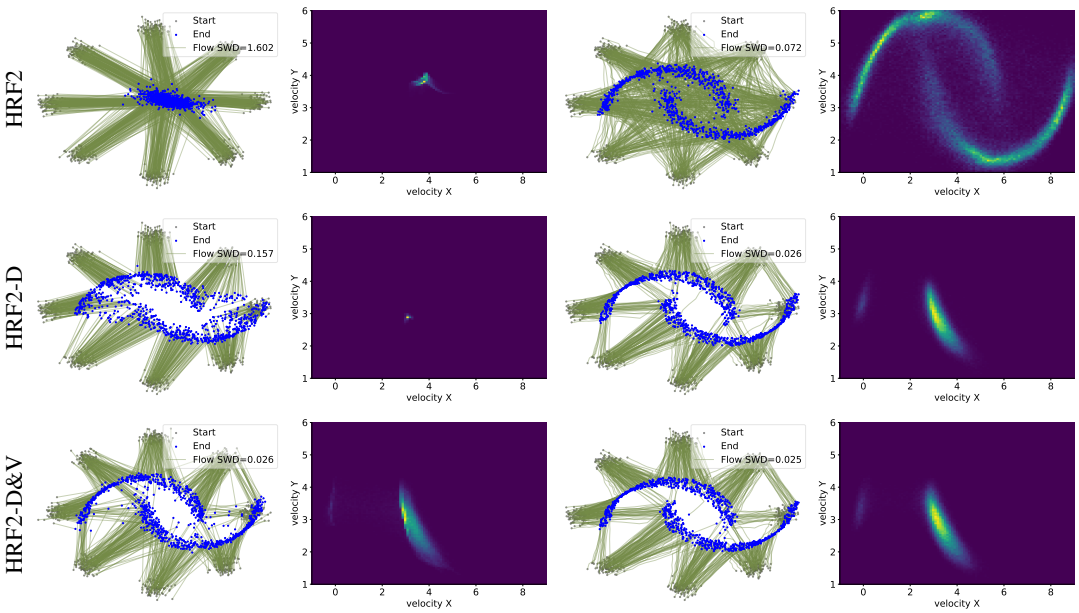
Figure 9: Results on 1D $\mathcal{N} \to 2\mathcal{N}$ data. The three rows correspond to HRF2, HRF2 with data coupling, HRF2 with data & velocity coupling. (a) and (b) are generated data distribution with total NFEs 1 and 100. (c) is velocity distribution at $(x_t, t) = (-1, 0)$.



(a) Data distribution total NFEs 1 (b) Data distribution total NFEs 100 (c) Velocity distribution

Figure 10: Results on 1D $\mathcal{N} \to 5\mathcal{N}$ data. The three rows correspond to HRF2, HRF2 with data coupling, HRF2 with data & velocity coupling. (a) and (b) are generated data distribution with total NFEs 1 and 100. (c) is velocity distribution at $(x_t, t) = (0, 0)$.

we extend the 1D findings in Figure 1 to 2D data. The velocity distribution at $t = 0$ is simply a



(a) Data distribution total NFE=1 (b) Velocity distribution 1 step integration (c) Data distribution total NFEs=100 (d) Velocity distribution 100 step integration

Figure 11: Results on $8\mathcal{N} \to$ moon dataset. Three rows are HRF2, HRF2 with data coupling, HRF2 with data & velocity coupling. (a) and (c) are trajectories (green) of sample particles flowing from source distribution (grey) to target distribution (blue) with total NFEs 1 and 100. (b) and (d) are velocity distributions at the center of the bottom left Gaussian mode at $t = 0$. Data coupling simplifies the velocity distribution and velocity coupling reduces the required sampling steps.

