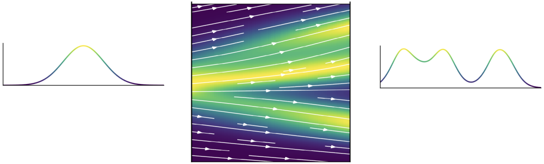
ICML
International Conference
On Machine Learning

Improving Flow Matching by Aligning Flow Divergence

Yuhao Huang, Taos Transue, Shih-Hsin Wang, William Feldman
Hong Zhang & Bao Wang



THE UNIVERSITY OF UTAH®
Argonne NATIONAL LABORATORY

Introduction to Flow Matching

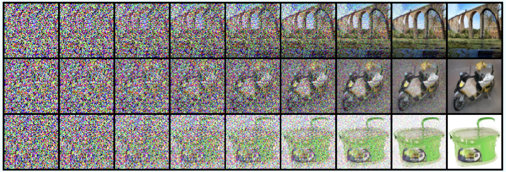
Learn a flow map $\psi_t : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$ via the following ODE

$$\frac{d}{dt}\psi_t(\boldsymbol{x}) = \boldsymbol{u}_t(\psi_t(\boldsymbol{x})) \approx \boldsymbol{v}_t(\psi_t(\boldsymbol{x}); \boldsymbol{\theta}), \tag{1}$$

with IC $\psi_0(\boldsymbol{x}) = \boldsymbol{x} \sim p_{\text{prior}}$. It maps prior distribution $p_0 = p_{\text{prior}}$ to data distribution $p_1 \approx p_{\text{data}}$ via:

$$p_t(\boldsymbol{x}) = p_0\left(\psi_t^{-1}(\boldsymbol{x})\right)\det\left[\frac{\partial\psi_t^{-1}}{\partial\boldsymbol{x}}(\boldsymbol{x})\right], \ \forall \boldsymbol{x} \in p_0, \forall t \in [0,1]. \tag{2}$$

$* \sim p_0 = p_{\text{prior}}$



$* \sim p_T \approx p_{\text{data}}$

$$\frac{d}{dt}\psi_t(\boldsymbol{x}) = \boldsymbol{v}_t(\psi_t(\boldsymbol{x}); \boldsymbol{\theta})$$

Figure: Generative Flow Map

Introduction to Flow Matching

Flow Matching for Generative Model

Flow Matching Loss:

$$\mathcal{L}_{FM}(\boldsymbol{\theta}) := \mathbb{E}_{t,p_t(\boldsymbol{x})}\left[\left\|\boldsymbol{v}_t(\boldsymbol{x},\theta) - \boldsymbol{u}_t(\boldsymbol{x})\right\|^2\right] \tag{3}$$

where $p_t(\boldsymbol{x})$ and $\boldsymbol{u}(\boldsymbol{x})$ are intractable.(Lipman et al.2023) proposed **Conditional Flow Matching** loss:

$$\mathcal{L}_{CFM}(\boldsymbol{\theta}) := \mathbb{E}_{t\sim\mathcal{U}[0,1],p_{\text{data}}(\boldsymbol{x}_1),p_t(\boldsymbol{x}|\boldsymbol{x}_1)}\left[\left\|\boldsymbol{v}_t(\boldsymbol{x},\theta) - \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1)\right\|^2\right]. \tag{4}$$

where $p_t(\boldsymbol{x}|\boldsymbol{x}_1)$ and $\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1)$ are pre-defined, $s.t.$,

$$p_t(\boldsymbol{x}) = \int p_t(\boldsymbol{x} \mid \boldsymbol{x}_1)p_{\text{data}}(\boldsymbol{x}_1)d\boldsymbol{x}_1, \quad \boldsymbol{u}_t(\boldsymbol{x}) = \int \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1)\frac{p_t(\boldsymbol{x} \mid \boldsymbol{x}_1)p_{\text{data}}(\boldsymbol{x}_1)}{p_t(\boldsymbol{x})}d\boldsymbol{x}_1 \tag{5}$$

, e.g., **Optimal transport (OT)** path:

$$p_t(\boldsymbol{x}|\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_t(\boldsymbol{x}_T), \sigma_t(\boldsymbol{x}_T)^2\boldsymbol{I}), \text{ with } \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1) = \frac{\boldsymbol{x}_1 - (1-\sigma_{\min})\boldsymbol{x}}{1-(1-\sigma_{\min})t}. \tag{6}$$

, where $\boldsymbol{\mu}_t(\boldsymbol{x}) = t\boldsymbol{x}_T$ and $\sigma_t(\boldsymbol{x}) = 1-(1-\sigma_{\min})t$.

