Plausible and Feasible Long-Term Human Trajectory Prediction via Motion Field-Regularized Flow Matching

Yufei Zhu[1*], Jinhao Liang[2*], Sven Koenig[3], Ferdinando Fioretto[2], Martin Magnusson[1]

***Abstract*— Predicting multiple plausible future trajectories of people is essential for applications in robotics and intelligent environments. While recent generative approaches based on diffusion models and flow matching have advanced multi-candidate trajectory prediction, they typically pursue diversity among the various output trajectories generated without ensuring that each output is individually valid: while a candidate may be spatially diverse, it may also be physically infeasible or inconsistent with the environment's motion patterns. To address this limitation, we propose a framework that integrates a neural implicit motion field into conditional flow matching for long-term human trajectory prediction. The proposed method generates K candidate trajectories using a conditional flow matching backbone with a best-of-K training objective, and regularizes all candidates via a motion-field consistency loss that steers predictions towards plausible motion modes. To enforce physical consistency, this work introduces an SDF-based geometric constraint to discourage collisions with static obstacles. Crucially, this constraint is applied at inference-time within the learned vector velocity field, without requiring retraining and ensuring generalizability to unseen environments. Experiments on a real-world dataset demonstrate that our approach improves both prediction accuracy and physical feasibility of the predicted trajectories over extended horizons.**

I. Introduction

Accurate long-term prediction of future human trajectories is a core capability for autonomous systems, with applications in motion planning, human-robot interaction, automated driving, and intelligent monitoring [1]. Given an observed trajectory, the task is to forecast a set of plausible future paths that capture the inherent uncertainty in human motion. This is fundamentally challenging because human motion is complex, multimodal, and shaped by factors that are both endogenous and exogenous: it is driven by individual intent and dynamics as well as external influences such as social conventions and environmental structure [2, 3].

The challenge intensifies at longer prediction horizons. Over extended time scales (e.g., 20 seconds or more), the influence of the surrounding environment becomes the dominant factor guiding motion [3]. Short-term kinematics and immediate social interactions alone are insufficient to capture how corridors, intersections, and exits shape pedestrian flow [4, 5, 6]. Long-term predictions therefore demand models that incorporate environment-level structure into the generative process.

*Equal contribution
[1]Örebro University, Sweden yufei.zhu@oru.se
[2]University of Virginia, USA
[3]University of California, Irvine, USA

A variety of learning-based approaches have been proposed to tackle trajectory prediction. Early methods relied on recurrent sequence models [2], generative adversarial training [7], and variational autoencoders [8]. Transformer-based architectures [9, 10] subsequently improved the modeling of long-range temporal dependencies. More recently, denoising diffusion models [11, 12] and flow matching [13, 14] have emerged as powerful generative backbones for multi-candidate trajectory prediction. Flow matching learns a vector field that defines a continuous transformation from noise to data via an ordinary differential equation (ODE), offering efficient sampling with fewer integration steps [15, 16]. However, these generative approaches typically encourage diversity among output candidates, for instance through Minimum-of-K training objectives, without ensuring that each individual prediction is *plausible* with respect to the environment's motion patterns or *physically feasible* with respect to obstacle geometry. A candidate trajectory may be spatially diverse yet follow an unrealistic motion mode or pass through walls and furniture.

A separate line of research has shown that Maps of Dynamics (MoDs) [17] can capture valuable prior knowledge about how people typically move through a space. MoDs encode spatio-temporal motion patterns as a property of the environment, implicitly accounting for the topological structure. MoD-based trajectory predictors [18] leverage this structure and have demonstrated strong performance at long prediction horizons. Recent work on continuous neural representations [19] has made MoDs differentiable and queryable at arbitrary locations, opening the door to integration with gradient-based training pipelines. However, MoD-based predictors have not been combined with modern generative frameworks, leaving the complementary strengths of these two directions unexploited.