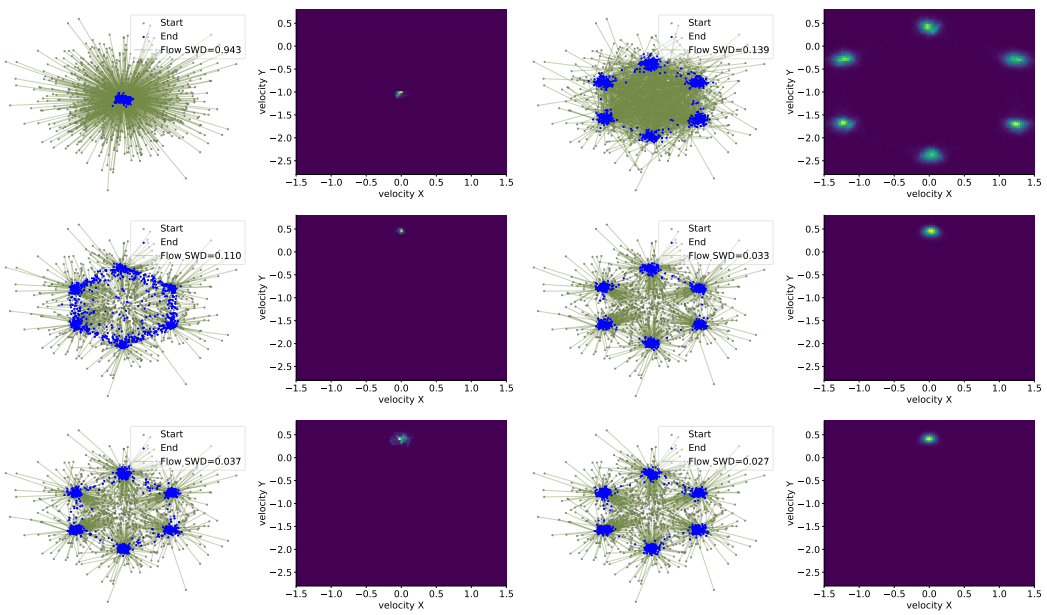
Table 4: FID performance on MNIST under different total NFEs. **Bold** for the best per NFE.

Method	**NFE** $= 5$	**NFE** $= 10$	**NFE** $= 20$	**NFE** $= 50$	**NFE** $= 100$
RF	19.187 ± 0.188	7.974 ± 0.119	6.151 ± 0.090	5.605 ± 0.057	5.563 ± 0.049
OT-CFM	13.977 ± 0.166	4.477 ± 0.099	2.763 ± 0.036	2.321 ± 0.038	2.346 ± 0.023
HRF2	15.798 ± 0.151	6.644 ± 0.076	3.408 ± 0.076	2.664 ± 0.058	2.588 ± 0.075
HRF2-D	10.167 ± 0.136	**3.823 ± 0.038**	**2.318 ± 0.053**	**1.929 ± 0.031**	**1.847 ± 0.011**
HRF2-D&V	**5.519 ± 0.112**	3.861 ± 0.089	3.720 ± 0.045	3.604 ± 0.016	3.423 ± 0.003

shifted version of the data distribution. Data coupling makes the velocity distribution more unimodal, effectively aligning with a portion of the target distribution. For example, in Figure 11(d), we observe that the velocity distribution primarily consists of the region of the target distribution closest to x_t. In contrast, velocity coupling does not modify the velocity distribution itself but significantly reduces the number of required sampling steps. As shown in Figure 11(b), a single integration step already produces a reasonable velocity distribution with joint data and velocity couplings. Across all these experiments, we consistently observe that data coupling simplifies the velocity distribution, while velocity coupling significantly reduces the required sampling steps. This corroborates the findings discussed in the main paper.

G.2 Image Data Results

We present more results on image data: Table 4 for MNIST, Table 5 for CIFAR-10, Table 6 for CelebA-HQ 256, and Table 7 for ImageNet 256, with qualitative MNIST and CIFAR-10 samples in Figure 13. Again, we consistently observe that data coupling enhances sampling quality for both low and high total NFEs, but collapses when total NFE is reduced to 1, while velocity coupling produces high-quality samples even under this extreme case.



