NeMo-map: Neural Implicit Flow Fields for Spatio-Temporal Motion Mapping

Anonymous authors
Paper under double-blind review

Abstract

Safe and efficient robot operation in complex human environments can benefit from good models of site-specific motion patterns. Maps of Dynamics (MoDs) provide such models by encoding statistical motion patterns in a map, but existing representations use discrete spatial sampling and typically require costly offline construction. We propose a continuous spatio-temporal MoD representation based on implicit neural functions that directly map coordinates to the parameters of a Semi-Wrapped Gaussian Mixture Model. This removes the need for discretization and imputation for unevenly sampled regions, enabling smooth generalization across both space and time. Evaluated on a large public dataset with long-term real-world people tracking data, our method achieves better accuracy of motion representation and smoother velocity distributions in sparse regions while still being computationally efficient, compared to available baselines. The proposed approach demonstrates a powerful and efficient way of modeling complex human motion patterns.

1 Introduction

Safe and efficient operation in complex, dynamic and densely crowded human environments is a critical prerequisite for deploying robots in various tasks to support people in their daily activities. Extending the environment model with human motion patterns using a *map of dynamics* (MoD) is one way to achieve unobtrusive navigation, compliant with existing site-specific motion flows (Palmieri et al., 2017; Swaminathan et al., 2022).

As illustrated in Fig. 1, incorporating MoDs into motion planning provides benefits in crowded environments, since they encode information about the expected motion outside of the robot's sensor range, allowing for less reactive behavior. In the example shown in Fig. 1, the oncoming pedestrian flow is initially outside the robot's observation radius. Without MoD awareness, the robot chooses a direct path to the goal but later becomes trapped in the oncoming crowd. In contrast, a planner informed by MoDs can exploit prior knowledge of human motion patterns to generate a trajectory that aligns with the expected flow, allowing the robot to reach the goal safely and efficiently. MoDs can also be applied to long-term human motion prediction (Zhu et al., 2023). As shown in the right of Fig. 1, MoDs help predict realistic trajectories that implicitly respect the complex topology of the environment, such as navigating around corners or avoiding obstacles.

Several approaches have been proposed for constructing MoDs. Early methods modeled human motion on occupancy grid maps, treating dynamics as shifts in occupancy (Wang et al., 2015; 2016). These approaches struggle with noisy or incomplete trajectory data. Later, velocity-based representations have been introduced, most notably the CLiFF-map (Kucner et al., 2017), which models local motion patterns with Gaussian mixture models, effectively captures multimodality in human flows and has been successfully used in both robot navigation and prediction tasks. The methods above are computed in batch, given a set of observations. Online learning methods have also been explored to update motion models as new observations arrive (Zhu et al., 2025), allowing robots to adapt to changing environments without costly retraining from scratch. Temporal MoDs have also been explored, including STeF-maps (Molina et al., 2022), which apply frequency-based models to encode periodic variations in the flow.

However, existing MoD representations require spatial discretization, with a manually selected map resolution for point locations and interpolation to estimate motion at arbitrary positions. This dis-

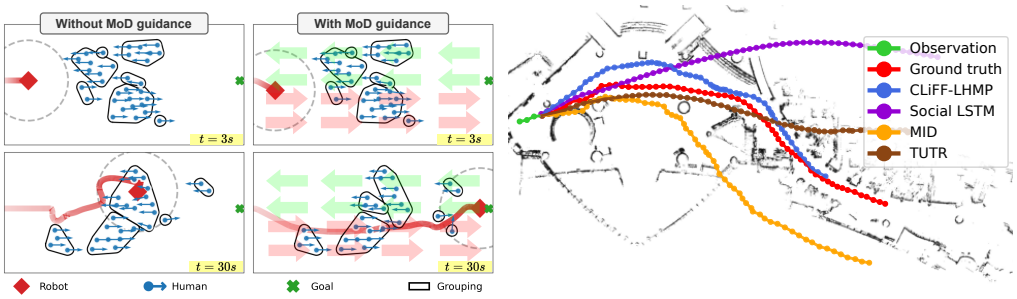

Figure 1: Indicative applications of Maps of dynamics (MoD) for motion planning and human motion prediction. **Left**: Illustration of how MoDs can support socially aware motion planning. The robot (red diamond) navigates toward the goal (green cross) in the presence of two opposing human flows, with the underlying MoD shown as colored arrows. In this scenario, the oncoming pedestrian flow moving in the opposing direction is initially outside the robot's observation radius (grey dashed circle). Without guidance from the MoD, the planner initially takes a direct path to the goal but eventually becomes stuck and collides with the oncoming flow. In contrast, when informed by the MoD, the robot aligns its trajectory with the motion patterns and reaches the goal efficiently and safely. **Right**: Human motion prediction with a 60 s horizon. The **red** line represents the ground truth trajectory and the **green** line represents the observed trajectory. With MoD guidance, CLiFF-LHMP Zhu et al. (2023) makes more accurate and realistic predictions than deep learning methods. While the trajectories predicted by Social LSTM (Alahi et al., 2016), TUTR (Shi et al., 2023) and MID (Gu et al., 2022) often are unfeasible (e.g., crossing the walls), CLiFF-LHMP predictions implicitly follow the topology of the environment.

cretization introduces information loss, reduces flexibility, and complicates tuning across different environments.