In this work, we bridge this gap by integrating a differentiable neural motion field into conditional flow matching for long-term human trajectory prediction. Our framework addresses two key properties that existing generative approaches lack. First, to promote *plausibility*, we introduce a Map-of-Dynamics training loss that regularizes all K trajectory candidates to align with the environment's observed motion patterns. Since the neural motion field is inherently multimodal, this loss steers different candidates toward different motion modes, encouraging mode-consistent diversity rather than random perturbations. Second, to enforce *physical feasibility*, we introduce an SDF-based geometric guidance applied during ODE integration at inference time. This training-free mechanism steers trajectories away from static

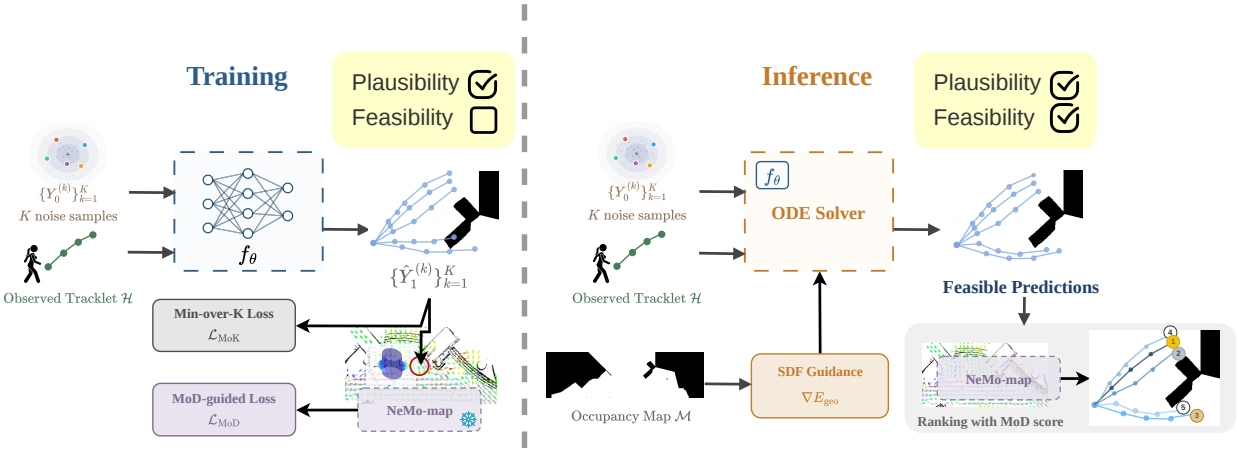

Fig. 1: Overview of the proposed framework. During training, the generative model is regularized by a Map-of-Dynamics Loss to encourage **plausible**, mode-consistent predictions. During sampling, an SDF-based geometric guidance is introduced to steer trajectories away from obstacles, encouraging physical **feasibility**. The resulting candidates are **ranked** by MoD-consistency score.

obstacles using the SDF gradient, requiring no retraining and enabling generalization to unseen obstacle configurations.

We evaluate on the ATC dataset [20] over prediction horizons of 20–60 s and demonstrate improvements in prediction accuracy and physical feasibility over strong baselines including Transformer-based, flow matching, and MoD-based methods. The overall framework is illustrated in Fig. 1.

II. Method

We address long-term human trajectory prediction: given an observed tracklet $\mathcal{H} = \langle s_{t_0-O_p}, \dots, s_{t_0} \rangle$ of 2D positions and an occupancy grid map $\mathcal{M}$, we generate K future trajectory predictions $\{\hat{Y}^{(k)}\}_{k=1}^{K}$, each spanning T_p time steps. Our framework comprises three components: (i) a conditional flow matching (CFM) backbone (Sec. II-A), (ii) a Map-of-Dynamics (MoD) training loss for mode-consistent diversity (Sec. II-B), and (iii) geometric feasibility guidance at inference (Sec. II-C).