(a) Data distribution total NFEs 1 (b) Velocity distribution 1 step integration (c) Data distribution total NFEs 100 (d) Velocity distribution 100 step integration

Figure 12: Results on 2D $\mathcal{N} \to 6\mathcal{N}$ data. The three rows correspond to HRF2, HRF2 with data coupling, HRF2 with data & velocity coupling. (a) and (c) are trajectories (green) of sample particles flowing from source distribution (grey) to target distribution (blue) with total NFEs 1 and 100. (b) and (d) are velocity distributions at $(0, 1)$ at $t = 0$.



(a) HRF2-D single step (b) HRF2-D many steps (c) HRF2-D&V single step (d) Performance

Figure 13: Results on MNIST (top) and CIFAR-10 (bottom). (a) HRF2-D at a single step; (b) HRF2-D with many steps; (c) HRF2-D&V at a single step; (d) FID with respect to total NFEs. As in the main text, data coupling alone degrades at a single step, whereas HRF2-D&V stays coherent with very few steps.

G.2.1 Performance Gains across Datasets of Increasing Complexity

We evaluate performance gains across datasets of increasing complexity (MNIST → CIFAR-10 → CelebA-HQ). As shown in Table 10, HRF2-D consistently improves generation quality, with

Table 5: FID performance on CIFAR-10 under different total NFEs. **Bold** for the best per NFE.

Method	**NFE** $= 5$	**NFE** $= 10$	**NFE** $= 20$	**NFE** $= 50$	**NFE** $= 100$
RF	36.209 ± 0.142	14.113 ± 0.092	8.355 ± 0.065	5.514 ± 0.034	4.588 ± 0.013
OT-CFM	23.111 ± 0.010	12.564 ± 0.016	8.553 ± 0.002	5.911 ± 0.005	4.952 ± 0.012
HRF2	30.884 ± 0.104	12.065 ± 0.024	7.129 ± 0.027	4.847 ± 0.028	4.334 ± 0.054
HRF2-D	22.817 ± 0.072	10.969 ± 0.025	6.860 ± 0.022	**4.739 ± 0.006**	**4.301 ± 0.022**
HRF2-D&V	**6.315 ± 0.057**	**5.739 ± 0.017**	**5.332 ± 0.009**	5.142 ± 0.024	5.078 ± 0.044

Table 6: FID performance on CelebA-HQ 256 under different total NFEs. **Bold** for the best per NFE.

Method	**NFE** $= 5$	**NFE** $= 10$	**NFE** $= 20$	**NFE** $= 50$	**NFE** $= 100$
RF	38.641 ± 0.126	16.876 ± 0.088	10.027 ± 0.060	7.395 ± 0.021	6.850 ± 0.064
OT-CFM	29.646 ± 0.093	12.879 ± 0.083	7.426 ± 0.042	5.545 ± 0.023	5.236 ± 0.034
HRF2	34.246 ± 0.107	15.391 ± 0.074	9.291 ± 0.042	6.927 ± 0.041	6.450 ± 0.062
HRF2-D	32.918 ± 0.085	13.424 ± 0.022	7.048 ± 0.033	**5.529 ± 0.021**	**4.961 ± 0.019**
HRF2-D&V	**7.612 ± 0.015**	**6.931 ± 0.038**	**6.560 ± 0.039**	6.330 ± 0.022	6.248 ± 0.023

no clear diminishing trend as the evaluation resolution increases. Moreover, as shown in Table 11, HRF2-D&V significantly boosts performance at fixed low NFE, achieving up to 74.3% improvement on CelebA-HQ. These results demonstrate that our method scales well and remains effective on high-dimensional datasets.

G.2.2 Sub-Optimal Checkpoints

Since HRF2-D is used to generate the training set for HRF2-D&V, one should expect that the quality of HRF2-D impacts the performance of HRF2-D&V. Our empirical results show that HRF2-D&V is relatively robust to the specific checkpoint, as long as the result quality is reasonable. In Table 12, we report the FID of the generated CIFAR-10 images at different training stages. It shows that using a suboptimal HRF2-D checkpoint yields similar performance to using the best checkpoint. This suggests that the second-stage training is robust to such variations.