To address these challenges, in this work, instead of representing motion patterns on a discrete grid, we propose a *continuous map of dynamics* using implicit neural representation. We learn a neural function that maps *spatio-temporal coordinates* to parameters of a local motion distribution. Implicit neural representations have emerged as powerful tools for encoding continuous functions, providing compact and differentiable models with strong generalization. Leveraging these properties, this formulation allows the model to smoothly generalize across space and time, while maintaining multimodality in places where flows tend to go in more than one direction since it produces a wrapped Gaussian mixture model of expected motion given a query location and time.

We evaluate our approach on real-world datasets of human motion and show that continuous MoDs not only improve representation accuracy but can also drastically reduce map construction time. Our method yields smoother and more consistent velocity distributions, resulting in more accurate representations of human motion patterns. In contrast to baseline approaches that rely on time-consuming per-cell motion modeling, it computes nearly two orders of magnitude faster than CLiFF-map Kucner et al. (2017). Unlike the faster but discretised representation of STeF-map Molina et al. (2022), our method preserves non-discretised directions, yielding results closer in spirit to CLiFF.

In summary, the main contribution of this work is an entirely novel representation of flow-aware maps of dynamics, named NeMo-map. In contrast to existing methods, NeMo allow *continuous spatio-temporal queries* to generate location- and time-specific *multimodal flow predictions*. As evidenced by our experimental validation on real-world human motion data, NeMo efficiently learns a highly accurate statistical representation of motion in large-scale maps.

2 Related Work

A *map of dynamics* (MoD) is a representation that augments the geometric map of an environment with statistical information about observed motion patterns. Unlike static maps, MoDs incorporate spatio-temporal flow information, allowing robots to reason about how humans typically move in a given environment.

MoDs can be built from various sources of input, such as trajectories (Bennewitz et al., 2005), dynamics samples, or information about the flow of continuous media (e.g., air or water) (Bennetts et al., 2017). Furthermore, these models can feature diverse underlying representations, including evidence grids, histograms, graphs, or Gaussian mixtures.

There are several types of MoDs described in the literature, generally striving to provide an efficient tool for storing and querying information about historical or expected changes in states within the environment. Occupancy-based methods focus on mapping human dynamics on occupancy grid maps, modeling motion as shifts in occupancy (Wang et al., 2015; 2016). Trajectory-based methods extract human trajectories and group them into clusters, with each cluster representing a typical path through the environment (Bennewitz et al., 2005). These approaches suffer from noisy or incomplete trajectories. To address this, Chen et al. (2016) formulate trajectory modeling as a dictionary learning problem and use augmented semi-nonnegative sparse coding to find local motion patterns characterized by partial trajectory segments.

MoDs can also be based on velocity observations. With velocity mapping, human dynamics can be modeled through flow models. Kucner et al. (2017) presented a probabilistic framework for mapping velocity observations, which is named Circular-Linear Flow Field map (CLiFF-map). CLiFF-map represents local flow patterns as a multi-modal, continuous joint distribution of speed and orientation, as further described in Sec. 3. A benefit of CLiFF-map is that it can be built from incomplete or spatially sparse velocity observations (de Almeida et al., 2024), without the need to store a long history of data or deploy advanced tracking algorithms. CLiFF-maps are typically built offline, for the reason of high computational costs associated with the building process. This constraint limits their applicability in real environments.

When building flow models, temporal information can also be incorporated. Molina et al. (2022) apply the Frequency Map Enhancement (FreMEn Krajník et al. (2017)), which is a model describing spatio-temporal dynamics in the frequency domain, to build a time-dependent probabilistic map to model periodic changes in people flow called STeF-map. The motion orientations in STeF-map are discretized. Another method of incorporating temporal information is proposed by Zhi et al. (2019). Their approach uses a kernel recurrent mixture density network to provide a multimodal probability distribution of movement directions of a typical object in the environment over time, though it models only orientation and not the speed of human motion.

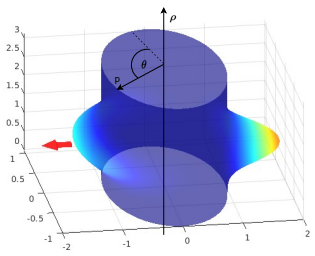

Figure 2: Probability density of a Semi-Wrapped Gaussian Mixture Model (SWGMM) with two components, visualized on a cylinder. Orientation θ is wrapped around the circular axis, while speed ρ extends along the vertical axis. The representation allows joint modeling of angular (orientation) and linear (speed) variables, capturing multimodality in motion patterns.

It is important to note that a map of dynamics is not a trajectory prediction model. Whereas trajectory predictors (e.g., LSTMs) aim to forecast the future state of agents by propagating state information forward in time from an initial state, our goal is fundamentally different. We seek to construct a spatio-temporal prior that encodes the distribution of motion patterns in the environment itself. This prior can be queried directly at any spatial coordinate and any time of day, providing motion statistics that can support downstream tasks such as planning or long-term prediction, but it does not by itself generate trajectories for individual agents.