A. Conditional Flow Matching Backbone

We employ CFM [13, 14] to learn a vector field $v_\theta(\cdot, \tau)$ that transports samples from a Gaussian prior p_0 to the trajectory distribution p_1, conditioned on $\mathcal{H}$, where $\tau \in [0, 1]$ denotes the flow time. Following [13], the conditional path between noise $Y_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and ground truth Y_1 is $Y_\tau = (1-\tau)Y_0 + \tau Y_1$, with conditional vector field $u_\tau = Y_1 - Y_0$.

a) Multi-candidate prediction: We generate K candidates jointly using a Minimum-over-K (MoK) objective [15, 21]. Following [15], we reparameterize in data-prediction space: the network predicts $\hat{Y}_1^{(k)} = f_\theta(Y_\tau^{(k)}, \tau \mid \mathcal{H})$, and the MoK loss selects the closest prediction:

$$\mathcal{L}_{\text{MoK}}(\theta) = \mathbb{E}\left[\min_{k=1,\dots,K} \frac{\|\hat{Y}_1^{(k)} - Y_1\|^2}{(1-\tau)^2}\right]. \tag{1}$$

Gradients flow only through the best-matching candidate; the MoD loss (Sec. II-B) regularizes all K outputs.

b) Inference.: Trajectories are obtained by integrating the flow ODE from τ=0 to τ=1 via N Euler steps. At each step, the velocity is recovered as $v_\theta = (\hat{Y}_1^{(k)} - Y_\tau^{(k)})/(1-\tau)$ and the state is updated as $Y_{\tau+\Delta\tau}^{(k)} = Y_\tau^{(k)} + \Delta\tau \cdot v_\theta$.

B. Map-of-Dynamics Loss

To encourage each candidate to follow a semantically distinct motion mode, we incorporate a pre-trained neural motion field as a structured prior. We use NeMo-map [19], which parameterizes the environment's motion patterns as a continuous implicit function Φ_ψ. At any position $\mathbf{p}$, it outputs a Semi-Wrapped Gaussian Mixture Model (SWGMM) over velocities:

$$p\big(\mathbf{v} \mid \Phi_\psi(\mathbf{p})\big) = \textstyle\sum_{j=1}^{J} w_j \, \mathcal{N}^{\text{SW}}_{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}(\mathbf{v}). \tag{2}$$

For each predicted trajectory, we compute velocities $\hat{v}_t^{(k)} = \hat{s}_{t+1}^{(k)} - \hat{s}_t^{(k)}$ in polar form and minimize their negative log-likelihood under the location-conditioned distribution:

$$\mathcal{L}_{\text{MoD}} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{T_p}\sum_{t}\big[-\log p(\hat{v}_t^{(k)} \mid \Phi_\psi(\hat{s}_t^{(k)}))\big]. \tag{3}$$

Since the motion field is multimodal, this loss steers different candidates toward different mixture components, promoting mode-consistent diversity. The full training loss is $\mathcal{L} = \mathcal{L}_{\text{MoK}} + \lambda_{\text{MoD}}\,\mathcal{L}_{\text{MoD}}$.

C. Geometric Feasibility Guidance

While the CFM with MoD loss generates diverse and mode-consistent trajectories, there is no explicit guarantee that the predicted paths are strictly feasible in the physical environment, potentially resulting in paths that intersect with the environmental geometry (e.g. walls, furniture). We address this by incorporating geometric guidance into the sampling process: at each ODE integration step, an SDF-based gradient signal steers the evolving trajectories away from obstacles.

1) Signed Distance Function: Given the occupancy grid map $\mathcal{M}$ of the environment, we precompute a signed distance function (SDF) $d : \mathbb{R}^2 \to \mathbb{R}$ using the Euclidean distance transform, such that:

$$d(\mathbf{p}) = \begin{cases} +\min_{\mathbf{q}\in\partial\mathcal{O}} \|\mathbf{p}-\mathbf{q}\|, & \text{if } \mathbf{p} \in \mathcal{F}, \\ 0, & \text{if } \mathbf{p} \in \partial\mathcal{O}, \\ -\min_{\mathbf{q}\in\partial\mathcal{O}} \|\mathbf{p}-\mathbf{q}\|, & \text{if } \mathbf{p} \in \mathcal{O}, \end{cases} \tag{4}$$

where $\mathcal{F}$, $\mathcal{O}$, and $\partial\mathcal{O}$ denote the free space, obstacle interior, and obstacle boundary, respectively. The SDF is computed once and stored as a grid that supports efficient bilinear interpolation for arbitrary query points.