G.2.3 Velocity Coupling with Reflow

In our velocity coupling setting, if we directly use (v_0, v_1) pairs generated from HRF2-D, it will be similar to the reflow process proposed by Liu (2022). We test and show in Table 13 that for CIFAR-10, using reflow for velocity coupling is slightly worse compared to velocity coupling with batch OT. Similar trends were observed on MNIST and CelebA-HQ data.

H Implementation Details

H.1 Synthetic Data

For synthetic data experiments, we employ a neural network architecture with two distinct stages. The first stage separately encodes spatial and temporal inputs with linear layers and Sinusoidal Positional Embeddings. The second stage concatenates the processed features and refines them through multiple linear layers to produce the final output. The model consists of 304,513 parameters, totaling 0.30M in size.

For data coupling, we train the model from scratch following Algorithm 1 strictly. A key consideration is the choice of batch sizes, as two different batch sizes are involved – one for batch OT and another for training. In 1D and 2D experiments, a large batch size is necessary for stable training, but using an excessively large batch size for batch OT is computationally inefficient. To address this, we set the batch size for batch OT to 100 while using a batch size of 1,000 for gradient computation. This means that in each training iteration, we perform batch OT on 100 data points 10 times to accumulate a full batch for gradient updates.

Table 7: FID on class-conditional ImageNet 256 (VAE latent space) as a function of total NFE (best classifier-free-guidance scale per NFE). HRF2-D applies data coupling; HRF2-D&V additionally applies velocity coupling. Lower is better. External few-step methods are compared in Table 8.

Total NFE	4	8	16	32	64	128
RF	26.81	8.34	4.27	4.14	4.16	4.09
OT-CFM	8.75	4.26	3.20	3.02	3.05	3.08
HRF2	10.84	5.24	3.43	3.04	2.96	2.99
HRF2-D	9.31	3.71	**2.64**	**2.45**	**2.40**	**2.40**
HRF2-D&V	**3.57**	**2.94**	2.76	2.60	2.58	2.60

Table 8: Comparison on ImageNet 256 with external methods, reporting FID against total NFE together with model size. The upper block lists diffusion baselines with fast DPM-Solver sampling and our HRF2-D&V, and the lower block lists other low-NFE models. All external numbers are quoted from the literature and use different backbones and sampling protocols, so they are not directly comparable, and we give model size for context. Our HRF2-D&V is the smallest model. Dashes mark total NFEs that are not reported. Lower is better.

Method	Params	4	8	16	32	64	128
LDM + DPM-Solver (Ni et al., 2024)	400M	11.74	4.56	-	-	-	-
U-ViT-H + DPM-Solver (Ni et al., 2024)	501M	8.45	3.37	-	-	-	-
DiT-XL + DPM-Solver (Ni et al., 2025)	675M	9.71	5.18	3.31	-	-	-
SiT-XL (Ni et al., 2025)	675M	9.33	4.90	2.99	-	-	-
ShortCut (Frans et al., 2025)	675M	7.8	-	-	-	-	3.8
IMM (Zhou et al., 2025)	675M	2.51	1.99	1.90	1.89	-	-
iSM (Nguyen et al., 2025)	675M	2.05	1.93	-	-	-	1.88
HRF2-D&V (ours)	322M	3.57	2.94	2.76	2.60	2.58	2.60

For velocity coupling, we use the HRF2-D model from the previous step as the base model to generate (v_0, v_1) pairs at a fixed space-time location (x_t, t), following Algorithm 2. During training, we observed that the performance depends on the quality of the base model. To mitigate this, we save multiple checkpoints of HRF2-D and select the best-performing checkpoint via a validation dataset as the base model for velocity coupling.

Computational requirements during training are shown in Table 14. In the low-dimensional setting, batch OT becomes more time-consuming than the training itself. As a result, HRF2-D trains significantly slower than HRF2. In contrast, HRF2-D&V uses precomputed velocity pairs and therefore does not require batch OT during training. Moreover, it operates with a smaller batch size (1000) than HRF2 (5000), resulting in lower memory usage and faster training.