3 Methodology

3.1 Probabilistic Modeling of Human Motion

Our spatio-temporal map of dynamics produces probability distributions over human motion velocities. A velocity $\mathbf{v}$ is defined by the pair of *speed* (a positive linear variable $\rho \in \mathbb{R}^+$) and *orientation* (a circular variable $\theta \in [0, 2\pi)$).

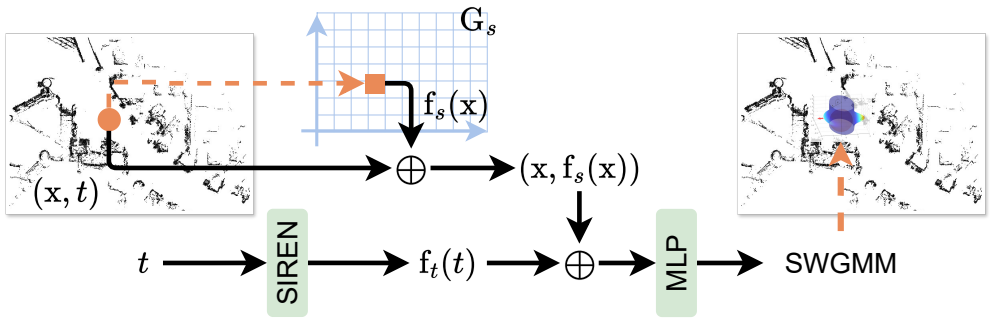

Figure 3: Method overview. A spatio-temporal query $(\mathbf{x}, t)$ is mapped to parameters of a Semi-Wrapped Gaussian Mixture Model (SWGMM). The spatial coordinate $\mathbf{x}$ is used to interpolate features from a learnable spatial grid $\mathbf{G}_s$, and the temporal coordinate t is encoded using a SIREN network. The spatial features $\mathbf{f}_s(\mathbf{x})$, temporal encoding $\mathbf{f}_t(t)$, and raw coordinates are concatenated and passed through an MLP, which outputs the parameters of a SWGMM, providing a continuous, multimodal probabilistic representation of motion dynamics at the queried location and time.

To capture the statistical structure of such data, we model human motion patterns with a *Semi-Wrapped Gaussian Mixture Model* (SWGMM), similar to the CLiFF-map representation (Kucner et al., 2017). While a von Mises distribution would be effective for purely angular variables, is not suitable when combining circular and linear components. Roy et al. (2012) proposed the von Mises-Gaussian mixture model (VMGMM) to jointly represent one circular variable and linear variables. However, their model assumes independence between the circular and linear dimensions, which limits its ability in capturing real-world correlations. To overcome this, SWGMM (Roy et al., 2016) jointly models circular-linear variables and allows correlations between them.

An SWGMM models velocity $\mathbf{v} = [\rho, \theta]^\top$ as a mixture of J Semi-Wrapped Normal Distributions (SWNDs):

$$p(\mathbf{v} \mid \boldsymbol{\xi}) = \sum_{j=1}^{J} w_j \mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j}(\mathbf{v}), \tag{1}$$

where $\boldsymbol{\xi} = \{\xi_j = (w_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\}_{j=1}^{J}$ denotes a finite set of components of the SWGMM. Each w_j is a mixing weight and satisfies $0 \leq w_j \leq 1$), $\boldsymbol{\mu}_j$ the component mean, and $\boldsymbol{\Sigma}_j$ the covariance. An SWND $\mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}$ is formally defined as

$$\mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}(\mathbf{v}) = \sum_{k \in \mathbb{Z}} \mathcal{N}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\left([\rho, \theta]^\top + 2\pi[0, k]^\top\right), \tag{2}$$

where k is a winding number. In practice, the PDF can be approximated adequately by taking $k \in \{-1, 0, 1\}$ (Mardia & Jupp, 2008).

The SWGMM density function over velocities can be visualized as a function on a cylinder, as shown in Fig. 2. Orientation values θ are wrapped on the unit circle and the speed ρ runs along the length of the cylinder. This formulation yields a flexible and interpretable probabilistic representation of local human motion, capturing multimodality and correlations between orientation and speed.

3.2 Learning Continuous Motion Fields

Previous MoD approaches, such as CLiFF-maps and STeF-maps, rely on discretizing the environment into cells and fitting local probability models. Discretization leads to information loss and prevents querying at arbitrary locations. We address this by introducing a *continuous* map of dynamics parameterized by a *neural implicit* representation. The method overview is shown in Fig. 3.

In many real-world environments, human motion patterns exhibit strong daily periodicity, such as morning and evening rush hours or lunchtime activity. Motivated by this structure, we model time

as a periodic variable and condition the MoD on the time of day. This assumption allows the representation to capture long-term temporal variations without requiring sequential rollouts, and enables efficient queries of motion dynamics at arbitrary spatio-temporal coordinates (x, y, t).

