LEARNING SPATIO-TEMPORAL CONTINUOUS PRESENCE AND MOTION FIELDS FROM ONBOARD VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Mobile robots operating in crowded environments depend on an accurate model of human motion for efficient and safe planning and navigation. Maps of dynamics (MoDs) model motion as a persistent property of the environment, representing either its direction or its presence. We introduce PreMo-map, an MoD conditioned on what the robot observes while patrolling, estimating both the rate at which pedestrians are detected at a location and the distribution of their velocities there. Both quantities are evaluated at continuous coordinates and learned from individual detections rather than from a precomputed grid, and positions inside and beyond the field of view are estimated by separate heads. Measured on a large-scale real-world dataset, PreMo-map outperforms state-of-the-art MoDs within the field of view and remains competitive over the rest of the environment, while also returning a rate that published MoDs do not provide.

1 INTRODUCTION

Long-term deployment in human-populated environments requires anticipating how people move through a space at a given time (Singamaneni et al., 2024). Maps of dynamics (MoDs) address this requirement by modeling motion as a persistent property of the environment rather than of the agents that pass through it (Kucner et al., 2023). Existing MoDs divide into two complementary classes according to the quantity they predict: directional MoDs model the distribution of velocities at a location, its flow (Kucner et al., 2017), and non-directional MoDs the presence of motion there (Jovan et al., 2016). The joint estimation of presence and direction of motion remains underexplored in the literature.

Independently of the quantity modelled, both classes are fitted from the observations accumulated at each location and queried by position (Kucner et al., 2013) and time (Molina et al., 2021). The locations are in most cases the cells of a partition fixed, which sets the resolution of the estimate in advance. However a continuous field over spatio-temporal coordinates removes that discretization (Zhu et al., 2026). We consider a mobile robot navigating a given environment, with an onboard camera providing a limited field of view (FoV) only, therefore only having partial observations on the human motion around it. The map is nonetheless required over the whole environment, since it serves to describe motion at the locations the robot does not currently observe. Estimating it from an onboard observation therefore raises two questions: whether the observation improves the estimate inside the FoV, and whether an estimate produced in this way remains accurate over the rest of the environment. We therefore fit and evaluate the two regions separately.

We introduce PreMo-map, which learns to estimate the rate and the flow at a location jointly, where existing MoDs provide one or the other. We do so by conditioning the estimate on short video clips the robot acquires from an egocentric perspective during operation and assuming a limited FoV of the environment. Given a short clip, the robot pose, and a time, the model estimates two quantities at any position, evaluated at continuous coordinates and without a spatial partition. The first is the rate at which detections occur, from which the probability of presence follows. The second is the flow, represented as a semi-wrapped Gaussian mixture whose components are fixed heading bands shared by every query. Both are learned directly from individual detections under the likelihood of a point process marked by velocity, so no aggregated map is constructed as a supervision target and the two quantities are estimated on a common scale. A frozen self-supervised video encoder (Assran et al., 2025) encodes the clip, and a transformer fuses it with the pose and the time. Positions inside and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

beyond the FoV are estimated by separate heads, each trained on the detections falling in its own region.

Experiments on a large-scale real-world pedestrian dataset, with the onboard video rendered from the recorded trajectories, show that within the FoV PreMo-map estimates both flow and presence more accurately than existing methods, and that beyond it the estimate remains competitive.

We make the following contributions:

- **Conditioning a map of dynamics on onboard video.** We show that conditioning a map of dynamics on a short clip from the robot's onboard camera, together with its pose, improves online estimates of pedestrian flow inside the camera's field of view over existing frameworks.

- **Joint estimation of rate and flow.** In contrast to existing methods, we use a single point-process likelihood over individual detections, to estimate jointly, for the first time, the rate of motion at a location alongside its flow.

2 Related Work

Early MoDs model the occurrence of motion at a location, recorded as a change in occupancy (Wang et al., 2015) or as routes extracted from observed paths (Bennewitz et al., 2005; Chen et al., 2016). Directional MoDs model the distribution of velocities, represented at a cell as a semi-wrapped Gaussian mixture over speed and heading (Kucner et al., 2017), a density on the cylinder that carries the several directions of travel a location sustains (Roy et al., 2016), fitted from individual velocity measurements without complete trajectories or a stored observation history. Related constructions model angular uncertainty alone (Senanayake & Ramos, 2018) or the dependence between adjacent cells (Kucner et al., 2013). Non-directional MoDs model the amount of traffic a location carries, as the rate of a counting process over the detections at a cell (Jovan et al., 2016).

Spectral representations recover periodic structure from sparse, irregular observations (Krajnik et al., 2014; Krajník et al., 2017), giving a time-conditioned directional map when applied to per-cell orientation histograms (Molina et al., 2021), and wrapped-time constructions extend the basis to multi-day recurrence (Vintr et al., 2019). Learned temporal models relax the periodic assumption, either carrying a latent motion pattern forward through a learned transition (Shi & Kucner, 2025a; Shi et al., 2025) or adding a recurrent term to the estimate (Yan et al., 2025; Zhi et al., 2019). A per-cell mixture is instead updated incrementally as observations arrive, so the map reflects the most recent observations and takes no time argument (Zhu et al., 2025b). A continuous alternative to the grid learns an implicit neural field from a spatio-temporal coordinate to the parameters of a semi-wrapped mixture, making the distribution available at any position and any time (Zhu et al., 2026).

Where a location has been observed rarely, a Gaussian process infers its field by interpolating the observations a mobile robot gathers around it (Stuede & Schappler, 2022), and where even those are absent, the static geometry of the environment serves as the prior, a floor plan predicting people flow across environments (Verdoja et al., 2024) or the occupancy distribution of walking humans from a building map alone (Doellinger et al., 2018). In all of this work the query takes a position and a time, and no observation of the moment enters it. PreMo-map is conditioned on the clip and the robot pose at the moment of the query, and estimates the rate jointly with the flow.

Trajectory forecasting propagates the state of an identified agent forward from its observed history, from recurrent predictors to latent-variable and diffusion models (Alahi et al., 2016; Salzmann et al., 2020; Gu et al., 2022), so a forecast is defined only where an agent is tracked, while a map of dynamics is defined at every location, whether or not an agent occupies it. Macroscopic variants predict crowd density and velocity from sensing that covers the whole environment (Kiss et al., 2021; Dugas et al., 2022). The two combine when a map supplies the prior over the velocities a forecaster samples (Zhu et al., 2025a; Shi & Kucner, 2025b).