Relation to Continuity Equation

Continuity Equation

From (Villani, 2009), we can know the vector field $\boldsymbol{u}_t$ generates a probability path p_t satisfies the **continuity equation**:

$$\begin{aligned} &\frac{\partial p_t(\boldsymbol{x})}{\partial t} + \nabla \cdot \Big(p_t(\boldsymbol{x}) \boldsymbol{u}_t(\boldsymbol{x}) \Big) = 0 \\ \Leftrightarrow &\frac{\partial p_t(\boldsymbol{x})}{\partial t} = -\Big(\nabla \cdot \boldsymbol{u}_t(\boldsymbol{x}) \Big) p_t(\boldsymbol{x}) - \boldsymbol{u}_t(\boldsymbol{x}) \cdot \nabla p_t(\boldsymbol{x}) \end{aligned} \tag{7}$$

with $p_0(\boldsymbol{x}) = p_{\text{prior}}(\boldsymbol{x})$ and $p_1(\boldsymbol{x}) = p_{\text{data}}(\boldsymbol{x})$.

Similarly, the learned $\boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta})$ with estimated path $\hat{p}_t(\boldsymbol{x})$ and $p_0(\boldsymbol{x}) = p_{\text{prior}}(\boldsymbol{x})$ satisfy

$$\frac{\partial \hat{p}_t(\boldsymbol{x})}{\partial t} = -\Big(\nabla \cdot \boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta}) \Big) \hat{p}_t(\boldsymbol{x}) - \boldsymbol{v}_t(\boldsymbol{x}; \boldsymbol{\theta}) \cdot \nabla \hat{p}_t(\boldsymbol{x}) \tag{8}$$

Notice: $|\boldsymbol{u} - \boldsymbol{v}|$ is controlled by flow matching loss but $|\nabla \cdot \boldsymbol{u} - \nabla \cdot \boldsymbol{v}|$ not!

Error Defined by Continuity Equation

Error of Approximated Probability Path

Let $\epsilon_t := p_t - \hat{p}_t$ be the error, satisfying the following transport equation

$$\begin{cases} \partial_t \epsilon_t + \nabla \cdot \Big(\epsilon_t \boldsymbol{v}_t \Big) = L_t, \\ \epsilon_0(x) = 0, \end{cases} \tag{9}$$

where $L_t = -p_t \big[\nabla \cdot (\boldsymbol{u}_t - \boldsymbol{v}_t) + (\boldsymbol{u}_t - \boldsymbol{v}_t) \cdot \nabla \log p_t \big]$.

Duhamel's formula:

$$\epsilon_t \big(\phi_t(\boldsymbol{x}) \big) \cdot \det \nabla \phi_t(\boldsymbol{x}) = - \int_0^t p_s \Big[\big(\nabla \cdot (\boldsymbol{u}_s - \boldsymbol{v}_s) \big) + (\boldsymbol{u}_s - \boldsymbol{v}_s) \cdot \nabla \log p_s \Big] \cdot \det \nabla \phi_s(\boldsymbol{x}) ds \tag{10}$$

where $\phi(\boldsymbol{x})$ is the flow induced by $\boldsymbol{v}_t$ in $\frac{d}{dt}\phi_t(\boldsymbol{x}) = \boldsymbol{v}_t(\phi_t(\boldsymbol{x}); \boldsymbol{\theta})$, and $\det \nabla \phi(\boldsymbol{x})$ denotes the determinant of the Jacobian matrix.