During the evaluation, we select the best checkpoint from a validation set for all models (HRF2, HRF2-D, HRF2-D&V). For each seed, we conduct the experiment three times, yielding three best models per seed. Each model is then evaluated three times, resulting in nine experimental results per seed. Finally, we report the mean and standard deviation over three different seeds, totaling 27 experimental results.

H.2 Image Data

We adapt and modify the model architectures from Zhang et al. (2025) for MNIST and CIFAR-10 data and the model architecture from Dao et al. (2023) for the CelebA-HQ data.

MNIST. For MNIST, we use the standard UNet. The ResNet blocks in the UNet function similarly to the model used for synthetic data. They process spatial and temporal inputs separately using convolutional and linear layers, respectively. The processed features are then concatenated and passed through a series of linear layers to capture space-time dependencies. The model consists of 1.07M parameters.

Table 9: Empirical conditional dispersion of the acceleration on ImageNet 256. Each entry is a model's converged acceleration-matching training loss, which by the L^2-projection identity (Lemma E.1) is an empirical upper estimate of the conditional dispersion $\mathcal{C}_2$ of Eq. (13) (the irreducible objective value). The three models share the same 322M architecture and the same loss normalization, mean-squared error over all latent dimensions. The matching target is $a = (x_1 - x_0) - v_0$ for HRF2 and HRF2-D and $a = v_1 - v_0$ under the learned velocity coupling for HRF2-D&V, so each loss estimates $\mathcal{C}_2$ under that model's coupling (γ, κ). Data coupling leaves $\mathcal{C}_2$ almost unchanged, consistent with Theorem 3.4, in which $\mathcal{C}_2$ is dominated by $\operatorname{tr} \Sigma_{\pi_0}$ when velocities are sampled independently, while velocity coupling drives it toward zero (Theorem 3.6).

Method	Coupling (γ, κ)	Converged loss $\approx \mathcal{C}_2(\gamma, \kappa)$
HRF2	(indep., indep.)	1.134
HRF2-D	(OT, indep.)	1.120
HRF2-D&V	(OT, OT)	**0.010**

Table 10: Best FID comparison of baseline and HRF2-D on different datasets at NFE$= 500$. HRF2-D achieves consistent improvements over the baseline.

Dataset	Image Size	Baseline (best)	HRF2-D (best)	HRF2-D Improv.
MNIST	1×28×28	2.296	1.847	19.6%
CIFAR-10	3×32×32	3.706	3.578	3.5%
CelebA-HQ	3×256×256	5.094	4.624	9.2%
ImageNet 256	3×256×256	2.96	2.40	18.9%

CIFAR-10. For CIFAR-10, the model consists of two UNets: a large UNet for processing v_τ and τ, and a smaller UNet (one-fourth the size) for processing x_t and t. The outputs of each ResNet block in the smaller UNet are input to the corresponding ResNet blocks in the larger UNet, facilitating information exchange between different scales. The model consists of 44.81M parameters.

CelebA-HQ. For CelebA-HQ, we first encode images into a latent space using the pretrained VAE encoder from Stable Diffusion (Rombach et al., 2022). We then use DiT (Peebles & Xie, 2023) as the backbone to process v_τ in this latent space. To condition the velocity prediction on x_t, we inject x_t into each DiT block via cross-attention layers, while keeping the main DiT architecture unchanged. The time embedding is also modified by replacing embedding(t) with embedding(t) + embedding(τ) to incorporate time information in both time axes.