2) SDF-Guided Sampling: To ensure geometric feasibility, our goal is to sample from a target posterior distribution conditioned on both the observed history $\mathcal{H}$ and a geometric feasibility constraint d. The constrained posterior can be formulated as $p(Y \mid \mathcal{H}, d) \propto p_\theta(Y \mid \mathcal{H}) p(d \mid Y)$, where $p_\theta(Y \mid \mathcal{H})$ is the unconstrained learned distribution [22].

We define the feasibility likelihood as $p(d \mid Y) \propto \exp\big(-E_{\text{geo}}(Y)\big)$. Given a safety margin $\epsilon \geq 0$, the geometric energy over a trajectory $Y = \langle s_{t_0+1}, \ldots, s_{t_0+T_p} \rangle$ is formulated as:

$$E_{\text{geo}}(Y) = \sum_{t=t_0+1}^{t_0+T_p} \max\big(0,\ \epsilon - d(s_t)\big)^2. \tag{5}$$

This energy landscape is zero when all waypoints are safely in free space ($d(s_t) \geq \epsilon$), and increases when colliding with an obstacle.

To sample from the constrained posterior, we incorporate the gradient of the log-likelihood into the sampling process directly. Specifically, we modify the standard Euler integration step by incorporating this guidance signal, yielding the constrained training-free inference update rule:

$$Y^{(k)}_{\tau+\Delta\tau} = Y^{(k)}_{\tau} + \Delta\tau \cdot \Big[v_\theta\big(Y^{(k)}_{\tau}, \tau \mid \mathcal{H}\big) - \lambda_{\text{geo}} \nabla_Y E_{\text{geo}}\big(Y^{(k)}_{\tau}\big) \Big], \tag{6}$$

where λ_{geo} controls the guidance strength.

The gradient $\nabla_Y E_{\text{geo}}$ acts on each waypoint independently: for a point s_t with $d(s_t) < \epsilon$, it pushes s_t along $\nabla d(s_t)$ (towards free space); for points already in free space, the gradient is zero and the update reduces to standard Euler integration. This mechanism ensures that the motion dynamics learned by the generative model are preserved, while adhering to environmental constraints without the need for model retraining.

III. Experiments

A. Setup

We evaluate on the ATC dataset [20], which contains pedestrian trajectories recorded in a large shopping mall in Japan covering approximately $900\,\text{m}^2$ over 92 days. The data is downsampled to 1 Hz with an observation horizon of 3 s and prediction horizons from 20 s to 60 s, randomly split into 80%/10%/10% for training/validation/testing. We compare against TUTR [10] (Transformer-based), MoFlow [15] (flow matching), and two MoD-based methods: TC-CLiFF-LHMP [18] and STeF-LHMP [23]. All methods generate K=5 candidates.

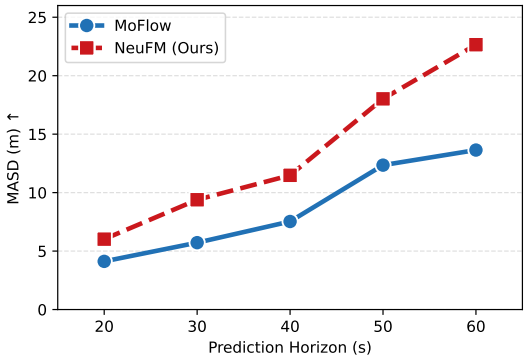

Fig. 2: Map-aware average self-distance (MASD) over prediction horizons.