Error Defined by Continuity Equation

TV Error from Continuity Equation

Theorem Under mild assumptions, there exists a constant $C > 0$ such that

$$\mathrm{TV}(p_t, \hat{p}_t) \leq \frac{1}{2}\mathbb{E}_{t,p_t(\boldsymbol{x})}\left[\left|\nabla \cdot \boldsymbol{u}_t(\cdot) - \nabla \cdot \boldsymbol{v}_t(\cdot;\boldsymbol{\theta})\Big)\right|\right] + \frac{C}{2}\mathbb{E}_{t,p_t(\boldsymbol{x})}\left[\left|\boldsymbol{u}_t(\cdot) - \boldsymbol{v}_t(\cdot;\boldsymbol{\theta})\right|\right] \quad (11)$$

Note: Second term is already flow matching loss.

Divergence Loss

Flow Divergence Matching Loss

The TV error bound gives the following divergence loss:

$$\mathcal{L}_{\mathrm{DM}}(\boldsymbol{\theta}) := \mathbb{E}_{t,p_t(\boldsymbol{x})}\left[\left|\nabla \cdot (\boldsymbol{u}_t - \boldsymbol{v}_t) + (\boldsymbol{u}_t - \boldsymbol{v}_t) \cdot \nabla \log p_t\right|\right] \tag{12}$$

which is also intractable.

Conditional Flow Divergence Loss

$$\begin{aligned}\mathcal{L}_{\mathrm{CDM}}(\boldsymbol{\theta}) := \mathbb{E}_{t,p_t(\boldsymbol{x}|\boldsymbol{x}_1),p(\boldsymbol{x}_1)}\Bigg[\Bigg|&\Big(\nabla \cdot \boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1) - \nabla \cdot \boldsymbol{v}_t(\boldsymbol{x},\theta)\Big) \\ &+\Big(\boldsymbol{u}_t(\boldsymbol{x}|\boldsymbol{x}_1) - \boldsymbol{v}_t(\boldsymbol{x},\theta)\Big) \cdot \nabla \log p_t(\boldsymbol{x}|\boldsymbol{x}_1)\Bigg|\Bigg].\end{aligned} \tag{13}$$

which is an upper bound of $\mathcal{L}_{\mathrm{DM}}(\boldsymbol{\theta})$.

Flow Matching with Divergence Loss

Improve Flow Matching with Aligning Flow Divergence

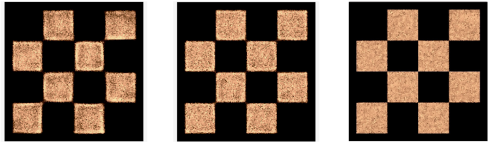
We propose the flow and divergence matching (FDM) loss:

$$\mathcal{L}_{\mathrm{FDM}} = \lambda_1 \mathcal{L}_{\mathrm{CFM}} + \lambda_2 \mathcal{L}_{\mathrm{CDM}}, \tag{14}$$

where $\lambda_1, \lambda_2 > 0$ are hyperparameters.

Density Modeling

Synthetic Data – Checkerboard



(a) FM (OT) (b) FDM (OT) (c) Ground Truth

Model	FM (OT)	FDM (OT)	FM (VP)	FDM (VP)
Likelihood (↑)	$2.38_{\pm .02}$	$\mathbf{2.53}_{\pm .02}$	$2.34_{\pm .02}$	$2.46_{\pm .02}$

Table 1. Likelihood estimation of models on the checkerboard test set. Here, "OT" denotes the optimal transport path and "VP" denotes the variance-preserving path. Unit: $\times 10^{-2}$

Image Data – CIFAR10

Model	NLL(↓)	FID(↓)
FM(OT)	2.99	6.35
FDM(OT)	2.85	5.62

Table 2. Negative log-likelihood and sample quality (FID scores) estimation on CIFAR-10.