Problem statement. Given a dataset $\mathcal{D}$ of $\mathcal{N}$ spatio-temporal motion samples:

$$\mathcal{D} = \{(\mathbf{x}_i, t_i, \mathbf{v}_i)\}_{i=1}^{N},$$

where $\mathbf{x}_i \in \mathbb{R}^2$ is the spatial coordinate, $t_i \in [0, 1]$ is the normalized time of a day, and $\mathbf{v}_i = [\rho_i, \theta_i]^\top$ is observed velocity, we learn a continuous function Φ_θ that maps a spatio-temporal coordinate $(\mathbf{x}, t)$ to SWGMM parameters:

$$\Phi_\theta(\mathbf{x}, t) = \Big\{ w_j(\mathbf{x}, t),\ \boldsymbol{\mu}_j(\mathbf{x}, t),\ \boldsymbol{\Sigma}_j(\mathbf{x}, t) \Big\}_{j=1}^{J}, \tag{3}$$

where J is the number of mixture components, weights $w_j \geq 0$ and $\sum_{j=1}^{J} w_j = 1$. Each of the j components models the joint velocity $\mathbf{v} = [\rho, \theta]^\top$ with a Semi-Wrapped Normal Distribution $\mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\Sigma},\boldsymbol{\mu}}$. At inference time, querying Φ_θ at any coordinate yields the full set of SWGMM parameters, resulting in a continuous probabilistic representation of motion dynamics. This formulation enables the model to learn smooth, continuous motion fields while retaining the multimodal characteristic of human motion.

Architecture. In our neural representation, we parameterize Φ_θ with a fully connected multilayer perceptron (MLP), conditioned on both spatial and temporal features:

$$\underbrace{\mathbf{f}_s(\mathbf{x})}_{\text{spatial features}} \in \mathbb{R}^{C_s}, \qquad \underbrace{\mathbf{f}_t(t)}_{\text{temporal encoding}} \in \mathbb{R}^{C_t}.$$

For spatial features, a learnable grid $\mathbf{G}_s \in \mathbb{R}^{H \times W \times C_s}$ is queried at location $\mathbf{x}$ by bilinear interpolation, producing $\mathbf{f}_s(\mathbf{x})$. This captures local variations in motion patterns while remaining continuous in space.

For temporal encoding, we encode t with SIREN, the sinusoidal representation network (Sitzmann et al., 2020), which uses periodic activation functions throughout the network.

The MLP input concatenates the raw coordinates and the spatial and temporal features, $\mathbf{z} = [\,\mathbf{x},\ t,\ \mathbf{f}_s(\mathbf{x}),\ \mathbf{f}_t(t)\,]$, and outputs SWGMM parameters. This feature-conditioned representation enables the model to flexibly encode local variations in motion dynamics while maintaining global smoothness across both space and time.

Likelihood and training. For a spatio-temporal coordinate $(\mathbf{x}_i, t_i)$, the velocity likelihood under the predicted SWGMM is

$$p(\mathbf{v}_i \mid \Phi_\theta(\mathbf{x}_i, t_i)) = \sum_{j=1}^{J} w_j(\mathbf{x}_i, t_i)\, \mathcal{N}^{\mathrm{SW}}_{\boldsymbol{\mu}_j(\mathbf{x}_i,t_i),\, \boldsymbol{\Sigma}_j(\mathbf{x}_i,t_i)}(\mathbf{v}_i), \tag{4}$$

where $\mathcal{N}^{\mathrm{SW}}$ denotes the semi-wrapped normal distribution that wraps the angular component (see Eq. (2)). The model is trained by minimizing the negative log-likelihood of motion samples from the dataset under the probability density function (PDF) produced by the model:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log\, p\big(\mathbf{v}_i \mid \Phi_\theta(\mathbf{x}_i, t_i)\big). \tag{5}$$

4 Results

4.1 Dataset

To evaluate spatio-temporal maps of dynamics that capture changes of human motion patterns over time, it is essential to use datasets that span multiple days and reflect variations in human motion patterns throughout the day. Our experiments were conducted using a real-world dataset, ATC (Brščić et al., 2013), which provide sufficient multi-day coverage for evaluation. This dataset was collected

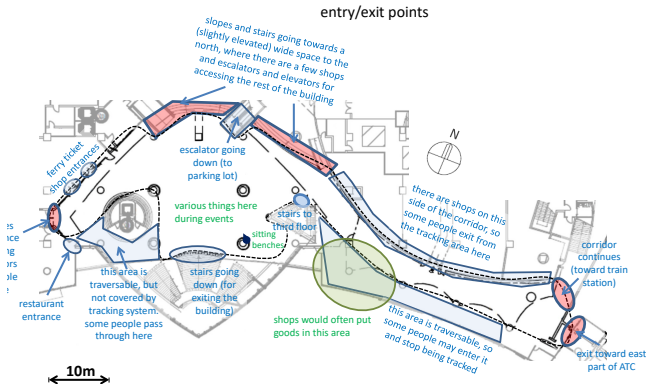

Figure 4: Layout of the ATC dataset environment (Brščić et al., 2013), showing the main east corridor and open areas annotated with semantic information such as entry and exit points, shops, seating areas, and stairs.

in a shopping mall in Japan using multiple 3D range sensors, recording pedestrian trajectories between 9:00 and 21:00, over a total of 92 days. ATC covers a large indoor environment, with a total area covered of approximately $900\,\mathrm{m}^2$. Because of the large scale of the dataset, we use first four days in the dataset (2012 Oct 24, 2012 Oct 28, 2012 Oct 31, and 2012 Nov 04) for experiments. The data from October 24 is used for training, while the other three days are used for evaluation. The observation rate is downsampled from over $10\,\mathrm{Hz}$ to $1\,\mathrm{Hz}$. After downsampling, the training set contains 717,875 recorded motion samples, and the test set contains 5,114,478 samples.