B. Prediction Accuracy

minADE$_K$ and **minFDE**$_K$ measure the Average Displacement Error and Final Displacement Error of the closest prediction (among K candidates) to the ground truth. Our method achieves the best or second-best performance across all metrics and prediction horizons. At longer horizons, our method consistently outperforms all baselines: at $T_s = 40\,\text{s}$, it reduces minADE$_5$ by 13.6% relative to MoFlow and minFDE$_5$ by 32.5%. This demonstrates the advantage of incorporating environment-level motion priors into the generative process. The learned motion patterns implicitly encode information about potential goal locations in the environment, providing guidance that becomes increasingly impactful at longer prediction horizons, where short-term kinematic cues alone are insufficient to capture the complexity of human motion.

C. Diversity

Beyond prediction accuracy, a key motivation is to generate trajectory candidates that provide meaningful coverage and diversity across plausible motion modes. To quantify diversity, we adopt the map-aware average self-distance (MASD) metric [24], which is used by MoFlow [15] for evaluating diversity. MASD computes the L2 distance between the two most distant trajectory samples among the K candidates, and then averages this maximum distance over the prediction horizon, with higher values indicating greater spread across the prediction set.

Fig. 2 reports MASD across prediction horizons. Our method consistently achieves higher MASD than MoFlow, with the gap growing from $+46\%$ at 20 s to $+66\%$ at 60 s. This indicates that the MoD-regularized training objective encourages the model to spread its K candidates across distinct motion modes. Combined with the improved MoD-consistency, this result supports that the achieved diversity is grounded in the environment's motion patterns.

		TUTR [ICCV'23]	MoFlow [CVPR'25]	TC-CLiFF-LHMP [RA-L'25]	STeF-LHMP [RA-L'25]	**NeuFM (Ours)**
20 s	minADE$_5$↓	1.86	**1.50**	2.63	2.68	1.69
	minFDE$_5$↓	3.59	2.99	4.75	5.14	**2.94**
40 s	minADE$_5$↓	4.62	3.17	5.12	5.74	**2.74**
	minFDE$_5$↓	9.43	7.08	10.66	12.47	**4.78**
60 s	minADE$_5$↓	8.86	4.65	8.98	10.45	**4.56**
	minFDE$_5$↓	18.97	11.06	20.47	24.62	**7.19**

TABLE I: Quantitative results on the ATC dataset across prediction horizons $T_s \in \{20, 40, 60\}$ s with $K = 5$ trajectory candidates. Observation horizon $O_s = 3$ s. All displacement errors are in meters. Best results are in **bold**, second best are underlined. ↓ indicates lower is better.

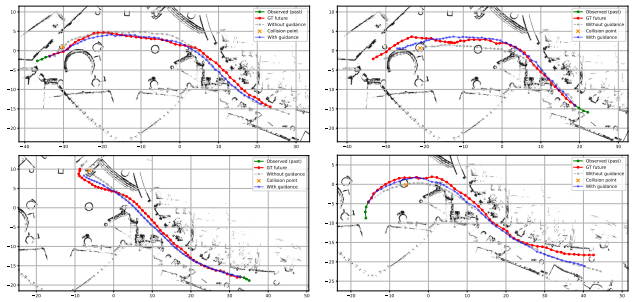

Fig. 3: Qualitative effect of SDF-based geometric guidance on predicted trajectories. Each figure shows the observed past (**green**), ground-truth future (**red**), and the best-of-K prediction without guidance (**grey dashed**) and with guidance (**blue**). **Orange crosses** mark waypoints where the unguided trajectory collides with obstacles (SDF < 0). The geometric guidance steers the predicted trajectory away from walls and pillars during ODE integration, encouraging collision-free paths without retraining.

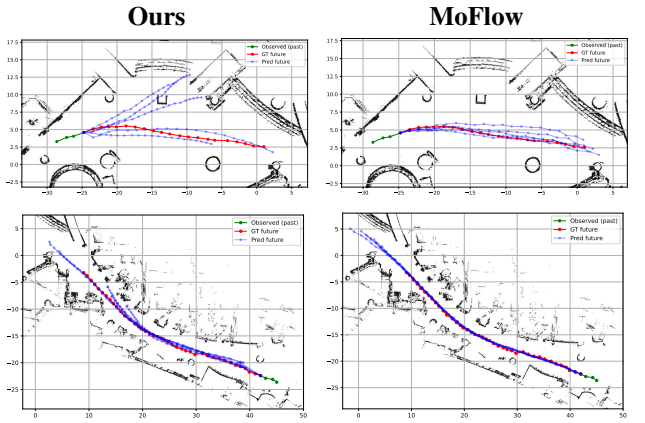

Fig. 4: Qualitative comparison between our method and MoFlow. Our method generates trajectories that cover distinct plausible motion modes, whereas MoFlow concentrate predictions in a single direction.