Sequential Data Sampling with Guidance – DNA Sequence

Method	MSE (↓)
Bit Diffusion (One-hot Encoding)(Albergo et al., 2023)	3.95E-2
DDSM (Albergo et al., 2023)	3.34E-2
Large Language Model (Stark et al., 2024)	3.33E-2
Linear FM (Stark et al., 2024)	2.82$_{\pm 0.02}$E-2
Linear FDM (ours)	2.78$_{\pm 0.01}$E-2
Dirichlet FM (Stark et al., 2024)	2.68$_{\pm 0.01}$E-2
Dirichlet FDM (ours)	**2.59**$_{\pm 0.02}$E-2

Table 4. Evaluation of transcription profile guided promoter DNA sequence design of different models.

- Train the models guided by a profile by providing it as additional input to the vector field;
- Evaluate generated sequences using mean- squared error (MSE) between their predicted and original regulatory activity.

Spatiotemporal Data – Dynamical System

Model	Lorenz		FitzHugh-Nagumo	
	$p(\boldsymbol{x}_1)$ (↓)	$p(\boldsymbol{x}_1\|E)$ (↓)	$p(\boldsymbol{x}_1)$ (↓)	$p(\boldsymbol{x}_1\|E)$ (↓)
Diffusion	0.0314	0.1001	0.0277	0.1192
FM	0.0348	0.0972	0.0314	0.2164
FDM	**0.0306**	**0.0914**	**0.0266**	**0.1168**

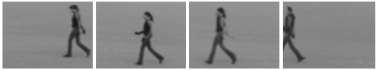
Table 5. TV distances of the models from the trajectory distribution $p(\boldsymbol{x}_1)$ and from the distribution conditioned on an event $p(\boldsymbol{x}_1|E)$. Here, Diffusion results follow from (Finzi et al., 2023), while FM and FDM are based on our implementation, which builds on the code provided by Finzi et al. (2023).

Model	Lorenz		FitzHugh-Nagumo	
	$p(\boldsymbol{x}_1)$	$p(\boldsymbol{x}_1\|E)$	$p(\boldsymbol{x}_1)$	$p(\boldsymbol{x}_1\|E)$
Diffusion	0.0056	0.2774	0.0260	0.3011
FM	0.0081	**0.2560**	0.0280	0.3468
FDM	**0.0049**	0.3045	**0.0280**	**0.2084**

Table 7. KL divergence between the histograms of the event constraint value $C(\boldsymbol{x}_1)$ for event trajectories $\boldsymbol{x}_1$ in the dataset of trajectories computed by an ODE solver and event trajectories sampled with event guidance from the models.

Sample trajectories in dynamical systems using the initial states from the first few steps as additional inputs to the model, with or without event guidance.

Spatiotemporal Data – Video Prediction

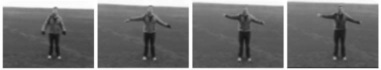


Walking; a-FM

Walking; a-FDM

Boxing; b-FM

Boxing; b-FDM



Hand Waving; c-FM

Hand Waving; c-FDM

Method	FVD(↓)	PSNR(↑)	Time(s/iter)
SRVP (Franceschi et al., 2020)	222	29.7	–
SLAMP (Akan et al., 2021)	228	29.4	–
Latent FM (Davtyan et al., 2023)	180	30.4	0.18
Latent FDM (ours)	**155.5**$_{\pm 5}$	**31.2**	0.27

Table 8. KTH dataset evaluation. The evaluation protocol is to predict the next 30 frames given the first 10 frames.

Method	FVD(↓)	MEM(GB)	Time(hours)
TriVD-GAN-FP (Luc et al., 2020)	103	1024	280
Video Transformer (Weissenborn et al., 2019)	94	512	336
LVT (Rakhimov et al., 2020)	126	128	48
RaMViD (Diffusion) (Höppe et al., 2022)	84	320	72
Latent FM (Davtyan et al., 2023)	146	24.2	25
Latent FDM (ours)	**123**$_{\pm 4.5}$	35	36

Table 9. BAIR dataset evaluation. We adopt the standard evaluation setup, where the model predicts 15 future frames conditioned on a single initial frame. MEM stands for peak memory footprint.

Autoregressive next-frame generation (prediction) guided by several preceding frames, provided as additional inputs to the flow matching vector regressor.

Thank you!