3 METHOD

3.1 MODELLING SPATIO-TEMPORAL DYNAMICS

We consider the long-term deployment of a robot equipped with an onboard camera in a human-populated environment. At time t the robot collects a *clip* of the n most recent frames $\mathcal{C}_t = \big(\mathbf{I}_{t-(n-1)\delta}, \ldots, \mathbf{I}_t\big)$, acquired at interval δ, and a pose $\mathbf{r}_t = (\mathbf{c}_t, \alpha_t) \in \mathbb{R}^2 \times [0, 2\pi)$. We assume a stream of detections $\mathcal{Z} = \big\{(t_m, \mathbf{x}_m, \mathbf{v}_m)\big\}_m$ over the whole environment for supervision, where m indexes the detections, $\mathbf{x}_m \in \mathbb{R}^2$ is the position of an agent on the ground plane at time t_m and $\mathbf{v}_m \in \mathbb{R}^2$ its velocity. Each detection defines a heading $\theta_m = \operatorname{atan2}(v_{m,y}, v_{m,x}) \in [0, 2\pi)$ and a speed $\rho_m = \|\mathbf{v}_m\| \in \mathbb{R}_{>0}$. Once fitted, the estimate is computed from the clip, the robot pose and the time alone.

Rate and Flow Estimation We model the configuration of the agents at a time t and an arbitrary position $\mathbf{x} \in \mathbb{R}^2$ as a point process marked by heading θ and speed ρ (Daley & Vere-Jones, 2003), with intensity $\Lambda(\mathbf{x}, \theta, \rho, t) \geq 0$. Marginalizing the intensity over heading and speed yields the *rate* $\lambda(\mathbf{x}, t)$, and division by the rate defines the *flow distribution* $p(\theta, \rho \mid \mathbf{x}, t)$, the distribution of the velocity of an agent at that position and time,

$$\lambda(\mathbf{x}, t) = \int_0^{2\pi} \int_0^{\infty} \Lambda \, d\rho \, d\theta, \qquad p(\theta, \rho \mid \mathbf{x}, t) = \frac{\Lambda(\mathbf{x}, \theta, \rho, t)}{\lambda(\mathbf{x}, t)}. \tag{1}$$

The flow distribution is a density on the cylinder $S^1 \times \mathbb{R}_{>0}$, defined wherever $\lambda(\mathbf{x}, t) > 0$. The rate is expressed in detections per unit area per unit time, and its integral over a region and an interval is an expected count. The probability that a region of area A contains at least one detection over an interval Δt follows,

$$p_{\text{present}} = 1 - \exp(-\lambda A \Delta t). \tag{2}$$

Flow Dynamics Model We model the flow distribution with a semi-wrapped Gaussian mixture over the polar decomposition of a velocity (Roy et al., 2016),

$$\hat{p}(\theta, \rho) = \sum_{k=1}^{K} \pi^{(k)} \, G^{(k)}(\theta) \, L^{(k)}(\rho), \qquad G^{(k)}(\theta) = \sum_{w=-W}^{W} \mathcal{N}\big(\theta + 2\pi w; \, \mu_\theta^{(k)}, \, \sigma_\theta^2\big). \tag{3}$$

The weights $\pi^{(k)} \geq 0$ sum to one over the K components, and the angular factor $G^{(k)}$ wraps a normal density over $W = 1$ winding either side, which is accurate for $\sigma_\theta \ll 2\pi$ (Mardia & Jupp, 2008). The radial factor $L^{(k)}$ is a normal density on speed with mean $\mu_\rho^{(k)}$ and standard deviation $\sigma_\rho^{(k)}$, left untruncated below zero, where the speeds of interest place negligible mass.

We fix the $K = 8$ component headings and their spread,

$$\mu_\theta^{(k)} = \frac{2\pi(k-1)}{K}, \quad k = 1, \ldots, K, \qquad \sigma_\theta = \pi / K. \tag{4}$$

Component k then denotes the same heading band at every query, making the weight vector $\boldsymbol{\pi}$ comparable across positions and times. We define a spread of half the spacing, so adjacent components overlap and a heading falling between two of them receives mass from each. Each component has its own speed mean and standard deviation, so the model describes a different speed distribution for each direction of travel.

3.2 ARCHITECTURE

Similarly to NeMo-map (Zhu et al., 2026), PreMo-map represents the map of dynamics as a continuous field that can be queried at any position, instead of a grid of per-cell estimates like prior work. However, while NeMo-map's field depends only on position and time, PreMo-map conditions its learning on short video clips the robot acquires from an egocentric perspective during operation and with limited FoV. Unlike existing methods, we jointly estimate the rate at each position alongside the flow distribution. Figure 1 shows the architecture.

PreMo-map introduces two key elements that enable this joint prediction:

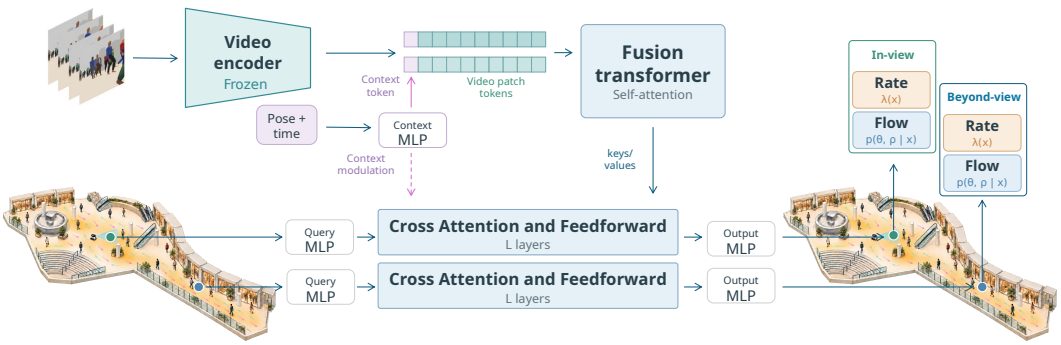

Figure 1: **Overview of PreMo-map.** Onboard video, robot pose and time condition continuous estimates of pedestrian detection rate $\lambda(\mathbf{x})$ and flow distribution $p(\theta, \rho \mid \mathbf{x})$, with separate readouts for positions inside and beyond the camera's FoV.

Encoding and Fusion First, we encode the clip with a frozen V-JEPA 2.1 video encoder (Mur-Labadia et al., 2026) into P patch tokens $\mathbf{F} \in \mathbb{R}^{P \times d}$, to each of which a learned positional embedding is added. The robot position $\mathbf{c}_t$, its heading, encoded as $(\sin \alpha_t, \cos \alpha_t)$, and the time of day, encoded as J sine and cosine harmonics of the phase $\phi = (t \bmod T)/T$, of period T, are concatenated and mapped by a two-layer perceptron to a single context token. The context token and the patch tokens are passed through a transformer encoder, giving a fused sequence $\mathbf{H} \in \mathbb{R}^{(1+P) \times d}$.