D. Qualitative Results

a) Geometric feasibility.: Fig. 3 illustrates the effect of SDF-based guidance. Without guidance, predicted trajectories may pass through walls and pillars (grey dashed, with orange collision markers). With guidance enabled (blue), trajectories are steered into free space during ODE integration without retraining, enabling adaptation to unseen obstacle configurations.

b) Multimodal coverage.: Fig. 4 presents a side-by-side comparison of trajectory predictions generated by our method and MoFlow. While MoFlow tends to concentrate its K predictions along a single direction, Our method produces candidates that spread across semantically distinct motion modes. In the first row ($T_s = 20$ s), our method generates trajectories heading towards different exits such as stairs and escalators, whereas MoFlow clusters all predictions along one path. Similarly, in the second and third rows ($T_s = 40$ s), our method captures multiple corridor branches, reflecting the multimodal nature of pedestrian motion in the shopping mall environment. This improved diversity can be attributed to the MoD loss, which encourages each candidate to align with a mixture component of the environment's motion field, promoting mode-consistent diversity rather than random perturbations. Across the examples, the proposed method maintains competitive accuracy while achieving meaningfully diverse predictions aligned with the environment layout.

IV. Conclusion

We presented a framework for long-term human trajectory prediction that integrates a neural implicit motion field into conditional flow matching. A Map-of-Dynamics loss regularizes all trajectory candidates during training to align with environment-specific motion modes, promoting plausible and mode-consistent diversity. At inference, an SDF-based geometric guidance steers predictions away from obstacles without retraining, enabling generalization to unseen environments. Experiments on the ATC dataset show improvements in accuracy, diversity, and physical feasibility.

a) Limitations and future work.: The current framework predicts each person's trajectory independently, without modeling interactions between multiple agents. In crowded scenarios, social forces and collision avoidance between pedestrians can significantly influence individual motion, and ignoring these interactions may lead to implausible predictions in dense settings. Additionally, the geometric guidance relies on a precomputed SDF of static obstacles and does not account for dynamic obstacles such as other pedestrians or moving objects. Future work will extend the framework to multi-agent prediction, incorporating inter-agent interactions into both the generative process and the guidance mechanism. We also plan to explore learned or dynamic obstacle representations to handle time-varying environments, and to evaluate generalization across diverse indoor and outdoor settings.

REFERENCES

[1] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras. "Human motion trajectory prediction: A survey". In: *Int. J. of Robotics Research* 39.8 (2020), pp. 895–935.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. "Social LSTM: Human Trajectory Prediction in Crowded Spaces". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 961–971. DOI: 10.1109/CVPR.2016.110.

[3] Z. Cao, H. Gao, K. Mangalam, Q. Cai, M. Vo, and J. Malik. "Long-term human motion prediction with scene context". In: *ECCV*. 2020.

[4] K. Mangalam, Y. An, H. Girase, and J. Malik. "From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting". In: *ICCV*. 2021, pp. 15213–15222. DOI: 10.1109/ICCV48922.2021.01495.

[5] S. H. Kiss, K. Katuwandeniya, A. Alempijevic, and T. Vidal-Calleja. "Constrained Gaussian Processes With Integrated Kernels for Long-Horizon Prediction of Dense Pedestrian Crowd Flows". In: *IEEE Robotics and Automation Letters* 7.3 (2022), pp. 7343–7350. DOI: 10.1109/LRA.2022.3177849.

[6] N. Gorlo, L. Schmid, and L. Carlone. "Long-Term Human Trajectory Prediction Using 3D Dynamic Scene Graphs". In: *IEEE Robotics and Automation Letters* 9.12 (2024), pp. 10978–10985. DOI: 10.1109/LRA.2024.3482169.