4.2 Baselines

Circular-Linear Flow Field Map (CLiFF-map) CLiFF-map (Kucner et al., 2017) represents motion patterns by associating each discretized grid location with an SWGMM fitted from local observations. The environment is divided into a set of grid locations, and each grid location aggregates motion samples within a fixed radius. The SWGMM parameters at each grid location are estimated via expectation-maximization (EM) (Dempster et al., 1977), with the number and initial positions of mixture components determined using mean shift clustering (Cheng, 1995). When training the CLiFF-map, the convergence precision is set to 1e–5 for both mean shift and EM algorithms, with a maximum iteration count of 100. The grid resolution is set to $1\,\mathrm{m}$. To evaluate different hours, we train separate CLiFF-maps for each hour using the motion samples observed during that time.

Spatio-Temporal Flow Map (STeF-map) STeF-map (Molina et al., 2022) is a spatio-temporal map of dynamics that models the likelihood of human motion directions using harmonic functions. Each grid location maintains k_{stef} temporal models, corresponding to k_{stef} discretized orientations of people moving through that location over time. By modeling periodic patterns, STeF-map captures long-term temporal variations in crowd movements and can predict activities at specific times of day under the assumption of periodicity in the environment. Following Molina et al. (2022), we set $k_{\mathrm{stef}} = 8$ in the experiments, and the model orders for training STeF-map, i.e. the number of periodicities, is set to 2.

Online CLiFF-map Online CLiFF-map (Zhu et al. (2025)) extends the static CLiFF model by updating the SWGMM parameters incrementally as new motion observations become available. Each grid location maintains an SWGMM, which is initialized upon first receiving observations and subsequently updated using the stochastic expectation-maximization (sEM) algorithm (Cappé & Moulines (2009)). In sEM, the expectation step of the original EM algorithm is replaced by a stochastic approximation step, while the maximization step remains unchanged. Like the static CLiFF-map, Online CLiFF-map outputs SWGMM parameters at each grid location, but supports continuous adaptation over time. In the experiments, we follow a spatio-temporal setting by generating an online CLiFF-map for each hour. Observations collected in an hour interval are treated as the new data batch for updating the SWGMMs, producing a temporally adaptive representation of motion dynamics.

4.3 Implementation Details

The output of the network Φ_θ parameterizes an SWGMM over speed and orientation. For J mixture components, the network predicts $6J$ raw values per query coordinate. Each component j is defined by: a mixture weight w_j, obtained by applying a softmax over the raw weights; a mean speed $\mu_{j,s} = \max(0, \tilde{\mu}_{j,s})$ and mean orientation $\mu_{j,a} = \tilde{\mu}_{j,a} \bmod 2\pi$; variances $\sigma^2_{j,s} = \exp(\mathrm{clamp}(\tilde{v}_{j,s}, -10, 10))$ and $\sigma^2_{j,a} = \exp(\mathrm{clamp}(\tilde{v}_{j,a}, -10, 10))$; and a correlation coefficient $\rho_j = 0.99 \tanh(\tilde{\rho}_j)$. Altogether, the network defines a valid SWGMM with parameters as in Eq. (3), where $\boldsymbol{\mu}_j = \big(\mu_{j,s}, \mu_{j,a}\big)$ and $\boldsymbol{\Sigma}_j$ is the covariance matrix with diagonal entries $\sigma^2_{j,s}$, $\sigma^2_{j,a}$ and correlation ρ_j. In the experiments, J is set to 3 and coordinates are normalized to $[-1, 1]$. Spatial input is processed by an MLP with hidden sizes $[128, 64]$ and ReLU activations. Temporal input is processed by a two-layer SIREN (sine activations with $\omega_0^{(1)} = 30$ in the first layer and $\omega_0^{(h)} = 1$ in the hidden layer). The two streams are fused via FiLM modulation (Perez et al., 2018). The fused representation is passed to a linear head producing $6J$ outputs. Models are trained using the Adam optimizer with learning rate 10^{-3} for 100 epochs. An ablation of alternative temporal encodings is provided in Sec. 4.6.

4.4 Quantitative Results

To quantitatively evaluate the accuracy of modeling human motion patterns (MoD quality), we use the negative log-likelihood (NLL). An MoD represents human motion as a probability distribution over velocity conditioned on a spatio-temporal coordinate (x, y, t), implemented as either an SWGMM (our method and CLiFF-maps) or a histogram (STeF-map). To evaluate representation accuracy, we use test data consisting of observed human motions in the same environment. For each test sample (x, y, t), we query the MoD to obtain the predicted distribution and compute the likelihood of the observed motion under this distribution. A higher likelihood indicates that the predicted distribution better aligns with the observed data. We report NLL for numerical stability and easy comparison, so lower NLL values correspond to more accurate motion representations, i.e., higher quality MoDs.