Query Readout Second, a query readout evaluates the field at an arbitrary position $\mathbf{x}$ against $\mathbf{H}$. Since only $2.7\,\%$ of the training detections fall within the FoV, a single readout would be dominated by queries beyond it. We therefore instantiate it in four parallel branches over the same $\mathbf{H}$, separating the flow distribution from the rate and the positions inside the FoV from those beyond it. Each query is routed geometrically, and each branch is trained only on the queries routed to it. Within a branch, a *query MLP* embeds the position, expressed both in world coordinates, normalized over the extent of the environment, and in the robot frame. The embedding $\mathbf{q}$ then passes through L layers, each cross-attending to $\mathbf{H}$ and applying a feedforward block. An *output MLP* maps $\mathbf{q}$ to the branch's estimate at $\mathbf{x}$. For a flow branch, the K mixture weights, normalized by a softmax, the K speed means, and the K speed spreads through an exponential. For a rate branch, $\lambda(\mathbf{x})$ also through an exponential (Bishop, 1994). Every readout layer is further conditioned on the context by feature-wise modulation (Perez et al., 2018; Rebain et al., 2023), producing a scale, a shift and a residual gate for the attention half and for the feedforward half of each layer. Writing (γ, β, g) for one such triple, the attention half of layer i computes

$$\mathbf{q} \leftarrow \mathbf{q} + g_i, \mathrm{attn}\big((1+\gamma_i), \mathrm{LN}(\mathbf{q}) + \beta_i, ; \mathbf{H}\big), \tag{5}$$

and the feedforward half likewise with its own triple. The normalization layers carry no affine parameters, and the layer producing the triples is initialized at zero, so each readout layer is initially the identity (Peebles & Xie, 2023).

We set $L = 4$, following V-JEPA 2's attentive-probe protocol (Assran et al., 2025; Cudlenco et al., 2026). Every attention block, in the fusion stage and in the readout alike, normalizes each head's queries and keys before their inner product, which bounds the attention logits (Zhai et al., 2023; Wortsman et al., 2024).

The environment is far larger than the camera's sector, so a single readout trained on both regions is dominated by the queries outside the FoV. We instantiate the readout in four branches over the same fused sequence, separating the flow distribution from the rate and the queries inside the camera's FoV from those outside it. The routing is geometric, and each branch is trained only on the queries routed to it.

3.3 Likelihood and Training

PreMo-map is supervised directly by the stream of individual detections. The rate and the flow distribution are therefore fitted jointly under a single objective, and a two-stage curriculum first grounds the estimate in the FoV before extending it to the whole environment.

Likelihood Given a clip ending at time t, we denote by $\mathcal{R} \subset \mathbb{R}^2$ a region of the environment and by $\mathcal{S} = [t, t + \Delta t)$ the interval beginning where the clip ends. The rate is a density in time, so the likelihood of the point process of (1) is taken over an interval of duration Δt rather than at an instant. For a marked Poisson process of intensity $\Lambda = \lambda\, p$, the log-likelihood of the detections in $\mathcal{R} \times \mathcal{S}$ is (Daley & Vere-Jones, 2003)

$$\ell = \sum_{\{m\,:\,(\mathbf{x}_m, t_m) \in \mathcal{R}\times\mathcal{S}\}} \Big[\log \hat{\lambda}(\mathbf{x}_m) + \log \hat{p}(\theta_m, \rho_m \mid \mathbf{x}_m)\Big] - \int_{\mathcal{S}}\int_{\mathcal{R}} \hat{\lambda}(\mathbf{x})\, d\mathbf{x}\, d\tau, \tag{6}$$

with $\hat{\lambda}$ and $\hat{p}$ the network's estimates. Fitting maximizes ℓ, or equivalently minimizes $-\ell$, which separates additively into a flow term and a rate term,

$$\mathcal{L}_{\text{flow}} = -\sum_m \log \hat{p}(\theta_m, \rho_m \mid \mathbf{x}_m), \qquad \mathcal{L}_{\text{rate}} = \int_{\mathcal{S}}\int_{\mathcal{R}} \hat{\lambda}\, d\mathbf{x}\, d\tau \;-\; \sum_m \log \hat{\lambda}(\mathbf{x}_m), \tag{7}$$

the flow term the density assigned to the observed velocities, and the rate term the expected count over $\mathcal{R} \times \mathcal{S}$ less the log rate at the detections, so that it rises with the rate at positions where nothing was detected.

The integral term of $\mathcal{L}_{\text{rate}}$ is evaluated by quadrature on a regular grid of spacing g over $\mathcal{R}$, with the rate held constant on each cell,

$$\int_{\mathcal{S}}\int_{\mathcal{R}} \hat{\lambda}(\mathbf{x})\, d\mathbf{x}\, d\tau \;\approx\; \Delta t \sum_i \hat{\lambda}(\mathbf{x}_i)\, a_i, \tag{8}$$

with $a_i = g^2$ the cell area. The spacing g is a quadrature parameter, independent of the spacing at which the field is queried. Each clip evaluates g_{n} cells of the grid, traversed cyclically and weighted by the measure $\Delta t, |\mathcal{R}|/g_{\text{n}}$ that subset represents, so one pass over the data reproduces the integral over $\mathcal{R}$. Each interval retains at most M_{kept} of its M detections, drawn at random and weighted by M/M_{kept}.

Training A detection is assigned to the latest clip whose interval $\mathcal{S}$ contains it, so no clip is fitted against a detection it observed, and a detection falling more than Δt after the last clip is discarded. Each training example is one detection: the position where the readout is queried and the heading and speed the flow term is evaluated on. The encoder is frozen throughout, and the remaining parameters are fitted as a curriculum in two stages. The first stage supervises the FoV alone and withholds every input except the clip, the robot heading and the query's position in the robot frame. The second stage restores the world position and the time of day, extends the supervision to the whole environment with the fusion stage frozen, and splits the readout into the four branches of Section 3.2, each initialized from the readout of the first stage.

4 Experiments

4.1 Dataset and Evaluation Protocol

Dataset Unlike MoDs fitted per location from recorded trajectories, PreMo-map learns the field from egocentric observations, requiring a dataset that simultaneously offers (i) a *moving* sensing platform, (ii) egocentric RGB data, (iii) annotated or recoverable pedestrian trajectories over the whole environment, and (iv) multi-session temporal structure (repeated visits and time-of-day variation). No existing dataset offers all four. We therefore build a *grounded simulation* on ATC (Brščić et al., 2013), a large-scale pedestrian tracking dataset spanning multiple days. The site's own localization grid is extruded into a Gazebo world and the recorded tracks are replayed through it at $1\,\text{Hz}$; a platform with a limited FoV is then driven along a serpentine coverage loop, rendering an egocentric clip of 8 frames at $3\,\text{Hz}$ and recording its pose at every step. Only the observer is simulated, whereas the trajectories, their timing and the detection stream used for supervision and evaluation are the dataset's own, and pedestrian motion is inherited from the recording rather than modelled. The resulting dataset pairs over $62\,000$ egocentric clips across the four recorded days with the robot pose and the pedestrian detections over the whole environment. We release it together with the simulation upon acceptance, providing a benchmark for learning MoDs from onboard observations.

