MoFlow: One-Step Flow Matching for Human Trajectory Forecasting via Implicit Maximum Likelihood Estimation based Distillation

Yuxiang Fu[1,2] Qi Yan[1,2] Lele Wang[1] Ke Li[4] Renjie Liao[1,2,3]

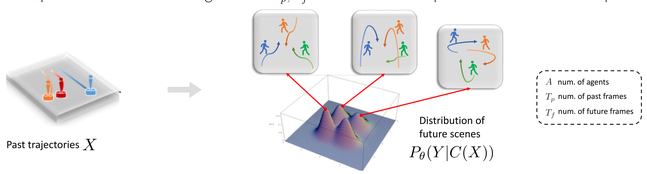
[1]University of British Columbia [2]Vector Institute for AI [3]Canada CIFAR AI Chair [4]Simon Fraser University

Webpage

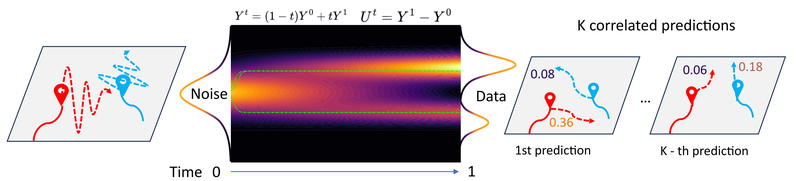
CVPR Nashville JUNE 11-15, 2025

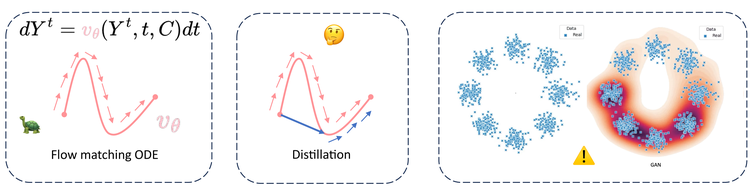
Background

Problem formulation: Given the past trajectory history $X = [\boldsymbol{x}_1^\top; \boldsymbol{x}_2^\top; \ldots; \boldsymbol{x}_A^\top] \in \mathbb{R}^{A \times 2T_p}$ of the agents, we want to model the distribution of the future scenes $Y \sim P_\theta, Y \in \mathbb{R}^{A \times 2T_f}$ jointly so that the generated samples are socially and physically compliant. Note that A represents the number of agents and T_p, T_f are the number of past and future frames respectively.



Limitations of Flow Models and distillation techniques





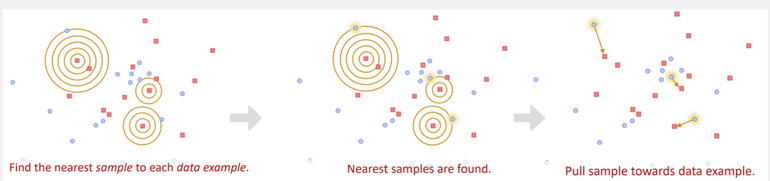
- The numerical ODE solver requires numerous iterative network evaluations, making the inference **slow** and **expensive**.
- Most distillation methods fail to preserve good quality of samples from teacher model.
- Consistency Distillation requires samples Y^{t_n} from the teacher model at different noise levels for all timestamps.

Implicit Maximum Likelihood Estimator

The Implicit Maximum Likelihood Estimator (IMLE) is defined as

$$\hat{\theta}_{\text{IMLE}} := \arg\min_\theta \mathbb{E}_{\tilde{\boldsymbol{x}}_1^\theta, \ldots, \tilde{\boldsymbol{x}}_m^\theta} \left[\sum_{i=1} \min_{j \in [m]} \|\tilde{\boldsymbol{x}}_j^\theta - \boldsymbol{x}_i\|_2^2 \right]$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are our data examples and $\tilde{\boldsymbol{x}}_1^\theta, \ldots, \tilde{\boldsymbol{x}}_m^\theta$ denotes the i.i.d. samples from P_θ.