Table 1 reports the accuracy results. Our method achieves the lowest NLL (0.775 ± 2.052), outperforming all baselines. Online CLiFF-map, CLiFF-map, and STeF-map exhibit significantly higher NLLs, with paired t-tests showing $p < 0.001$ under the null hypothesis that baseline NLL is less than or equal to ours. The reductions relative to our method are respectively $+0.752$ (online CLiFF), $+1.189$ (CLiFF), and $+4.801$ (STeF), all with 95% confidence intervals.

Compared with STeF-maps, methods based on SWGMM, such as ours and CLiFF-map, offer two key advantages. They jointly model speed and orientation, whereas STeF-maps do not include speed information. In addition, SWGMMs represent orientation continuously rather than through a discretized 8-bin histogram as in STeF-map. These aspects lead to a more accurate representation of human motion and contribute to the improved performance.

Limitations of CLiFF-maps are from discretizing the environment into grid cells, with each cell storing a locally fitted SWGMM. This grid-based design limits spatial resolution and introduces discontinuities at cell boundaries in both space and time. In particular, dividing time into hourly intervals is a coarse approximation that can produce abrupt changes, since human motion patterns do not necessarily shift at exact hour boundaries. In contrast, our method models the MoD as a continuous neural implicit representation. This enables smooth generalization across space and time, supports queries at arbitrary spatio-temporal coordinates, and provides a compact representation that avoids the memory overhead of storing distributions for every grid cell.

We also compare the map building time of the baselines against our approach as shown in Table 2. For the baselines, the training time corresponds to convergence on all grid cells, while for our method it corresponds to the neural network training time. Our method trains in 19 minutes, substantially faster than CLiFF-map (over 30 hours) while achieving higher accuracy. These results highlight the practicality of continuous MoDs for real-time applications, combining both accuracy and efficiency.

Table 1: Accuracy evaluation on the ATC dataset using average negative log-likelihood (NLL), where lower values indicate better accuracy. We report mean ± standard deviation, together with the reduction in NLL relative to our method and the corresponding 95% confidence interval (CI).

Method	NLL↓	NLL reduction (vs Ours)	95% CI
Ours	**0.775 ± 2.052**	–	–
Online CLiFF-map	1.527 ± 4.156	+0.752	[0.749, 0.755]
CLiFF-map	1.964 ± 4.953	+1.189	[1.185, 1.192]
STeF-map	5.576 ± 9.314	+4.801	[4.794, 4.809]

Table 2: Training and inference times for map building on the ATC dataset. Lower values indicate faster performance. Experiments were conducted on a desktop computer equipped with an Intel i9-12900K CPU and an NVIDIA GeForce RTX 3060 GPU running Ubuntu 20.04.

Method	Train time (minute)↓	Inference time (second)↓
Ours	19.26	1.363×10^{-6}
Online CLiFF-map	23.859	1.914×10^{-3}
CLiFF-map	1831	1.914×10^{-3}
STeF-map	0.815	5.665×10^{-5}

4.5 Qualitative Results

Examples of NeMo-map are shown in Fig. 5. The model is queried at regular spatial intervals at three different times of day, at locations where human motion appears in the training dataset. Across the day, the map adapts smoothly to changes in human motion patterns. For example, in the east corridor (right side of the ATC map), the flow is directed left/upwards in the morning, shifts direction at noon, where pedestrians keep left when facing oncoming flows, and turns into right/downwards in the evening. (These patterns are most clearly seen when displaying only the SWGMM mixture component with the largest weight, in the bottom row, but please note that the map maintains a representation of the full multimodal distribution at all times.) The generated flow fields capture such temporal variations and implicitly align with the environment's topology, even though no explicit map was provided during training. For instance, speeds decrease near resting benches, motion flows pass through exits, and flows follow the corridors.

4.6 Ablation Study

We perform an ablation study on alternative methods for temporal encoding. In our method, we use a SIREN network to process the temporal input. For comparison, we evaluate two alternative mappings of time t into a temporal feature vector $\mathbf{f}_t(t)$:

- **Temporal grid.** A learnable grid $\mathbf{G}_t \in \mathbb{R}^{K \times C_t}$ that captures daily periodicity, where K is the number of discretized time bins (set to 24). The grid feature corresponding to each time bin is concatenated with the spatial feature and passed through an MLP with hidden sizes $[128, 64]$ and ReLU activations.
- **Fourier features.** The time input t is mapped into a periodic embedding using Fourier features Tancik et al. (2020); Mildenhall et al. (2020). For F frequencies, we construct

$$\mathbf{f}_t(t) = \left[\sin(2^n 2\pi t),\ \cos(2^n 2\pi t)\right]_{n=0}^{F-1}.$$

  This representation enables the model to capture time-dependent variations at multiple resolutions. The implementation is identical to the temporal grid variant, except the time grid is replaced by Fourier features with $F = 4$, yielding an 8-dimensional temporal embedding.