Evaluation Protocol We adopt the evaluation protocol of NeMo-map (Zhu et al., 2026), training on one day and testing on three held-out days. While the training loss in (7) incorporates terms for both presence and flow, we drive model selection by the flow term alone. This is because the direction of motion is an inherent property of the environment, largely shared across days, whereas the presence fluctuates over the days, depending on both place and time.

We measure the density each estimator assigns to the observed velocity, and the probability it assigns to the observed presence event, both as a mean negative log likelihood in nats, a standard measure of a predicted distribution against an observation (Gneiting & Raftery, 2007). The **flow likelihood** $\mathcal{L}_{\text{flow}}$ is computed per detection, alongside those of its heading and speed marginals, $\mathcal{L}_\theta$ and $\mathcal{L}_\rho$. We also report the Mean Absolute Circular Error (MAE) of the heading and the speed error as diagnostics. The heading compared is the circular mean of the mixture, $\operatorname{atan2}\big(\sum_k \pi^{(k)} \sin \mu_\theta^{(k)}, \sum_k \pi^{(k)} \cos \mu_\theta^{(k)}\big)$, and the speed the weighted mean of the component means, $\sum_k \pi^{(k)} \mu_\rho^{(k)}$. **Presence** is the event that a grid cell contains at least one detection over the window $[t, t + \Delta t)$, and its probability is defined by (2) with A the cell's area. We set $\Delta t = 1\,\text{s}$ and a cell spacing of $2\,\text{m}$ throughout. We further decompose the presence log loss $\mathcal{L}_{\text{pres}}$ into reliability and resolution (Murphy, 1973).

All quantities are reported over the two regions the FoV separates, in and beyond view. $97.7\,\%$ of the evaluated detections lie beyond it, the *whole-environment* results are the mean of the two weighted by that fraction and is not reported separately.

4.2 Baselines

We compare PreMo-map's flow distribution against four state-of-the-art methods.

- **STeF-map** (Molina et al., 2021) accumulates per-cell harmonic models over eight discretized orientations and carries no speed.
- **CLiFF-map** (Kucner et al., 2017) represents flow with a semi-wrapped Gaussian mixture (SWGMM) over heading and speed per cell and carries no time argument.
- **Online CLiFF-map** (Zhu et al., 2025b) updates the previous representation incrementally with incoming observations, yielding recency rather than a queried time.
- **NeMo-map** (Zhu et al., 2026) represents the SWGMM as a continuous field over position and time of day.

Since no available MoD estimates a rate, we compare against two further baselines:

- **Constant base rate** corresponds to the site's occupancy at every cell. Fitted on the cells it is scored on, it carries neither resolution nor calibration error.
- **FreMEn** (Krajník et al., 2017), the same spectral model over a cell's occupied fraction that STeF-map carries over orientations.

4.3 Results

Flow Prediction. Table 1 reports every baseline on the three held-out evaluation days. Inside the FoV PreMo-map achieves the best result on all likelihoods (with $1.450/1.098/0.329$ nats against $1.592/1.306/0.338$ for NeMo-map), showing the added value of the visual information to identify the heading. Beyond the FoV, NeMo-map's $1.848/1.477$ outperforms PreMo-map's $1.959/1.595$ on both $\mathcal{L}_{\text{flow}}$ and $\mathcal{L}_\theta$ with the roles being reversed on speed $\mathcal{L}_\rho$ (0.417 against 0.394).

Presence Estimation. Inside the FoV PreMo-map outperforms both baseline with an $\mathcal{L}_{\text{pres}}$ of 0.1630 against 0.2238 for FreMEn and 0.2541 for the constant, and 22.8×10^{-3} of resolution against FreMEn's 4.0×10^{-3}: the observation, and not the identity of the cell, is what separates a busy cell from a quiet one there. Beyond the FoV the two are on the same level, 0.0481 against 0.0486, so where the robot has no observation a per-cell history is equally informative. FreMEn reduces here to a constant occupancy per cell: at $\Delta t = 1\,\text{s}$ a cell is occupied in only $1.38\,\%$ of windows, too sparse for a single training day to reveal any periodicity, so its model selection retains no periodic component.

Table 1: **Flow prediction** on ATC (Brščić et al., 2013), three evaluation days pooled. Negative log likelihood ↓ of an observed velocity per channel: the joint ($\mathcal{L}_{\text{flow}}$), heading ($\mathcal{L}_\theta$) and speed ($\mathcal{L}_\rho$). Heading-only baselines carry no speed model, so their speed and joint channels are not applicable (—). Lower is better.

	In view			Beyond view		
Method	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_\theta$	$\mathcal{L}_\rho$	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_\theta$	$\mathcal{L}_\rho$
Uniform flow (reference)	2.281	1.838	0.443	2.241	1.838	0.403
STeF-map (Molina et al., 2021)	—	1.565	—	—	1.631	—
CLiFF-map (Kucner et al., 2017)	2.249	1.698	0.521	2.773	1.987	0.764
Online CLiFF-map (Zhu et al., 2025b)	1.915	1.507	0.408	2.216	1.697	0.517
NeMo-map (Zhu et al., 2026)	1.592	1.306	0.338	**1.848**	**1.477**	0.417
PreMo-map *(ours)*	**1.450**	**1.098**	**0.329**	1.959	1.595	**0.394**

Table 2: **Presence estimation** on ATC (Brščić et al., 2013): $\mathcal{L}_{\text{pres}}$ over the cells and windows of the three evaluation days, with *reliability* (Rel), *resolution* (Res) (in units of 10^{-3}) and *expected calibration error* (ECE). — defines *not applicable*.

Model	Cells	Occup.	$\mathcal{L}_{\text{pres}}$ ↓	Rel↓	Res↑	ECE↓
In view						
Constant base rate	129 845	7.02 %	0.2541	0.0	0.0	—
FreMEn (Krajník et al., 2017) occupancy	129 845	7.02 %	0.2238	1.6	4.0	0.033
PreMo-map *(ours)*	129 845	7.02 %	**0.1630**	2.0	**22.8**	0.025
Beyond view						
Constant base rate	28 594 191	1.35 %	0.0716	0.0	0.0	—
FreMEn (Krajník et al., 2017) occupancy	28 594 191	1.35 %	**0.0481**	0.2	1.0	0.007
PreMo-map *(ours)*	28 594 191	1.35 %	0.0486	0.2	**1.1**	0.007

PreMo-map outperforms all baselines within the FoV on both flow and presence, and beyond it is outperformed only by NeMo-map on flow and matches FreMEn on presence, while being the only method jointly estimating both. Its improvement against the state-of-the-art within the FoV exceeds its deficit beyond it: the heading likelihood 0.208 nats (16 %) lower in view and 0.118 (8 %) higher beyond, and the joint flow likelihood 0.142 nats (9 %) lower against 0.111 nats (6 %) higher.