Contributions

1. We present a novel **Mo**tion prediction **Flow** matching model for trajectory prediction tasks. Also, we design a novel flow matching loss that encourages learning a diverse set of future trajectories that well capture the multi-modality of human trajectories.
2. We propose a **one-step** distillation method for flow models based on IMLE. Our approach accelerates the standard conditional flow matching inference by **100x** without sacrificing quality of generated trajectories.
3. Both MoFlow and the proposed distillation method deliver SOTA performance over three human motion datasets, producing diverse trajectories that are physically and socially plausible.

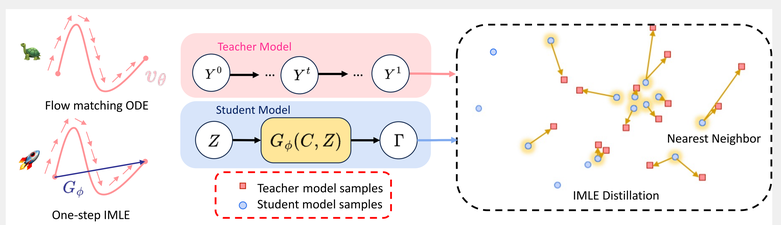
MoFlow Objective

Our model outputs K scene-level waypoint predictions, denoted by $\{S_i\}_{i=1}^K, S_i \in \mathbb{R}^{A \times 2T_f}$, alongside corresponding **classification logits** $\{\zeta_i\}_{i=1}^K, \zeta_i \in \mathbb{R}$. Building upon original Flow Matching loss, we derive

$$\tilde{\mathcal{L}}_{\text{FM}} = \mathbb{E}_{Y^0, Y^1, t} \left[\|S_{j^*} - Y^1\|_2^2 + \text{CE}(\zeta_{1:K}, j^*) \right], \quad j^* = \arg\min_j \|S_j - Y^1\|_2^2,$$

where $\text{CE}(\cdot, \cdot)$ denotes the cross-entropy loss.

IMLE Distillation



We adapt IMLE to a new objective tailored for our conditional generation task, expressed as follows

$$\hat{\phi}_{\text{IMLE}} := \arg\min_\phi \mathbb{E}_{Z_1, \ldots, Z_m} \left[\min_{j \in [m]} \mathcal{L}_{\text{IMLE}}(\hat{Y}_{1:K}^1, G_\phi(C, Z_j)) \right], \mathcal{L}_{\text{IMLE}}(\hat{Y}_{1:K}^1, \Gamma) = \frac{1}{K} \left(\sum_{i=1}^K \min_j \|\hat{Y}_i^1 - \Gamma^{(j)}\| + \sum_{j=1}^K \min_i \|\hat{Y}_i^1 - \Gamma^{(j)}\| \right),$$

where our conditional IMLE generator G_ϕ uses a noise vector among $Z_1, \ldots, Z_m \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ and context C to generate K-component trajectories $\Gamma \in \mathbb{R}^{K \times A \times 2T_f}$, matching the shape of $\hat{Y}_{1:K}^1$ from our teacher model.