ImageNet 256. For ImageNet 256, we operate in the $16 \times 16 \times 16$ latent space of the KL-16 VAE from MAR (Li et al., 2024). We decouple the two levels into a large backbone for the location (x_t, t) and a small head for the velocity (v_τ, τ). An outer transformer backbone (a bidirectional Transformer following MAR (Li et al., 2024), with 1024-dimensional embeddings, 16 layers, and 16 heads) encodes (x_t, t) into a per-token conditioning z. A separate joint inner network – a DiT (Peebles & Xie, 2023) operating over all 256 latent tokens (width 1024, depth 6) and conditioned on z via adaptive layer norm – models the velocity flow over τ. The full model has about 322M parameters and is class-conditional over the 1000 ImageNet classes, trained with a classifier-free-guidance dropout of 0.1; we use $T = 128$. For data coupling we apply whole-latent mini-batch OT between the source-noise and data latents with a batch size of 256, and train HRF2-D from scratch for 400 epochs. For velocity coupling we generate (v_0, v_1) pairs offline at each (x_t, t) by integrating the pretrained inner ODE for 32 steps and applying 2-Wasserstein OT over 256 velocity samples per location, then fine-tune only the inner network on these pairs while keeping the outer backbone frozen and reused from HRF2-D, so about 118M of the 322M parameters are updated rather than retraining the model from scratch. FID is reported against the ADM reference statistics over 50K EMA samples.

For training RF and OT-CFM on MNIST and CIFAR-10, we follow the procedures and hyperparameter settings from Tong et al. (2024) and Lipman et al. (2023). For HRF2 on the same datasets, we adopt the training setup from Zhang et al. (2025). For all models on CelebA-HQ, we follow the procedures and hyperparameters from Dao et al. (2023).

Table 11: FID comparison of baseline and HRF2-D&V at NFE=5. Our HRF2-D&V model significantly outperforms the best baseline with large relative improvements. For ImageNet 256 the closest total NFE on the T=128 sampling grid is 4, which we report here.

Dataset	Image Size	Best Baseline@NFE=5	HRF2-D&V@NFE=5	HRF2-D&V Improv.
MNIST	1×28×28	13.977	5.519	60.5%
CIFAR-10	3×32×32	23.111	6.315	72.7%
CelebA-HQ	3×256×256	29.646	7.612	74.3%
ImageNet 256	3×256×256	8.75	3.57	59.2%

Table 12: Evaluation of HRF2-D&V model using HRF2-D at different training stages (steps). Reported values are the corresponding metric scores with NFE set to 100 and 5, respectively.

Training Stage (steps)	HRF2-D NFE=100	HRF2-D&V NFE=5
Very early checkpoint (100k)	6.935	9.326
Later checkpoint (300k)	4.672	6.833
Latest checkpoint (400k)	4.301	6.315

Table 13: Performance comparison with and without OT for velocity coupling across different NFE values. Our OT-based method consistently improves results over the velocity coupling with reflow.

NFE	5	10	20	50	100	500
w/o OT (reflow)	6.632 ± 0.074	5.929 ± 0.057	5.669 ± 0.036	5.398 ± 0.007	5.273 ± 0.017	5.261 ± 0.027
w/ OT (ours)	6.315 ± 0.057	5.739 ± 0.017	5.332 ± 0.009	5.142 ± 0.024	5.078 ± 0.044	5.095 ± 0.032

For data coupling, we train the model from scratch following Algorithm 1. Both the batch OT and training batch sizes are set to 128 for MNIST and CIFAR-10 and 256 for CelebA-HQ.

Velocity-pair generation is performed offline and is independent across (x_t, t) locations, so it parallelizes across hardware and adds no sequential dependency to training.

We train the UNet for MNIST and CIFAR-10 on 1 NVIDIA L40S GPU and the DiT for CelebA-HQ on 8 NVIDIA L40S GPUs. Computational requirements, including training time and memory usage, are shown in Tables 15 to 17.

For each model, we conduct five evaluation runs, and report the means and standard deviations.

We use the `emd` function from the Python Optimal Transport (`pot`) library to compute exact OT. While the theoretical worst-case complexity is $O(n^3)$, we empirically observe much lower runtime scaling. As shown in Table 19, on CIFAR-10 data, the OT time grows sub-quadratically over batch sizes from 32 to 256, and remains negligible compared to a single training step (<2%).

I Adaptive Solvers

Our sampler consists of two nested integrations, and the inner integration follows a standard flow matching update. Since this inner step is independent of the hierarchical coupling structure, it can be replaced by any higher-order or adaptive ODE solver without modifying the formulation. This makes our framework compatible with existing adaptive solvers such as dopri5.