[7] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. "Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks". In: *CVPR*. 2018.

[8] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone. "Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data". In: *ICCV*. Dec. 2020, pp. 683–700. ISBN: 978-3-030-58522-8. DOI: 10.1007/978-3-030-58523-5_40.

[9] F. Giuliari, I. Hasan, M. Cristani, and F. Galasso. "Transformer Networks for Trajectory Forecasting". In: *ICPR*. 2021, pp. 10335–10342. DOI: 10.1109/ICPR48806.2021.9412190.

[10] L. Shi, L. Wang, S. Zhou, and G. Hua. "Trajectory Unified Transformer for Pedestrian Trajectory Prediction". In: *ICCV*. Oct. 2023, pp. 9675–9684.

[11] T. Gu, G. Chen, J. Li, C. Lin, Y. Rao, J. Zhou, and J. Lu. "Stochastic Trajectory Prediction via Motion Indeterminacy Diffusion". In: *CVPR*. 2022, pp. 17113–17122.

[12] W. Mao, C. Xu, Q. Zhu, S. Chen, and Y. Wang. "Leapfrog Diffusion Model for Stochastic Trajectory Prediction". In: *CVPR*. 2023, pp. 5517–5526.

[13] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. "Flow Matching for Generative Modeling". In: *ICLR*. 2023.

[14] A. Tong, K. FATRAS, N. Malkin, G. Huguet, Y. Zhang, J. Rector-Brooks, G. Wolf, and Y. Bengio. "Improving and generalizing flow-based generative models with minibatch optimal transport". In: *TMLR*. 2024.

[15] Y. Fu, Q. Yan, L. Wang, K. Li, and R. Liao. "MoFlow: One-Step Flow Matching for Human Trajectory Forecasting via Implicit Maximum Likelihood Estimation based Distillation". In: *CVPR*. 2025.

[16] Q. Yan et al. "TrajFlow: Multi-modal Motion Prediction via Flow Matching". In: *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*. 2025. URL: https://arxiv.org/abs/2506.08541.

[17] T. P. Kucner et al. "Survey of maps of dynamics for mobile robots". In: *Int. J. of Robotics Research* (2023).

[18] Y. Zhu, A. Rudenko, T. P. Kucner, A. J. Lilienthal, and M. Magnusson. "Long-Term Human Motion Prediction Using Spatio-Temporal Maps of Dynamics". In: *IEEE Robotics and Automation Letters* 10.11 (2025), pp. 12229–12236. DOI: 10.1109/LRA.2025.3619831.

[19] Y. Zhu, S.-M. Yang, A. Rudenko, T. P. Kucner, A. J. Lilienthal, and M. Magnusson. "NeMo-map: Neural Implicit Flow Fields for Spatio-Temporal Motion Mapping". In: *ICLR*. 2026. URL: https://arxiv.org/abs/2510.14827.

[20] D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita. "Person tracking in large public spaces using 3-D range sensors". In: *IEEE Trans. on Human-Machine Systems* 43.6 (2013), pp. 522–534.

[21] A. Bhattacharyya, B. Schiele, and M. Fritz. "Accurate and Diverse Sampling of Sequences Based on a "Best of Many" Sample Objective". In: *CVPR*. 2018, pp. 8485–8493. DOI: 10.1109/CVPR.2018.00885.

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. "Score-based generative modeling through stochastic differential equations". In: *arXiv preprint arXiv:2011.13456* (2020).

[23] S. Molina, G. Cielniak, and T. Duckett. "Robotic Exploration for Learning Human Motion Patterns". In: *IEEE Trans. on Robotics and Automation (TRO)* 38.2 (2022), pp. 1304–1318. DOI: 10.1109/TRO.2021.3101358.

[24] S. Suo, S. Regalado, S. Casas, and R. Urtasun. "TrafficSim: Learning to Simulate Realistic Multi-Agent Behaviors". In: *CVPR*. 2021.