4.4 Ablations

To identify what each input contributes to the estimate, we conduct two sets of ablations. The first withholds the clip, the position and the clock, in turn and in combination, and retrains the model, which measures the information an input carries beyond the remaining ones. The second replaces the clip supplied to the trained model with one fixed clip reused at every step, an empty corridor or a crowd, which measures the model's dependence on the video at inference.

Input Components The full conditioning achieves the lowest heading likelihood in both regions, 1.196 nats in view and 1.656 beyond it, and the lowest joint likelihood in view, 1.638. No other conditioning achieves both: the clip and the position reach 1.199 in view and 1.667 beyond it, the clip and the clock 1.208 and 1.731, and the clip alone 1.206 and 1.718. Removing any one input degrades the estimate in one region or one channel. Removing the clock raises the in-view speed likelihood from 0.416 to 0.455 nats, and removing the position raises the beyond-view heading likelihood from 1.656 to 1.731.

Removing the clip has the largest effect: in view it raises the joint likelihood by 0.319 nats, the heading likelihood by 0.366 and the presence log loss by 0.077; it also increases the heading error from 43.6° to 82.8° against the 90° of chance. The heading inside the FoV is therefore better recovered from video observations than from position and time of day alone. Beyond the FoV the same removal changes the heading likelihood by 0.018 nats, and the heading error there lies between 78.7° and 82.6° for every conditioning, so the benefit of the clip is concentrated in view.

Table 3: **Ablation study over input components** on the first held-out day. Negative log likelihood of an observed velocity per channel: the joint ($\mathcal{L}_{\text{flow}}$), heading ($\mathcal{L}_\theta$) and speed ($\mathcal{L}_\rho$); Mean Absolute circular Error (MAE) of the predicted heading; $\mathcal{L}_{\text{pres}}$ is the presence log loss (2). Lower is better.

	In view					Beyond view				
Conditioned on	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_\theta$	$\mathcal{L}_\rho$	$\mathcal{L}_{\text{pres}}$	MAE	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_\theta$	$\mathcal{L}_\rho$	$\mathcal{L}_{\text{pres}}$	MAE
Clip, position, clock	**1.638**	**1.196**	0.416	0.228	43.6	2.137	**1.656**	0.510	0.072	**79.1**
Clip, position	1.715	1.199	0.455	0.224	**43.0**	2.150	1.667	0.502	**0.071**	80.1
Clip, clock	1.641	1.208	0.406	**0.215**	43.4	2.166	1.731	0.462	0.078	82.0
Clip only	1.725	1.206	0.446	0.226	43.6	**2.113**	1.718	**0.413**	0.074	82.6
Position, clock	1.957	1.562	**0.385**	0.305	82.8	2.261	1.674	0.613	0.079	78.7

Table 4: **Ablation study over visual information** on the first held-out day. The clip is replaced by one fixed clip reused at every step, an empty corridor or a crowd. Results reported on the three days pooled. A heading MAE of 90° is chance. Lower is better.

	In view				Beyond view			
Clip	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_\theta$	$\mathcal{L}_\rho$	MAE	$\mathcal{L}_{\text{flow}}$	$\mathcal{L}_\theta$	$\mathcal{L}_\rho$	MAE
Camera clip	**1.450**	**1.098**	**0.329**	**40.2**	1.959	1.595	0.394	**77.8**
Fixed, empty	4.087	2.609	1.022	89.5	**1.939**	**1.588**	**0.378**	78.7
Fixed, crowded	3.206	2.398	0.838	89.2	2.007	1.636	0.410	79.6

In view the four conditionings receiving the clip hold within 0.012 nats of heading likelihood of each other, a spread much smaller than the 0.366 nats separating them from the conditioning without it.

Visual Information Replacing the clip supplied to the trained model raises the in-view heading likelihood by 1.300 nats with the crowded clip and 1.511 with the empty one, and raises the heading error to 89.2° and 89.5° against the 90° of chance. The speed channel follows, from 0.329 to 1.022 nats. The in-view estimate is therefore read from the clip acquired at the query.

Beyond the FoV the empty clip matches the camera's on every channel, to within 0.020 nats of likelihood and 0.9° of heading error. The out-of-view estimate is therefore insensitive to which clip is supplied, consistent with those queries being answered by a separate readout.

The two fixed clips separate from each other by 0.211 nats of heading likelihood and 0.881 of joint likelihood in view, while both remain at chance on the heading error. The content of the clip moves the likelihood, and recovering the heading requires the clip acquired at that moment.

4.5 Spatial Reach

We restrict the evaluated detections to a disc of radius R around the robot to quantify the difference between PreMo-map and NeMo-map (Zhu et al., 2026) as a function of distance from the robot (Table 5).

Inside the FoV PreMo-map outperforms Nemo-map at every radius, by 0.331 nats of heading within $2\,\text{m}$, 0.362 within $4\,\text{m}$ and 0.205 within $8\,\text{m}$, the camera's range. Beyond the FoV its heading likelihood is above NeMo-map's, between 0.065 and 0.113 nats, while the speed likelihood is still below it throughout, between 0.037 and 0.068.

The radius selects the population as well as the distance: NeMo-map is conditioned on neither a camera nor a pose, and its own joint likelihood rises with the radius, from 1.687 nats within $2\,\text{m}$ to 2.159 within $32\,\text{m}$, so what changes along the axis includes which pedestrians are averaged.

The advantage of PreMo-map is therefore determined by visibility rather than proximity: within the FoV it improves on NeMo-map's heading likelihood by 0.205 nats ($15\,\%$) even at the $8\,\text{m}$ range of the camera, whereas outside it, within $2\,\text{m}$ of the robot, it is outperformed by 0.102 ($8\,\%$).

Table 5: **Spatial reach of the observation** on the first held-out day. Detections are restricted to a disc of radius R around the robot; in view means inside the camera's wedge.

		$\mathcal{L}_{\text{flow}} \downarrow$			$\mathcal{L}_{\theta} \downarrow$			$\mathcal{L}_{\rho} \downarrow$		
	n	NeMo-map	PreMo-map	Diff.	NeMo-map	PreMo-map	Diff.	NeMo-map	PreMo-map	Diff.
Inside the field of view										
2 m	3927	1.738	1.457	+0.281	1.376	1.045	+0.331	0.418	0.388	+0.031
4 m	15 569	1.708	1.358	+0.350	1.347	0.985	+0.362	0.417	0.350	+0.067
8 m	55 219	1.778	1.638	+0.140	1.402	1.196	+0.205	0.426	0.416	+0.010
Beyond the field of view										
2 m	15 744	1.687	1.769	−0.082	1.325	1.427	−0.102	0.404	0.357	+0.047
4 m	57 974	1.796	1.879	−0.083	1.422	1.527	−0.105	0.415	0.378	+0.037
8 m	190 620	1.900	1.981	−0.081	1.471	1.584	−0.113	0.469	0.429	+0.041
16 m	650 312	2.116	2.130	−0.014	1.563	1.640	−0.077	0.589	0.521	+0.068
32 m	1 505 767	2.159	2.168	−0.009	1.594	1.660	−0.065	0.598	0.539	+0.059
Unrestricted	2 469 894	2.113	2.137	−0.024	1.586	1.656	−0.070	0.561	0.510	+0.051

5 Conclusion

We introduced PreMo-map, a Map of dynamics (MoD) learned from the video a robot acquires during operation. Conditioned on short egocentric clips and the robot pose, it estimates online the flow of pedestrian motion together with the rate at which pedestrians occur. Both are jointly learned from individual detections under a single point-process likelihood, which no previous MoD provides.