To illustrate this compatibility, we compare fixed step Euler sampling and adaptive dopri5 sampling on CIFAR-10 data (see Table 20). The results confirm that the hierarchical formulation does not restrict the choice of numerical solver and that adaptive solvers like dopri5 can be applied directly to the inner update. When an adaptive solver is used, HRF2-D continues to achieve the lowest FID among the compared methods. This shows that the benefits of hierarchical coupling are preserved regardless of the numerical solver.

Table 14: Computational requirements during training on synthetic datasets.

Training	1D data			2D data		
	HRF2 (0.30M)	HRF2-D (0.30M)	HRF2-D&V (0.30M)	HRF2 (0.32M)	HRF2-D (0.32M)	HRF2-D&V (0.32M)
Time (s/iter)	0.0028	0.0581	0.0025	0.0029	0.0588	0.0027
Memory (MB)	658	658	566	660	660	568
Param. Counts	304,513	304,513	304,513	321,154	321,154	321,154

Table 15: Computational cost and model size for different methods on MNIST.

MNIST	RF (1.08M)	OT-CFM (1.08M)	HRF2 (1.07M)	HRF2-D (1.07M)	HRF2-D&V (1.07M)
Time (s/iter)	0.045	0.046	0.046	0.046	0.046
Memory (MB)	2546	2546	2546	2546	2546
Param. Counts	1,075,361	1,075,361	1,065,698	1,065,698	1,065,698

Table 16: Computational cost and model size for different methods on CIFAR-10.

CIFAR-10	RF (35.75M)	OT-CFM (35.75M)	HRF2 (44.81M)	HRF2-D (44.81M)	HRF2-D&V (44.81M)
Time (s/iter)	0.166	0.169	0.196	0.202	0.200
Memory (MB)	7480	7480	9220	9220	9220
Param. Counts	35,746,307	35,746,307	44,807,843	44,807,843	44,807,843

Table 17: Computational cost and model size for different methods on CelebA-HQ.

CelebA-HQ 256	RF (457.06M)	OT-CFM (457.06M)	HRF2 (616.20M)	HRF2-D (616.20M)	HRF2-D&V (616.20M)
Time (s/iter)	0.418	0.420	0.688	0.688	0.688
Memory per GPU (GB)	26.65	26.65	39.89	39.89	36.79
Param. Counts	457,062,416	457,062,416	616,197,186	616,197,186	616,197,186

Table 18: Computational cost and model size for different methods on ImageNet 256.

ImageNet 256	RF (204.50M)	OT-CFM (204.50M)	HRF2 (322.58M)	HRF2-D (322.58M)	HRF2-D&V (322.58M)
Time (s/iter)	0.179	0.181	0.243	0.245	0.243
Memory per GPU (GB)	31.9	31.9	43.4	43.4	43.4
Param. Counts	204,503,056	204,503,056	322,582,544	322,582,544	322,582,544

Table 19: Comparison of OT computation time and training step time across different batch sizes. OT time remains negligible compared to the cost of one training step.

Batch Size	32	64	128	256
OT time (s)	0.00078	0.00117	0.00215	0.00670
One train step time (s)	0.07	0.09	0.18	0.37

Table 20: FID performance comparison between fixed step Euler sampling and adaptive dopri5 sampling on CIFAR-10 under different total NFE settings. **Bold** for the best.

Total NFEs	RF (35.75M)	OT-CFM (35.75M)	HRF2 (44.81M)	HRF2-D (44.81M)	HRF2-D&V (44.81M)
100	4.588 ± 0.013	4.952 ± 0.012	4.334 ± 0.054	**4.301 ± 0.022**	5.078 ± 0.044
500	3.887 ± 0.035	4.184 ± 0.086	3.706 ± 0.043	**3.578 ± 0.028**	5.095 ± 0.032
Adaptive	3.688 ± 0.077	3.601 ± 0.042	3.412 ± 0.058	**3.410 ± 0.027**	5.152 ± 0.009