A Spatio-Temporal Flow Matching Framework for Pedestrian Trajectory Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predicting pedestrian trajectories is essential for understanding human behavior and optimizing spatial planning. A key characteristic of pedestrian trajectories is their multimodality, which results from the diverse intentions of individuals. While recent studies have employed various techniques, such as clustering, tree enumeration, and Gaussian mixture models, to address this multimodality, a more natural and efficient approach is to directly model the distribution of trajectories. To address this need, we propose a spatio-temporal aware flow matching framework for pedestrian trajectory prediction. This framework empowers flow matching-based generative models by enabling them to analyze past trajectories of both the subject and their neighbors so as to model the distribution of future trajectories. Benchmarking results demonstrate the superiority of our proposed framework, highlighting its ability to achieve more accurate and efficient trajectory predictions compared to existing methods.

1 Introduction

Predicting pedestrian trajectories from observed paths is important for applications like spatial planning and video surveillance[17, 6]. However, it remains challenging due to three factors: First, complex spatial interactions – pedestrians interact with each other, such as friends walking together or strangers maintaining distance [9]. Second, temporal coherence – future predictions must align with past behaviors. Third, diverse intentions – individuals' varied goals create many possible future paths. These factors highlight the need for models that can address the inherent complexity of pedestrian trajectory prediction.

Previous research has effectively addressed the first two challenges in pedestrian trajectory prediction. First, models have focused on collective pedestrian behavior rather than individual behavior, incorporating the interactions among different pedestrians [3]. Second, these models have accounted for the temporal dimension, allowing them to use past trajectories within a time window to predict future trajectories in a manner consistent with historical data [16, 18]. However, the third challenge, which pertains to the multimodality of trajectory distributions, remains outstanding. Although various solutions have been proposed, each has notable limitations. For instance, models based on Variational Autoencoders (VAE) and Gaussian Mixture Models (GMM) [16, 2] impose strict constraints on the trajectory distribution family, which reduces expressiveness and limits precision. Generative Adversarial Networks (GANs) [3, 11] are typically difficult to train and are prone to mode collapse. Search-based models [12] suffer from inefficiency, with prediction quality heavily dependent on the resolution and exhaustiveness of the enumeration process. Recent approaches that cluster trajectories into distinct modes and represent predictions as weighted averages of these modes [13] offer increased efficiency but inevitably introduce discretization errors during both training and inference.

To effectively model the multimodal distribution of pedestrian trajectories, we propose TrajFM, a spatio-temporal-aware flow matching framework designed to predict arbitrarily shaped trajectory distributions. We evaluated TrajFM on two benchmark datasets for pedestrian trajectory prediction: ETH-UCY [4, 9] and SDD [10]. The evaluation results demonstrate that TrajFM outperforms current state-of-the-art models, thereby confirming the effectiveness of flow matching-based density estimation for trajectory prediction.

2 Method

2.1 Problem formulation

Pedestrian trajectory prediction aims to forecast future trajectories based on observed trajectories of pedestrians and their neighbors. In line with previous work, consider a traffic scene involving N pedestrians, where each pedestrian's trajectory spans T time steps. The position of the i-th pedestrian at time step t is denoted by $\boldsymbol{x}_{i,t} = (x_{i,t}, y_{i,t}) \in \mathbb{R}^2$.

The entire scene is represented by a collection of N trajectories: $\{\boldsymbol{x}_{i,t}\}_{i=1,\ldots,N;\, t=1,\ldots,T}$. For convenience, this collection can be expressed as a 3-D tensor $\mathbf{X} = [\boldsymbol{x}_{i,t}]_{N,T} \in \mathbb{R}^{N\times T\times 2}$, where the first dimension represents pedestrians, the second dimension represents time steps, and the third dimension represents the 2D coordinates. Given the observed trajectories for the first T_{obs} time steps, denoted as $\mathbf{X}_{\text{obs}} = \{\boldsymbol{x}_{i,t}\}_{i=1,\ldots,N;\, t=1,\ldots,T_{\text{obs}}}$, the model is tasked with predicting the trajectories for the remaining $T - T_{\text{obs}}$ time steps, denoted as $\mathbf{X}_{\text{future}} = \{\boldsymbol{x}_{i,t}\}_{i=1,\ldots,N;\, t=T_{\text{obs}}+1,\ldots,T}$.

2.2 Spatio-temporal feature extraction network

The spatio-temporal feature extraction network is designed to encode the observed past trajectories. It consists of two key modules: (1) the temporal attention module, which applies self-attention along the time dimension for each pedestrian, and (2) the spatial attention module, which performs self-attention among pedestrians at each time step. The temporal attention module focuses on capturing the temporal dynamics of each pedestrian's movement by considering the sequence of their past trajectories. Meanwhile, the spatial attention module addresses the interactions between pedestrians at each time step, enabling the model to account for social influences and spatial relationships. Together, these modules generate features that encapsulate both temporal and spatial contexts for each pedestrian, which are then utilized to predict the distribution of future trajectories.

The input to the network is a collection of observed pedestrian trajectories $\mathbf{X}_{\text{obs}} = [\boldsymbol{x}_{i,t}]_{N,T_{\text{obs}}} \in \mathbb{R}^{N,T_{\text{obs}},2}$. For each pedestrian i at each time step t, we first featurize the coordinate using a multi-layer perceptron (MLP) with 3 linear layers and the ReLU activation. To incorporate temporal information, we use the standard positional embedding layer [14] to encode the time step and add the positional embedding to the feature vector.

We combine multiple temporal attention modules and spatial attention modules in an alternating order to fully capture both temporal and inter-pedestrian spatial features. Finally, the feature vectors of the last observed time steps $\{\boldsymbol{h}_i = \boldsymbol{h}_{i,T_{\text{obs}}}\}_{i=1\ldots N}$ are used by the subsequent future trajectory density estimator to predict the trajectory of each pedestrian.

2.3 Flow matching for future trajectory modeling

Preliminary Flow matching is a method for learning a probability flow ψ_t that transforms a source distribution $p_0(\boldsymbol{x})$ into a target distribution $p_1(\boldsymbol{x})$. The probability flow ψ_t is governed by an ordinary differential equation defined by a time-dependent vector field u_t: $\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t = u_t(\boldsymbol{x}_t)$, where $\boldsymbol{x}_t = \psi_t(\boldsymbol{x}_0)$. To model the target distribution, a neural network $v_\theta(\boldsymbol{x}_t, t)$ can be employed to approximate the vector field u_t. However, the true vector field associated with the data distribution is intractable. To address this, the conditional flow matching framework [5] introduces a surrogate vector field that depends only on a specific prior sample $\boldsymbol{x}_0$ and a data sample $\boldsymbol{x}_t$: $u_t(\boldsymbol{x}_t|\boldsymbol{x}_0, \boldsymbol{x}_1) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{x}_t$. This surrogate field can be viewed as an interpolation between $\boldsymbol{x}_0$ and $\boldsymbol{x}_t$. The corresponding loss function aims to train the vector field network to approximate this tractable surrogate vector field, and is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{t,p_1(\boldsymbol{x}_1),p_0(\boldsymbol{x}_0)} \|v_\theta(\boldsymbol{x}_t, t) - u_t(\boldsymbol{x}_t|\boldsymbol{x}_1, \boldsymbol{x}_0)\|^2, \tag{1}$$

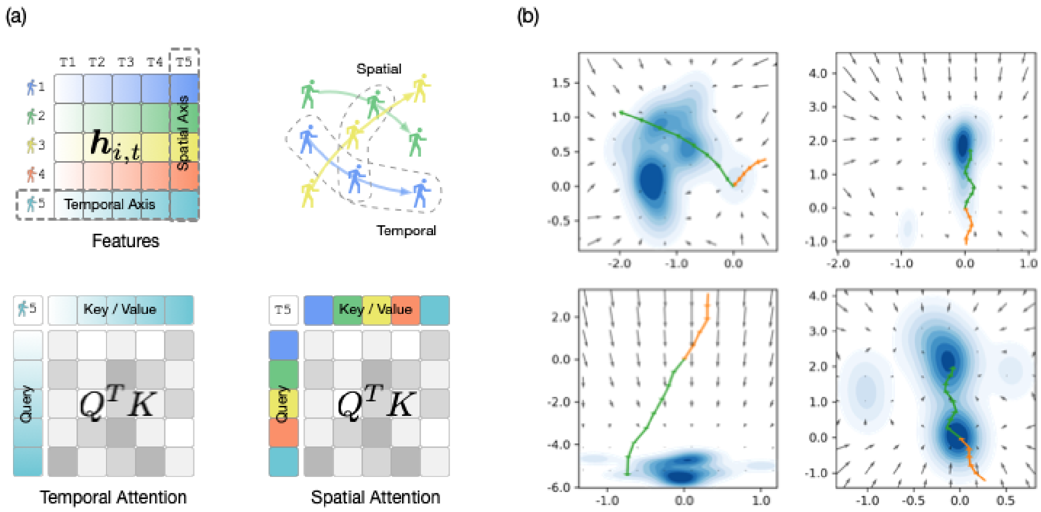

Figure 1: **(a)** Illustration of the spatial-temporal attention. **(b)** Distributions and associated vector fields of a pedestrian's location at a future time point. Orange arrows indicate the pedestrian's past trajectory and green arrows indicate the ground truth future trajectory.

where $t \sim \mathcal{U}(0, 1)$.

Formulation The future trajectory of the pedestrian i can be represented by a sequence of $T_{\text{future}} = T - T_{\text{obs}}$ coordinates: $[\boldsymbol{x}_{i,T_{\text{obs}}+1}, \ldots, \boldsymbol{x}_{i,T}]$. We center the coordinates around the last observed coordinate so that future trajectories are represented in relative terms, denoted by the vector $\boldsymbol{s}_i = [\boldsymbol{x}_{i,T_{\text{obs}}+1} - \boldsymbol{x}_{i,T_{\text{obs}}}, \ldots, \boldsymbol{x}_{i,T} - \boldsymbol{x}_{i,T_{\text{obs}}}] \in \mathbb{R}^{2 \times T_{\text{future}}}$ Formally, modeling the future trajectory distribution of pedestrian i is equivalent to modeling the following probability density:

$$p(\boldsymbol{s}_i | \mathbf{X}_{\text{obs}}), \quad i = 1 \ldots N. \tag{2}$$

Since the feature extraction network has encoded the condition $\mathbf{X}_{\text{obs}}$ into $\boldsymbol{h}_i$, the probability density function can be modified as:

$$p(\boldsymbol{s}_i | \boldsymbol{h}_i), \quad i = 1 \ldots N. \tag{3}$$

To define a vector field for the distribution, we choose the isotropic Gaussian $\mathcal{N}(0, I)$ as the prior, and define the flow connecting a prior sample $\boldsymbol{s}_i^{(0)}$ and a data sample $\boldsymbol{s}_i$ as linear interpolation. The linear interpolation favors a straight flow which contributes to the efficiency of both training and sampling as it is the shortest path between two points. Formally, the probability flow and its associated conditional vector field are defined as:

$$u_t(\boldsymbol{s}_i^{(t)} | \boldsymbol{s}_i^{(1)}, \boldsymbol{s}_i^{(0)}) = \boldsymbol{s}_i^{(1)} - \boldsymbol{s}_i^{(0)} = \frac{\boldsymbol{s}_i^{(1)} - \boldsymbol{s}_i^{(t)}}{1 - t}, \tag{4}$$

$$\psi_t(\boldsymbol{s}_i^{(0)} | \boldsymbol{s}_i^{(1)}) = t\boldsymbol{s}_i^{(1)} + (1 - t). \tag{5}$$

We use a simple three-layer MLP with ReLU activation to parameterize the vector field, which takes as input the current interpolant $\boldsymbol{s}_i^{(t)}$, the timestep t, and the embedding $\boldsymbol{h}_i$. The network is denoted by $v(\boldsymbol{s}_i^{(t)}, t, \boldsymbol{h}_i)$. The conditional flow matching objective for the i-th pedestrian is formulated as:

$$\mathcal{L}_i = \mathbb{E}_{p(\boldsymbol{s}_i^{(1)}), t, p(\boldsymbol{s}_i^{(0)} | \boldsymbol{h}_i),} \left\| v(\boldsymbol{s}_i^{(t)}, t, \boldsymbol{h}_i) - (\boldsymbol{s}_i^{(1)} - \boldsymbol{s}_i^{(0)}) \right\|^2, \tag{6}$$

and the final training loss is the average over all the observed pedestrians.

Sampling The sampling algorithm is outlined in Algorithm 1. First, the observed trajectories are encoded a single time using the transformer-based spatio-temporal network. An initial sample is then drawn from a prior Gaussian distribution. This sample is subsequently updated iteratively using the

Table 1: Benchmarking results on the ETH-UCY dataset using ADE and FDE scores. TrajFM outperforms the previous models in terms of average ADE and FDE scores. TrajFM also achieves the best scores in most cases.

Subset	ETH		HOTEL		UNIV		ZARA1		ZARA2		Average	
Metric	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE	ADE	FDE
Social GAN [3]	0.87	1.62	0.67	1.37	0.76	1.52	0.35	0.68	0.42	0.84	0.61	1.21
SoPhie [11]	0.70	1.43	0.76	1.67	0.54	1.24	0.30	0.63	0.38	0.78	0.51	1.15
STAR [18]	**0.36**	0.64	0.17	0.36	0.31	0.62	0.29	0.52	0.22	0.46	0.26	0.53
SGCN [8]	0.63	1.03	0.32	0.55	0.37	0.70	0.29	0.53	0.25	0.45	0.37	0.65
CAGN [2]	0.41	0.65	0.13	0.23	0.32	0.54	0.21	0.38	0.16	0.33	0.25	0.43
SIT [12]	0.39	0.62	0.14	0.22	0.27	0.47	0.19	0.33	0.16	0.29	0.23	0.38
SocialVAE [16]	0.47	0.76	0.14	0.22	0.25	0.47	0.20	0.37	0.14	0.28	0.24	0.42
SocialVAE+FPC [16]	0.41	0.58	0.13	0.19	**0.21**	0.36	0.17	0.29	**0.13**	0.22	0.21	0.32
PECNet [7]	0.54	0.87	0.18	0.24	0.35	0.60	0.22	0.39	0.17	0.30	0.29	0.48
AgentFormer [19]	0.45	0.75	0.14	0.22	0.25	0.45	0.18	0.30	0.14	0.24	0.23	0.39
MemoNet [15]	0.40	0.61	0.11	0.17	0.24	0.43	0.18	0.32	0.14	0.24	0.21	0.35
TUTR [13]	0.40	0.61	0.11	0.18	0.23	0.42	0.18	0.34	**0.13**	0.25	0.21	0.36
TrajFM	0.39	**0.57**	**0.10**	**0.13**	**0.21**	**0.32**	**0.15**	**0.24**	0.17	**0.22**	**0.20**	**0.30**

Euler method, based on the predicted vector field. Given that the vector field network is a lightweight MLP and the feature extraction network is evaluated only once at the beginning, the iterative sampling process remains efficient. The sample obtained after the final iteration represents the generated future trajectory.

3 Experiments

Dataset Following the standard set by previous work[2, 13], we use the ETH-UCY[4, 9] dataset to benchmark the performance of TrajFM and the baselines. ETH-UCY contains 5 subsets: ETH, HOTEL, UNIV, ZARA1, and ZARA2. In total, the dataset contains 1,536 pedestrian trajectories. In accordance with previous work[16, 15], we train five models with each model using a different subset for testing and the rest subsets for training.

Evaluation metrics Models are evaluated using the Average Displacement Error (ADE) and the Final Displacement Error (FDE) [1]. ADE measures the similarity between the predicted trajectory and the ground-truth trajectory. FDE emphasizes the difference between the predicted end point and the ground truth.

Result TrajFM demonstrates state-of-the-art performance in both average ADE and average FDE, as shown in Table 1. It improves upon the best ADE of previous methods, reducing it from 0.21 to 0.19, and decreases the FDE from 0.32 to 0.30. Specifically, TrajFM achieves the lowest FDE score across all five ETH-UCY subsets and the best ADE score on three out of five subsets. The ADE scores for TrajFM on the ETH and ZARA2 subsets are comparable to those of the leading baseline models.

Figure 2 in the appendix presents two examples of predicted trajectory distributions from the ETH-UCY dataset. Given the multi-dimensional nature of trajectory distributions, we employ two visualization techniques. First, we project the distributions onto a two-dimensional plane (shown in the first column of Figure 2). This projection displays the contour of the predicted trajectories, revealing that (1) TrajFM assigns high probability to the ground truth trajectory and (2) the branching shape of the distribution indicates the incorporation of multiple trajectory modes.

4 Conclusion

In this paper, we present TrajFM, a spatio-temporal flow matching framework specifically created for predicting pedestrian trajectories. TrajFM is a novel generative model for spatio-temporal data, and we have shown its effectiveness in capturing pedestrian movement patterns. This framework also holds potential for broader applications in general time series data generation.

References

[1] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 961–971, 2016.

[2] Jinghai Duan, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Liushuai Shi, and Gang Hua. Complementary attention gated network for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 542–550, 2022.

[3] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2255–2264, 2018.

[4] Alon Lerner, Yiorgos Chrysanthou, and Dani Lischinski. Crowds by example. In *Computer graphics forum*, volume 26, pages 655–664. Wiley Online Library, 2007.

[5] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[6] Matthias Luber, Johannes A Stork, Gian Diego Tipaldi, and Kai O Arras. People tracking with human motion predictions from social forces. In *2010 IEEE international conference on robotics and automation*, pages 464–469. IEEE, 2010.

[7] Karttikeya Mangalam, Harshayu Girase, Shreyas Agarwal, Kuan-Hui Lee, Ehsan Adeli, Jitendra Malik, and Adrien Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 759–776. Springer, 2020.

[8] Abduallah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-stgcnn: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14424–14432, 2020.

[9] Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, pages 261–268. IEEE, 2009.

[10] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Learning social etiquette: Human trajectory understanding in crowded scenes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 549–565. Springer, 2016.

[11] Amir Sadeghian, Vineet Kosaraju, Ali Sadeghian, Noriaki Hirose, Hamid Rezatofighi, and Silvio Savarese. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1349–1358, 2019.

[12] Liushuai Shi, Le Wang, Chengjiang Long, Sanping Zhou, Fang Zheng, Nanning Zheng, and Gang Hua. Social interpretable tree for pedestrian trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2235–2243, 2022.

[13] Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9675–9684, 2023.

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[15] Chenxin Xu, Weibo Mao, Wenjun Zhang, and Siheng Chen. Remember intentions: Retrospective-memory-based trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2022.

[16] Pei Xu, Jean-Bernard Hayet, and Ioannis Karamouzas. Socialvae: Human trajectory prediction using timewise latents. In *European Conference on Computer Vision*, pages 511–528. Springer, 2022.

[17] Masahiro Yasuno, Noboru Yasuda, and Masayoshi Aoki. Pedestrian detection and tracking in far infrared images. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 125–125. IEEE, 2004.

[18] Cunjun Yu, Xiao Ma, Jiawei Ren, Haiyu Zhao, and Shuai Yi. Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 507–523. Springer, 2020.

[19] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M Kitani. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.

A Additional Figures

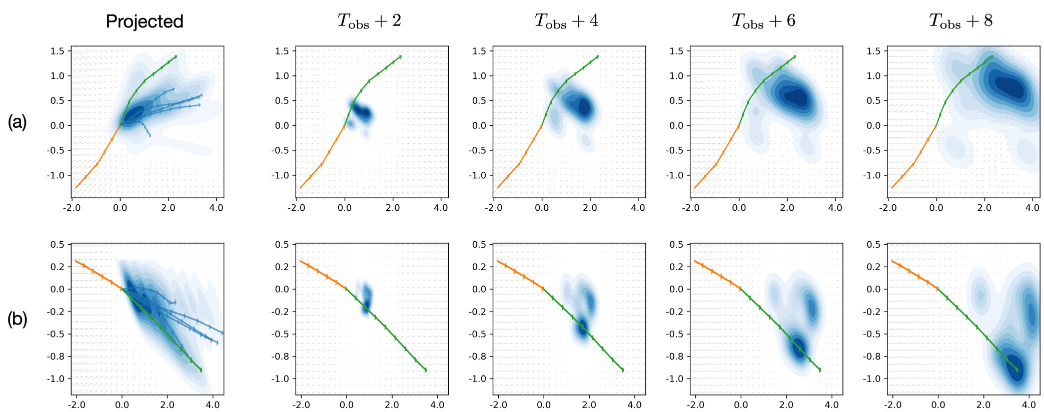

Figure 2: First column from the left: trajectory distributions and vector fields projected to the 2D plane. The rest of the columns: marginal distribution and its associated vector field of the trajectory at different time steps, interpreted as the distribution of the pedestrian coordinates at different time steps. Orange arrows indicate observed past trajectories, green arrows indicate ground truth future trajectories and blue arrows indicate predicted future trajectories.

Algorithm 1 Sampling Algorithm

Input: $\mathbf{X}_{\text{obs}}$ observed trajectories
Input: n integration steps

1: Extract spatio-temporal feature for observed trajectories: $\{\boldsymbol{h}_i\} = \text{Featurize}(\mathbf{X}_{\text{obs}})$
2: Sample from prior: $\boldsymbol{s}_i^{(0)} \sim \mathcal{N}(0, I)$
3: **for** $t \leftarrow 1$ to n **do**
4:     **for** $i \leftarrow 1$ to N in parallel **do**
5:         $\boldsymbol{s}_i^{(\frac{t}{n})} \leftarrow \text{Euler}\left(v(\boldsymbol{s}_i^{(\frac{t-1}{n})}, \frac{t}{n}, \boldsymbol{h}_i), \boldsymbol{s}_i^{(\frac{t-1}{n})}, \frac{1}{n}\right)$
6:     **end for**
7: **end for**
8: **return** $\boldsymbol{s}_i^{(0)}$