Table 3 summarizes the results of the ablation study on temporal encoding. Replacing SIREN with a temporal grid or Fourier features results in higher NLL, confirming the advantage of using SIREN for modeling continuous temporal dynamics. Among the alternatives, Fourier features outperform the temporal grid, but both remain less accurate than SIREN. The reductions in NLL relative to

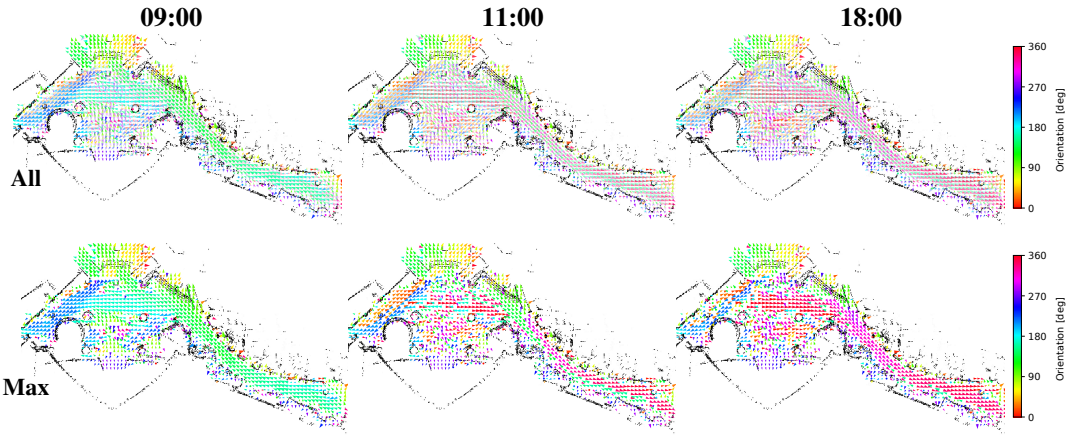

Figure 5: NeMo-map in the ATC dataset, for 09:00 (**left**), 11:00 (**middle**) and 18:00 (**right**), showing changes of motion patterns throughout the day. Predicted Semi-Wrapped Gaussian Mixture Models (SWGMMs) are visualized. At each location, arrow color encodes orientation and arrow length encodes speed. The **top** row shows multimodality by rendering all SWGMM components with transparency proportional to their weights, while the **bottom** row more clearly shows the dominant flow, only displaying the mixture component with the largest weight.

our method are 0.082 for the temporal grid and 0.063 for Fourier features, with 95% confidence intervals.

Table 3: Comparing alternative time encodings for NeMo-map, again using the ATC dataset and comparing average negative log-likelihood (NLL) where lower indicates better accuracy. We report mean $\pm$ standard deviation, together with the reduction in NLL relative to our method and the corresponding 95% confidence interval (CI). All models are trained using the Adam optimizer with learning rate 10^{-3} for 100 epochs.

Method	NLL↓	NLL reduction (vs Ours)	95% CI
Ours	**0.775 ± 2.052**	–	–
Temporal grid	0.857 ± 2.113	+0.082	[0.081, 0.083]
Fourier features	0.838 ± 2.105	+0.063	[0.062, 0.064]

5 Conclusions

We introduced the first-of-its-kind *continuous spatio-temporal map of dynamics* representation NeMo-map, a novel formulation of MoDs using implicit neural representations. In contrast to prior discretized methods such as CLiFF-map and STeF-map, our approach parametrizes a continuous neural function, which outputs the parameters of a Semi-Wrapped Gaussian Mixture Model at arbitrary spatio-temporal coordinates. The model enables smooth generalization across space and time, and provides a compact representation that avoids storing per-cell distributions.

Through experiments on the large-scale ATC dataset, we demonstrated that NeMo-map achieves subsantially higher accuracy (lower negative log-likelihood) than existing MoD baselines, while reducing map building time. Qualitative results further show that the learned flow fields capture multimodality, temporal variations, and environment topology without requiring explicit maps. Ablation studies confirmed the advantage of using SIREN-based temporal encoding over discrete or Fourier alternatives.

In summary, the results highlight continuous MoDs as a practical and scalable tool for modeling human motion dynamics. By combining accuracy, efficiency, and flexibility, the representation offers a powerful prior for downstream tasks such as socially aware navigation, long-term motion prediction, and localization in dynamic environments. In future work, we plan to extend this formulation with online update mechanisms to adapt continuously to evolving crowd behaviors, further bridging the gap toward long-term real-world deployment.

6 ETHICS STATEMENT

The dataset used for training are publicly available and fully anonymized, representing persons only as 2D positions without identifiers or visual data. Maps of dynamics further aggregate these trajectories into statistical motion patterns, so no personal information are retained.

7 REPRODUCIBILITY STATEMENT

For reproducibility, we provide the full training and evaluation code together with detailed instructions in the supplementary material. The package includes our main model as well as the variants used in the ablation study. Evaluation results with per-sample NLL values are also attached. In addition, we provide scripts for generating and visualizing maps of dynamics (MoDs), as shown in Fig. 5.

REFERENCES

A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social LSTM: Human trajectory prediction in crowded spaces. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, pp. 961–971, 2016.

Victor Hernandez Bennetts, Tomasz Piotr Kucner, Erik Schaffernicht, Patrick P. Neumann, Han Fan, and Achim J. Lilienthal. Probabilistic air flow modelling using turbulent and laminar characteristics for ground and aerial robots. *IEEE Robotics and Automation Letters*, 2(2):1117–1123, 2017.

M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun. Learning motion patterns of people for compliant robot motion. *Int. J. of Robotics Research*, 24(1):31–48, 2005.

D. Brščić, T. Kanda, T. Ikeda, and T. Miyashita. Person tracking in large public spaces using 3-d range sensors. *IEEE Trans. on Human-Machine Systems*, 43(6):522–534, 2013.

Olivier Cappé and Eric Moulines. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):593–613, 2009.

Y. F. Chen, M. Liu, and J. P. How. Augmented dictionary learning for motion prediction. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 2527–2534, 2016.

Yizong Cheng. Mean shift, mode seeking, and clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1995.

Tiago Rodrigues de Almeida, Yufei Zhu, Andrey Rudenko, Tomasz P. Kucner, Johannes A. Stork, Martin Magnusson, and Achim J. Lilienthal. Trajectory prediction for heterogeneous agents: A performance analysis on small and imbalanced datasets. *IEEE Robotics and Automation Letters*, pp. 1–8, 2024.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–38, 1977.

Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *Proc. of the IEEE Conf. on Comp. Vis. and Pat. Rec. (CVPR)*, 2022.

T. Krajník, J. P. Fentanes, J. M. Santos, and T. Duckett. Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Trans. on Robotics (TRO)*, 33 (4):964–977, 2017.

T. P. Kucner, M. Magnusson, E. Schaffernicht, V. H. Bennetts, and A. J. Lilienthal. Enabling flow awareness for mobile robots in partially observable environments. *IEEE Robotics and Automation Letters*, 2(2):1093–1100, 2017.

K. V. Mardia and P. E. Jupp. *Directional Statistics*. Wiley, 2008.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proc. of the Europ. Conf. on Comp. Vision (ECCV)*, 2020.

Sergi Molina, Grzegorz Cielniak, and Tom Duckett. Robotic exploration for learning human motion patterns. *IEEE Trans. on Robotics and Automation (TRO)*, 2022.

Luigi Palmieri, Tomasz P Kucner, Martin Magnusson, Achim J Lilienthal, and K. O. Arras. Kinodynamic motion planning on gaussian mixture fields. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, pp. 6176–6181. IEEE, 2017.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *Proc. of the AAAI Conf. on Artificial Intelligence (AAAI)*, 2018.

Anandarup Roy, Swapan K. Parui, and Utpal Roy. A mixture model of circular-linear distributions for color image segmentation. *International Journal of Computer Applications*, 58(9):6–11, 11 2012. ISSN 0975-8887.

Anandarup Roy, Swapan K. Parui, and Utpal Roy. SWGMM: a semi-wrapped gaussian mixture model for clustering of circular-linear data. *Pattern Anal. Appl.*, 19(3):631–645, 2016.

Liushuai Shi, Le Wang, Sanping Zhou, and Gang Hua. Trajectory unified transformer for pedestrian trajectory prediction. In *Proc. of the IEEE Int. Conf. on Computer Vision (ICCV)*, pp. 9675–9684, October 2023.

Vincent Sitzmann, Julien N.P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *Advances in Neural Inf. Proc. Syst. (NeurIPS)*, 2020.

Chittaranjan Srinivas Swaminathan, Tomasz Piotr Kucner, Martin Magnusson, Luigi Palmieri, Sergi Molina, Anna Mannucci, Federico Pecora, and Achim J. Lilienthal. Benchmarking the utility of maps of dynamics for human-aware motion planning. *Frontiers in Robotics and AI*, 9, 2022. ISSN 2296-9144.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Inf. Proc. Syst. (NeurIPS)*, 2020.

Zhan Wang, Patric Jensfelt, and John Folkesson. Modeling spatial-temporal dynamics of human movements for predicting future trajectories. In *Workshop Proc. of the AAAI Conf. on Artificial Intelligence "Knowledge, Skill, and Behavior Transfer in Autonomous Robots"*, 2015.

Zhan Wang, Patric Jensfelt, and John Folkesson. Building a human behavior map from local observations. In *Proc. of the IEEE Int. Symp. on Robot and Human Interactive Comm. (RO-MAN)*, pp. 64–70, 2016.

W. Zhi, R. Senanayake, L. Ott, and F. Ramos. Spatiotemporal learning of directional uncertainty in urban environments with kernel recurrent mixture density networks. *IEEE Robotics and Automation Letters*, 4(4):4306–4313, 2019.

Yufei Zhu, Andrey Rudenko, Tomasz P. Kucner, Luigi Palmieri, Kai O. Arras, Achim J. Lilienthal, and Martin Magnusson. CLiFF-LHMP: Using spatial dynamics patterns for long-term human motion prediction. In *Proc. of the IEEE Int. Conf. on Intell. Robots and Syst. (IROS)*, 2023.

Yufei Zhu, Andrey Rudenko, Luigi Palmieri, Lukas Heuer, Achim J. Lilienthal, and Martin Magnusson. Fast online learning of cliff-maps in changing environments. In *Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2025.

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

Parts of this manuscript were edited with the assistance of LLMs to improve grammar and clarity. All scientific content was written and verified by the authors.