Trained and evaluated on a reconstructed grounded simulation of the ATC dataset, within a robot's FoV, PreMo-map estimates both flow and presence more accurately than state-of-the-art MoDs. It improves on NeMo-map's heading likelihood by 0.208 nats ($16\,\%$) and on FreMEn's presence log loss by 0.061 ($27\,\%$). We conduct ablation studies at training and inference confirming that this advantage derives from the clip. Beyond the FoV, where the estimate relies on position and time of day, PreMo-map achieves performance comparable to state-of-the-art methods on both flow and presence.

In future work we will condition the estimate on the observations a robot accumulates along its route rather than on a single clip, extending the region over which the observation is informative.

Ethics Statement

The dataset used here records pedestrian trajectories in a public shopping centre, and the egocentric observations are rendered from those trajectories in simulation, so no recording of people was performed for this work.

Reproducibility Statement

Section A2 give the training procedure and the implementation details, and **??** reports the changes to the shared scoring implementation under which the published estimators are re-evaluated.

References

Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 961–971, 2016.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv:2506.09985*, 2025.

Maren Bennewitz, Wolfram Burgard, Grzegorz Cielniak, and Sebastian Thrun. Learning motion patterns of people for compliant robot motion. *The Int. J. of Robotics Research*, 24(1):31–48, 2005.

Christopher M Bishop. Mixture density networks. 1994.

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.

Y. F. Chen, M. Liu, and J. P. How. Augmented dictionary learning for motion prediction. In *IEEE Int. Conf. on Robot. Autom.*, pp. 2527–2534, 2016.

Nicolae Cudlenco et al. GTASA: Ground truth annotations for spatiotemporal analysis, evaluation and training of video models. In *European Conf. on Computer Vision*, 2026.

D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes, Volume I: Elementary Theory and Methods*. Springer, 2 edition, 2003.

Johannes Doellinger, Markus Spies, and Wolfram Burgard. Predicting occupancy distributions of walking humans with convolutional neural networks. *IEEE Robotics & Automation L.*, 3(3): 1522–1528, 2018.

Daniel Dugas, Kuanqi Cai, Olov Andersson, Nicholas Lawrance, Roland Siegwart, and Jen Jen Chung. Flowbot: Flow-based modeling for robot navigation. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 8799–8805, 2022.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *J. of the American Statistical Association*, 102(477):359–378, 2007.

Tianpei Gu, Guangyi Chen, Junlong Li, Chunze Lin, Yongming Rao, Jie Zhou, and Jiwen Lu. Stochastic trajectory prediction via motion indeterminacy diffusion. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pp. 17113–17122, 2022.

Ferdian Jovan, Jeremy Wyatt, Nick Hawes, and Tomas Krajnik. A poisson-spectral model for modelling temporal patterns in human data observed by a robot. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 4013–4018, 2016.

Stefan H Kiss, Kavindie Katuwandeniya, Alen Alempijevic, and Teresa Vidal-Calleja. Probabilistic dynamic crowd prediction for social navigation. In *IEEE Int. Conf. on Robot. Autom.*, pp. 9269–9275, 2021.

Tomas Krajnik, Jaime Pulido Fentanes, Grzegorz Cielniak, Christian Dondrup, and Tom Duckett. Spectral analysis for long-term robotic mapping. In *IEEE Int. Conf. on Robot. Autom.*, pp. 3706–3711, 2014.

Tomáš Krajník, Jaime P Fentanes, Joao M Santos, and Tom Duckett. Fremen: Frequency map enhancement for long-term mobile robot autonomy in changing environments. *IEEE Trans. on Robotics*, 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 & Automation L.*, 2(2):1093–1100, 2017.

Tomasz Kucner, Jari Saarinen, Martin Magnusson, and Achim J Lilienthal. Conditional transition maps: Learning motion patterns in dynamic environments. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 1196–1201, 2013.

Tomasz Piotr Kucner, Martin Magnusson, Sariah Mghames, Luigi Palmieri, Francesco Verdoja, Chittaranjan Srinivas Swaminathan, Tomáš Krajník, Erik Schaffernicht, Nicola Bellotto, Marc Hanheide, et al. Survey of maps of dynamics for mobile robots. *The Int. J. of Robotics Research*, 42(11):977–1006, 2023.

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

Sergi Molina, Grzegorz Cielniak, and Tom Duckett. Robotic exploration for learning human motion patterns. *IEEE Trans. on Robotics*, 38(2):1304–1318, 2021.

Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv:2603.14482*, 2026.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF Int. Conf. on Computer Vision*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conf. on Artificial Intell.*, 2018.

Daniel Rebain, Mark J. Matthews, Kwang Moo Yi, Gopal Sharma, Dmitry Lagun, and Andrea Tagliasacchi. Attention beats concatenation for conditioning neural fields. *Trans. on Machine Learning Research*, 2023.

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

Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conf. on Computer Vision*, pp. 683–700. Springer, 2020.

Ransalu Senanayake and Fabio Ramos. Directional grid maps: modeling multimodal angular uncertainty in dynamic environments. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 3241–3248, 2018.

Junyi Shi and Tomasz Piotr Kucner. Learning temporal maps of dynamics for mobile robots. *Robotics & Autonomous Syst.*, 184:104853, 2025a.

Junyi Shi and Tomasz Piotr Kucner. Leveraging maps of spatial motion patterns to enhance long-term adaptive trajectory prediction with diffusion models. In *European Conf. on Mobile Robots*, pp. 1–8, 2025b.

Junyi Shi, Qingyun Guo, and Tomasz Piotr Kucner. Event-triggered maps of dynamics: A framework for modeling spatial motion patterns in non-stationary environments. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 12924–12931, 2025.

Phani Teja Singamaneni, Pilar Bachiller-Burgos, Luis J. Manso, Anaís Garrell, Alberto Sanfeliu, Anne Spalanzani, and Rachid Alami. A survey on socially aware robot navigation: Taxonomy and future challenges. *The Int. J. of Robotics Research*, 43(10):1533–1572, 2024. doi: 10.1177/02783649241230562.

Marvin Stuede and Moritz Schappler. Non-parametric modeling of spatio-temporal human activity based on mobile robot observations. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 126–133, 2022.

Francesco Verdoja, Tomasz Piotr Kucner, and Ville Kyrki. Bayesian floor field: Transferring people flow predictions across environments. In *IEEE/RSJ Int. Conf. on Intell. Robots and Syst.*, pp. 12801–12807, 2024.

Tomáš Vintr, Sergi Molina, Ransalu Senanayake, George Broughton, Zhi Yan, Jiří Ulrich, Tomasz Piotr Kucner, Chittaranjan Srinivas Swaminathan, Filip Majer, Mária Stachová, et al. Time-varying pedestrian flow models for service robots. In *European Conf. on Mobile Robots*, pp. 1–7, 2019.

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

Mitchell Wortsman et al. Small-scale proxies for large-scale transformer training instabilities. In *Int. Conf. on Learning Representations*, 2024.

Zhibin Yan, Junyi Shi, and Tomasz Piotr Kucner. STeF-LSTM: A hybrid framework integrating periodic and sequential modeling for mapping motion patterns. In *European Conf. on Mobile Robots*, pp. 1–8, 2025.

Shuangfei Zhai, Tatiana Likhomanenko, Etai Littwin, Dan Busbridge, Jason Ramapuram, Yizhe Zhang, Jiatao Gu, and Josh Susskind. Stabilizing transformer training by preventing attention entropy collapse. In *Int. Conf. on Machine Learning*, 2023.

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 & Automation L.*, 4(4):4306–4313, 2019.

Yufei Zhu, Andrey Rudenko, Tomasz P Kucner, Achim J Lilienthal, and Martin Magnusson. Long-term human motion prediction using spatio-temporal maps of dynamics. *IEEE Robotics & Automation L.*, 2025a.

Yufei Zhu, Andrey Rudenko, Luigi Palmieri, Lukas Heuer, Achim J. Lilienthal, and Martin Magnusson. Fast online learning of CLiFF-maps in changing environments. In *IEEE Int. Conf. on Robot. Autom.*, 2025b.

Yufei Zhu, Shih-Min Yang, Andrey Rudenko, Tomasz P. Kucner, Achim J. Lilienthal, and Martin Magnusson. NeMo-map: Neural implicit flow fields for spatio-temporal motion mapping. In *Int. Conf. on Learning Representations*, 2026.

APPENDIX

A1 GENERATIVE AI DISCLOSURE

In this work, we used generative AI tools for improving grammar and clarity, and for rendering illustrations for the diagram figure. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A2 IMPLEMENTATION DETAILS

The readout parameterizes the flow mixture of (3) and the rate at the queried position. The component headings $\mu_\theta^{(k)}$ and their spread σ_θ are fixed by (4), so the free parameters of a query are the K weights, the K speed means, the K speed standard deviations and the rate. The weights $\pi^{(k)}$ pass through a softmax. The speed means $\mu_\rho^{(k)}$ are unconstrained. Each speed standard deviation $\sigma_\rho^{(k)}$ is an exponential of a raw value clamped to $[-4, 2]$ in the log domain, admitting spreads between 0.018 and $7.4\,\mathrm{m\,s^{-1}}$, and the rate λ is an exponential of a raw value clamped to $[-12, 5]$, between 6×10^{-6} and 148 detections per second per square metre. We set $K = 8$ components, $\sigma_\theta = \pi/8$ and $W = 1$ winding retained either side. The speed mean and spread are initialized at $0.94\,\mathrm{m\,s^{-1}}$ and $e^{-0.92}$, and the rate at the occupancy of the whole environment.

The clip is 8 frames at $3\,\mathrm{Hz}$, encoded by a frozen V-JEPA 2.1 ViT-gigantic at $384\,\mathrm{px}$. A two-layer perceptron maps the robot position, $(\sin\alpha_t, \cos\alpha_t)$ and $J = 2$ sine and cosine harmonics of the phase, at period $T = 24\,\mathrm{h}$ in local time, to a context token of width 64. The fusion stage is a transformer encoder of two layers with eight heads and feedforward width 512, with dropout 0.1 in its blocks and 0.2 on the frozen patch features. The readout is $L = 4$ cross-attention layers of query width 64 followed by a perceptron of hidden width 256, in four branches.

Both stages are trained with AdamW at learning rate 1×10^{-4} and weight decay 1×10^{-4} under a cosine schedule, for 30 epochs each and with no gradient clipping. Clips are shuffled every epoch and batched eight at a time, each carrying its own detections. The model conditioned on position and clock alone cannot be warm-started from the first stage and is trained from scratch for 60 epochs. Each interval $\mathcal{S}$ retains at most 128 of its detections, and the rate integral is evaluated on a grid of spacing $g = 2\,\mathrm{m}$ with $g_\mathrm{n} = 64$ cells per clip over $\Delta t = 1\,\mathrm{s}$. Validation uses $10\,\%$ of the training day,

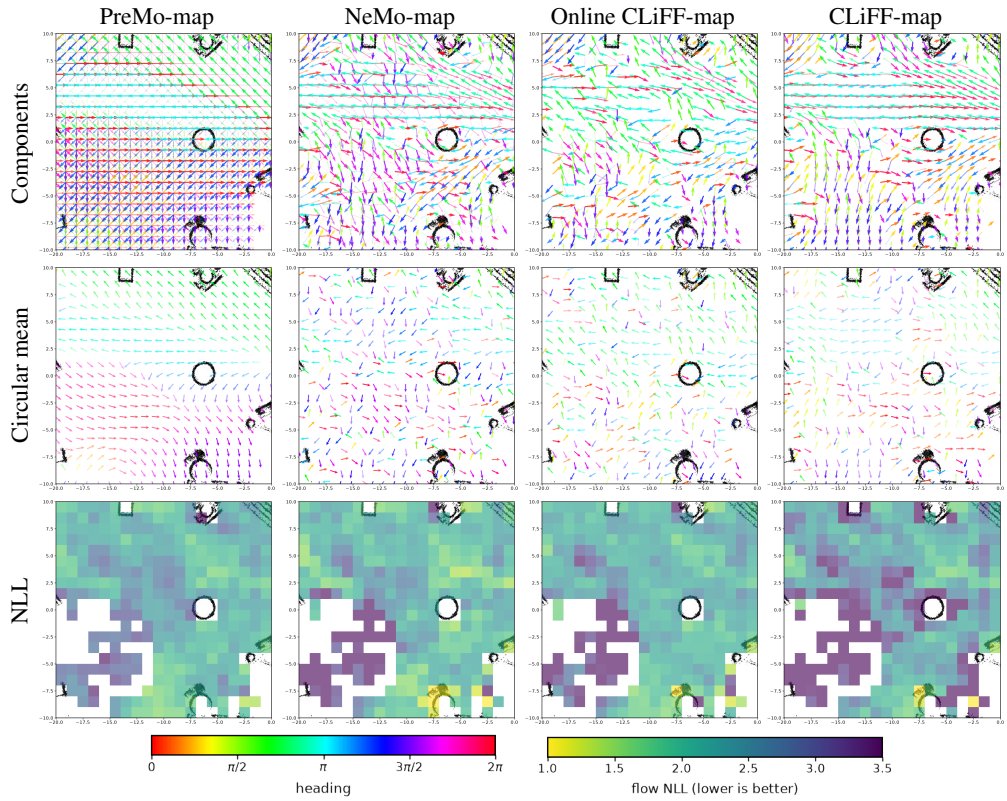

Figure A1: **Crossing flows in the central open area** at 12:30, the region $x \in [-20, 0]$, $y \in [-10, 10]$. **First row:** every mixture component from its cell centre, hue its heading, length its speed and opacity its weight within the cell. **Second row:** one arrow per cell at the weight-weighted circular mean of the heading, at a fixed length, its opacity the resultant length of the cell's heading density. **Third row:** per-cell flow likelihood, cells holding fewer than eight detections left blank.

drawn as 11 contiguous blocks separated from the training clips by a guard of 120 s. Every reported figure is a single run at seed 0.

A3 Evaluation Protocol

We distinguish three cases when evaluating a likelihood: an available but inaccurate estimate, an absent estimate, and a degenerate estimate.

Where an estimate is defined, the negative log likelihood contributes to the mean with an upper bound of 20.72 nats induced by assigning a density of 1×10^{-9} to every estimator. The bound is achieved for 0.03 % of NeMo-map's detections and for at most 4 % among the grid-based estimators.

For a cell absent from training, no estimate is defined. We use a fixed prediction, independent of position and time of $\log 2\pi = 1.838$ nats for heading and a normal density fitted to the speeds observed on the training day. An absent estimate is distinguished from an inaccurate estimate, and the proportion of detections for which an estimate is defined is reported as coverage, which ranges from 97.5 % to 100 % across the three grid-based estimators.

A component whose covariance is not positive definite does not define a valid distribution and is removed, after which the remaining weights in its cell are renormalized. A component with a narrow covariance is retained, since removing it would replace the estimator produced by the method with a repaired estimator. STeF-map returns a probability for each orientation bin rather than a density; this probability is therefore divided by the bin width before comparison.

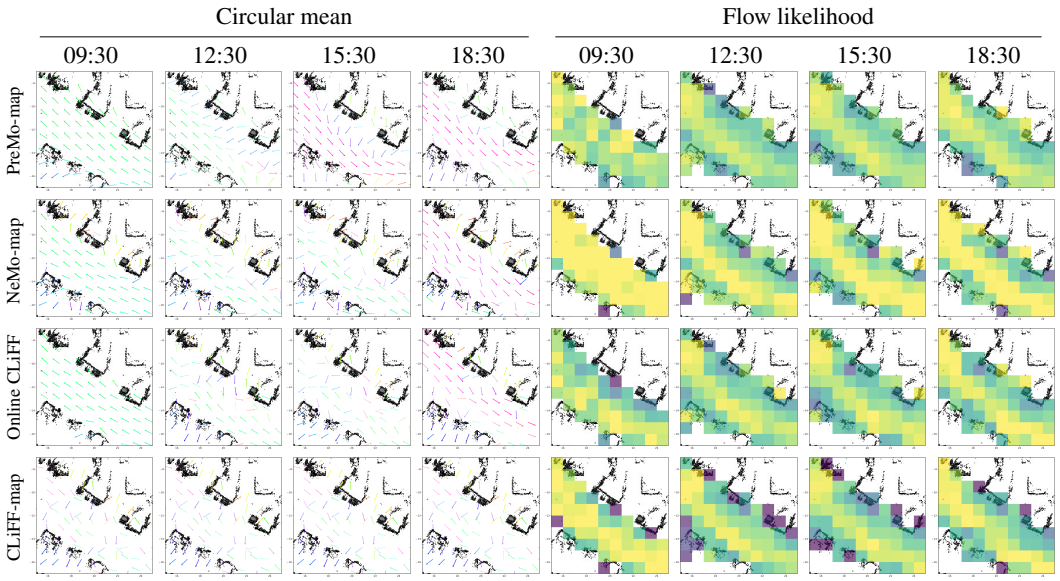

Figure A2: **East corridor through the day**, the region $x \in [15, 25]$, $y \in [-17, -7]$. **Left:** one arrow per cell at the weight-weighted circular mean of the heading. **Right:** per-cell flow likelihood over the detections of that hour. Colours and opacity the same as Figure A1.

Table A1: **Flow likelihood per evaluation day**, whole environment. STeF-map is absent, carrying no speed. Best per column in bold.

	$\mathcal{L}_{\text{flow}} \downarrow$		
Estimator	10-28 (Sun)	10-31 (Wed)	11-04 (Sun)
Uniform flow (reference)	2.342	2.263	2.200
CLiFF-map (Kucner et al., 2017)	3.183	1.771	2.551
Online CLiFF-map (Zhu et al., 2025b)	2.425	1.730	2.089
NeMo-map (Zhu et al., 2026)	**2.105**	**1.290**	**1.685**
PreMo-map *(ours)*	2.126	1.649	1.811

A4 Qualitative Results

We report qualitative examples of PreMo-map compared with the baselines in Figures A1 and A2.

Figure A1 shows the central open area at 12:30, where a horizontal flow intersects a vertical one. PreMo-map represents both flows within the same cells, with the component row jointly capturing the leftward, rightward, and downward bands through the intersection. The second row reports the weight-weighted circular mean for each method. Each baseline changes direction between neighbouring cells, whereas PreMo-map's mean direction varies gradually across the region. In the third row, the three baselines reach the upper end of the likelihood scale in the lower-left portion of the region, whereas PreMo-map remains at intermediate values. In the upper half, the four methods produce similar likelihoods.

Figure A2 shows the east corridor at four times, reported as the circular mean. PreMo-map reverses its dominant direction over the day: at 09:30 the mean points towards the upper left of the corridor and by 18:30 towards the lower right. CLiFF-map is fitted once over the training day and draws the same field at every time, so a reversal is outside what it can represent.

Figure A2 shows the east corridor at four times, represented by the circular mean. PreMo-map reverses its dominant direction over the course of the day: at 09:30, the mean points towards the upper left of the corridor, whereas by 18:30 it points towards the lower right. CLiFF-map is fitted once over the training day and therefore produces the same field at every time.

A5 Additional Results

We report $\mathcal{L}_{\text{flow}}$ over the whole environment for each evaluation day (see Table A1). On 10-28, PreMo-map has a flow NLL of 2.126 nats and 2.105 for NeMo-map, a difference of 0.021 nats. On 10-31, the corresponding values are 1.649 and 1.290, a difference of 0.359 nats, and on 11-04 they differ by 0.126 nats. Both methods have lower NLL than the uniform flow on all three days. CLiFF-map ranges from 1.771 to 3.183 nats and has higher NLL than the uniform flow on both Sundays; its online variant exceeds the uniform-flow NLL only on 10-28.