Quantitative Results on NBA/ETH-UCY/SDD Datasets

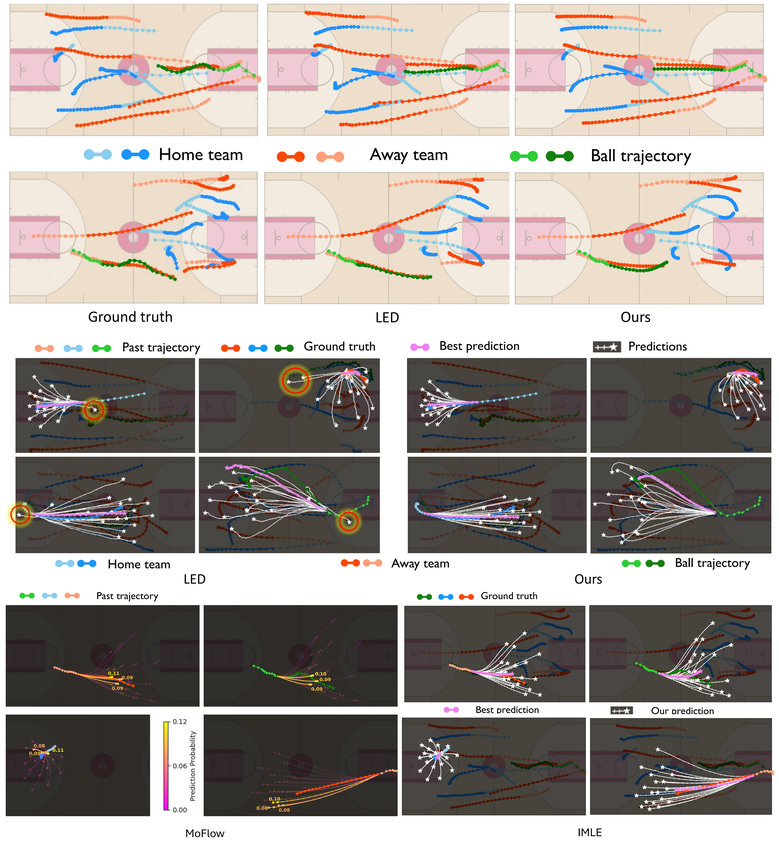
Time	MemoNet [54]	NPSN [4]	GroupNet [53]	MID [16]	LED† [33]	LED* [33]	MoFlow	IMLE
1.0s	0.38/0.56	0.35/0.58	0.26/0.34	0.28/0.37	0.21/0.28	**0.18**/0.27	**0.18/0.25**	**0.18/0.25**
2.0s	0.71/1.14	0.68/1.23	0.49/0.70	0.51/0.72	0.44/0.64	0.37/0.56	**0.34/0.47**	0.35/**0.47**
3.0s	1.00/1.57	1.01/1.76	0.73/1.02	0.71/0.98	0.69/0.95	0.58/0.84	**0.52/0.67**	**0.52/0.67**
Total (4.0s)	1.25/1.47	1.31/1.79	0.96/1.30	0.96/1.27	0.94/1.21	0.81/1.16	**0.71/0.87**	**0.71/0.87**

Subsets	MID [16]	GroupNet [53]	TUTR [42]	EqMotion [55]	EigenTraj [5]	LED* [33]	SingularTraj [6]	MoFlow	IMLE
ETH	0.39/0.66	0.46/0.73	0.40/0.61	0.40/0.61	0.36/0.53	0.39/0.58	**0.35/0.42**	0.40/0.57	0.40/0.58
HOTEL	0.13/0.22	0.15/0.25	**0.11**/0.18	0.12/0.18	0.12/0.19	**0.11/0.17**	0.13/0.19	**0.11/0.17**	0.12/0.18
UNIV	**0.22**/0.45	0.26/0.49	0.23/0.42	0.23/0.43	0.24/0.34	0.26/0.44	0.25/0.44	0.23/**0.39**	0.23/**0.39**
ZARA1	0.17/0.30	0.21/0.39	0.18/0.34	0.18/0.32	0.19/0.33	0.18/**0.26**	0.19/0.32	**0.15/0.26**	0.16/**0.26**
ZARA2	0.13/0.27	0.17/0.33	0.13/0.25	0.13/0.23	0.14/0.24	0.13/**0.22**	0.15/0.25	**0.12/0.22**	0.13/**0.22**
AVG	0.21/0.38	0.25/0.44	0.21/0.36	0.32/0.35	0.21/0.34	0.21/0.33	0.21/**0.32**	**0.20/0.32**	0.21/0.33

Method	ADE/FDE
Evo-Graph [23]	13.90/22.90
Y-net [32]	11.49/20.23
GroupNet [53]	9.31/16.11
CAGN [10]	9.42/15.93
NPSN [4]	8.56/11.85
MemoNet [54]	9.50/14.78
SocialVAE [57]	8.88/14.81
MID [16]	9.73/15.32
TUTR [42]	7.76/12.69
LED [33]	8.48/**11.66**
MoFlow	**7.50**/11.96
IMLE	7.85/12.86

Qualitative Assessments

NBA dataset:



ETH-UCY dataset: