A Potential Field-Based Approach to Adversarial Inverse Reinforcement Learning for Pedestrian Behavior Modeling

Daichi Ogawa[1], Eiji Hato*[1], and Kenan Zhang[2]

[1]Department of Civil Engineering, The University of Tokyo, Japan
[2]EPFL, School of Architecture, Civil and Environmental Engineering, Lausanne, Switzerland

Short Summary

End-to-end autonomous driving systems generate control signals directly from sensor inputs, implicitly assuming that the surrounding environment is exogenous. However, at unsignalized intersections, pedestrians actively adjust their actions in response to approaching vehicles, possibly leading to inefficient or unsafe control. To accurately capture interactions across all agents, we propose Mean-Field Adversarial Inverse Reinforcement Learning with Continuous Actions and Kernel Decomposition (MF-AIRL-C+K) to quantitatively evaluate pedestrian reactive behavior in the presence of surrounding pedestrians and vehicles, by extending adversarial inverse reinforcement learning to continuous action spaces and incorporating mean-field-based interaction. Specifically, restricting the utility function to a negative-definite quadratic form guarantees a unique and well-defined Gaussian optimal policy. In addition, the parameterization of the reward function based on kernel decomposition further enforces the condition of potential Mean-Field Games and enables interpretable spatial analysis. Experiments on pedestrian trajectories collected at JR Matsuyama Station show that MF-AIRL-C+K outperforms all baselines on all metrics.
Keywords. continuous action space, multimodal interaction, pedestrian behavior modeling, adversarial inverse reinforcement learning, potential mean-field game

1 Introduction

End-to-end (E2E) autonomous driving has advanced rapidly, with systems trained to generate vehicle control signals directly from raw sensor inputs (Bojarski et al., 2016). Despite strong performance in structured environments, E2E models implicitly assume that other road users are independent of the ego vehicle. At unsignalized intersections, this assumption breaks down because pedestrians actively adjust their speed and direction in response to approaching vehicles (Onkhar et al., 2022; Bazilinskyy et al., 2021). Neglecting such reactive behavior leads to systematic prediction errors and, in safety-critical settings, can cause an overly conservative planner to enter a paralysis state known as the robot freezing problem (Trautman & Krause, 2010).

Accurately modeling pedestrian reactive behavior requires a scalable framework for pedestrian-pedestrian and pedestrian-vehicle interaction. Pairwise approaches such as the Social Force Model (SFM) (Helbing & Molnár, 1995) and Social GAN (Gupta et al., 2018) incur a computational cost that scales quadratically with the number of agents, limiting applicability to large crowds at busy intersections. Mean-field games (MFG) (Lasry & Lions, 2007) resolve this challenge by replacing pairwise couplings with interaction between each individual and the aggregate density field, reducing per-agent complexity from $O(N^2)$ to $O(N)$. Existing MFG-based pedestrian models, however, rely on manually designed cost functions that cannot adapt to observational data and fail to capture the nuanced reactive patterns induced by vehicular presence.

Inverse reinforcement learning (IRL) (Ziebart et al., 2008) provides a principled remedy by recovering reward functions from demonstrated behavior. Prior IRL work on crowd modeling, including the discrete-action mean-field framework (MF-AIRL) (Chen et al.,

2025), discretizes the action space, causing approximation errors tied to grid resolution that make continuous velocity and heading direction difficult to represent faithfully. Extending Adversarial Inverse Reinforcement Learning (AIRL) (Fu et al., 2018) to continuous action spaces introduces a theoretical obstacle because AIRL defines the optimal policy proportional to the exponential of the action-value function, whose normalizing constant over the continuous action domain may diverge and leave the policy ill-defined. Existing AIRL frameworks do not resolve this divergence risk for continuous actions.

This paper proposes Mean-Field Adversarial Inverse Reinforcement Learning with Continuous Actions and Kernel Decomposition (MF-AIRL-C+K). The continuous-action extension resolves the normalizing constant divergence inherent in applying AIRL to continuous action spaces, guaranteeing a well-defined unique optimal policy throughout training. The kernel reward decomposition further makes the recovered spatial interaction structure interpretable, enabling post-hoc analysis of how the surrounding agent density shapes individual utility.

2 Methodology

Problem Formulation

We consider a population of pedestrians moving in a shared pedestrian–vehicle space. Each pedestrian is characterized by a state $x \in \mathbb{R}^2$ representing their current position, a continuous action $a = (v_x, v_y)^\top \in \mathbb{R}^2$ comprising Cartesian velocity components, a destination context c, and a mean field μ. In standard MFG formulations, the mean field captures only the density of the agents being modeled. Here we extend μ to encode both pedestrian density and vehicle occupancy distribution over a spatial grid, giving each agent a compact aggregate representation of the surrounding environment without explicit pairwise enumeration.

We formulate pedestrian interactions as a mean-field game (MFG) in the infinite-player limit. Each individual maximizes their expected utility while responding solely to the aggregate density μ rather than to specific neighbors. The objective of our framework is to recover the underlying utility function and policy from observed pedestrian trajectories.

Model Architecture

Mean-Field Adversarial Inverse Reinforcement Learning with Continuous Actions (MF-AIRL-C) consists of three neural networks. The policy network outputs the Gaussian policy parameters $\bar{a}_G(x, \mu, c)$ and $\Sigma_G(x, \mu, c)$. The utility network outputs the state-action-dependent utility function $g_\theta(x, a, \mu, c)$. The value network outputs the state-value function $h_\phi(x, \mu, c)$. All three networks receive the current state x, mean field μ, and destination context c as inputs.

Unlike standard AIRL, we restrict the utility function g_θ to be a negative-definite quadratic form in action a, with $r_\theta(x, \mu, c)$ as the state-dependent reward component.

$$g_\theta(x, a, \mu, c) = -\tfrac{1}{2}(a - \bar{a}_\theta)^\top \Sigma_\theta^{-1}(a - \bar{a}_\theta) + r_\theta(x, \mu, c). \tag{1}$$

Following the AIRL framework (Fu et al., 2018), the optimal policy takes the Boltzmann form given in Eq. (2), where the advantage function $f_{\theta,\phi}$ is defined in Eq. (3a). The quadratic structure of g_θ ensures that the partition function (the denominator of Eq. (2)) is finite, a property exploited in Proposition 1.

$$\pi^*(a \mid x, \mu, c) = \frac{\exp(f_{\theta,\phi}(x, a, \mu, x', \mu', c))}{\int \exp(f_{\theta,\phi}(x, a', \mu, x', \mu', c))\, da'}, \tag{2}$$

The advantage function $f_{\theta,\phi}$ and discriminator $D_{\theta,\phi}$ are defined as

$$f_{\theta,\phi}(x, a, \mu, x', \mu', c) = g_\theta(x, a, \mu, c) + \gamma h_\phi(x', \mu', c) - h_\phi(x, \mu, c), \tag{3a}$$

$$D_{\theta,\phi}(x, a, \mu, x', \mu', c) = \frac{\exp(f_{\theta,\phi})}{\exp(f_{\theta,\phi}) + \pi_G(a \mid x, \mu, c)}, \tag{3b}$$

where $\pi_G(a \mid x, \mu, c)$ is the probability density computed from the generator's output $\bar{a}_G$ and Σ_G. The discriminator does not require explicit computation of optimal choice probability according to the obtained advantage function, making it tractable for continuous action spaces.

Proposition 1

Proposition 1. *Under the MFG infinite-player assumption with Cartesian actions $a = (v_x, v_y)^\top$, when g_θ is a negative-definite quadratic form in a as in Eq. (1), applying a first-order Taylor expansion of h_ϕ about x, the partition function $\int \exp(f_{\theta,\phi})\, da$ is always finite and satisfies*

$$\int \exp(f_{\theta,\phi})\, da = 2\pi\, |\Sigma_\theta|^{1/2} \exp\big(r_\theta(x, \mu, c) + \gamma h_\phi(x, \mu', c) - h_\phi(x, \mu, c) + Q\big), \tag{4}$$

where, writing $\nabla h' \equiv \nabla_x h_\phi(x, \mu', c)$ for brevity, $Q = \gamma \Delta t\, \bar{a}_\theta^\top \nabla h' + \frac{(\gamma \Delta t)^2}{2} (\nabla h')^\top \Sigma_\theta\, \nabla h'$, and the Boltzmann optimal policy is uniquely $\pi^* = \mathcal{N}\big(\bar{a}^*,\, \Sigma_\theta\big)$ with $\bar{a}^* = \bar{a}_\theta + \gamma \Delta t\, \Sigma_\theta \nabla h'$.

Proof sketch. Under the MFG infinite-player assumption, each individual agent's action has no effect on the aggregate density, so μ' is treated as an observed variable independent of a. Furthermore, with Cartesian actions $a = (v_x, v_y)^\top$, the next state satisfies $x' = x + \Delta t\, a$, so the correction $\gamma h_\phi(x'(a), \mu', c) \approx \gamma h_\phi(x, \mu', c) + \gamma \Delta t\, (\nabla h')^\top a$ contributes only a linear term in a to the exponent. Completing the square shifts the mean from $\bar{a}_\theta$ to $\bar{a}^* = \bar{a}_\theta + \gamma \Delta t\, \Sigma_\theta \nabla h'$ without changing the covariance, and the integral evaluates to $2\pi |\Sigma_\theta|^{1/2}$ times a state-dependent exponential. The optimal policy is therefore $\mathcal{N}(\bar{a}^*, \Sigma_\theta)$, well-defined for all $\Sigma_\theta \succ 0$; the policy network, receiving inputs (x, μ, c), learns $\bar{a}^*$ directly through training convergence without architectural modification. In the limit $\Delta t \to 0$, $\bar{a}^* \to \bar{a}_\theta$ and the policy collapses to $\mathcal{N}(\bar{a}_\theta, \Sigma_\theta)$.

Remark (necessity). The quadratic form is also necessary, as $\log \pi^*(a) = g_\theta(a) - \text{const}$ implies the policy is Gaussian if and only if g_θ is quadratic in a.

Potential Functional Parameterization of the State Reward

We extend MF-AIRL-C with a learnable interaction kernel in the state reward, yielding MF-AIRL-C+K (MF-AIRL-C and Kernel Decomposition). Inspired by Lachapelle & Wolfram (2011), who introduced interaction kernels within a potential MFG framework to model pedestrian congestion and aversion, we parameterize the state-dependent scalar reward $r_\theta(x, \mu, c)$ as the Fréchet derivative of a potential functional $F_\theta[\mu]$, a design that automatically enforces the symmetry condition of potential MFGs and makes F_θ itself an interpretable learning outcome. We adopt a kernel decomposition

$$F_\theta[\mu] = \int \rho_\theta(x, c)\, \mu(x)\, dx + \tfrac{1}{2} \iint k_\theta(x - y)\, \mu_{\text{ped}}(x)\, \mu_{\text{ped}}(y)\, dx\, dy, \tag{5}$$

where $y \in \mathbb{R}^2$ is another position in the pedestrian space, $\mu_{\text{ped}}(x)$ denotes the pedestrian-only component of the mean field (i.e. μ restricted to pedestrian occupancy, excluding vehicles), $\rho_\theta(x, c)$ encodes location- and destination-specific preferences, and $k_\theta(x - y)$ is a learnable symmetric interaction kernel implemented as a convolutional layer over the spatial grid. When k_θ is even-symmetric ($k_\theta(-z) = k_\theta(z)$), the Fréchet derivative of F_θ with respect to μ at position x yields the state reward analytically as

$$r_\theta(x, \mu, c) = \frac{\delta F_\theta}{\delta \mu(x)} = \rho_\theta(x, c) + (k_\theta * \mu_{\text{ped}})(x). \tag{6}$$

This parameterization provides two advantages. First, the potential functional F_θ is a direct output of learning, enabling interpretable analysis of how aggregate density shapes individual utility. Second, the interaction kernel k_θ characterizes the Hessian of F_θ via $H(x, y) = \delta^2 F_\theta / \delta\mu(x)\,\delta\mu(y) = k_\theta(x - y)$, so the learned kernel provides an analytically tractable map of spatial interdependencies across the pedestrian space. The even-symmetry constraint guarantees $H(x, y) = H(y, x)$, consistent with the symmetry requirement of potential MFGs, and supports post-hoc quantitative analysis beyond raw reward values (see Figure 2).

Training Procedure

The networks are trained via alternating adversarial updates. The discriminator and generator parameters are each updated by minimizing the following losses.

$$\mathcal{L}_D = -\mathbb{E}_{\text{Data}}[\log D_{\theta,\phi}] - \mathbb{E}_{\pi_G}[\log(1 - D_{\theta,\phi})], \tag{7a}$$

$$\mathcal{L}_G = \mathbb{E}_{a\sim\pi_G}\left[\log \pi_G(a)\left(\overline{\log \pi_G(a)} - f_{\theta,\phi}(x, a, \mu, x'_a, \mu'_a, c) - 1\right)\right], \tag{7b}$$

$\mathcal{L}_G$ is derived via the score-function estimator (REINFORCE, Williams 1992) and serves as a surrogate for the standard AIRL generator maximization objective $\mathbb{E}_{\pi_G}[f_{\theta,\phi} - \log \pi_G]$. The moving-average baseline $\overline{\log \pi_G(a)}$ reduces gradient variance without introducing bias. At each time step, actions sampled from π_G are aggregated onto a spatial grid to construct an empirical density estimate, which is used directly as the mean field μ' for the next step.

3 Results and Discussion

Dataset

We use a proprietary pedestrian trajectory dataset collected at JR Matsuyama Station (including an unsignalized crosswalk) on September 28, 2022 by the University of Tokyo and Fukken Surveys and Designs Co., Ltd. Pedestrian trajectories were extracted via YOLO (Jocher & Qiu, 2026) detection, BoT-SORT (Aharon et al., 2022) tracking, projective transformation to UTM coordinates, and Kalman filter smoothing. Vehicle trajectories followed an analogous pipeline using YOLO instance segmentation and a kinematic Kalman filter.

Experimental Setup

We compare MF-AIRL-C+K against three baselines that differ in action space, interaction model, and learning paradigm. SFM (Helbing & Molnár, 1995) employs hand-crafted pairwise social forces with continuous actions and requires no learning. BC+MF uses maximum-likelihood behavior cloning with a mean-field interaction module, yielding a continuous-action learned policy without reward recovery. MF-AIRL (Chen et al., 2025) applies AIRL in a discrete action space with mean-field interaction, providing a direct comparison on the effect of action-space discretization. Additionally, an ablation study assesses the contribution of mean-field interaction and vehicle context by removing the mean-field input μ entirely (w/o MF) or by removing only the vehicle occupancy component from μ while retaining pedestrian density (w/o Veh.).
All methods are evaluated on Average Displacement Error (ADE), Final Displacement Error (FDE), and pedestrian collision rate.

Main Results

MF-AIRL-C+K achieves the lowest ADE, FDE, and collision rate among all methods (Table 1). Relative to SFM, the halved ADE confirms that data-driven reward learning

Table 1: **Trajectory prediction accuracy and pedestrian collision rate.** ADE and FDE are in metres, and collision rate is scaled by 10^{-3}. Best values shown in bold.

Method	ADE (m) ↓	FDE (m) ↓	Coll. Rate ($\times 10^{-3}$) ↓
SFM (Helbing & Molnár, 1995)	2.72	4.28	1.07
BC+MF	1.78	3.05	1.21
MF-AIRL (Chen et al., 2025)	2.43	2.38	0.92
MF-AIRL-C+K (proposed)	**1.38**	**2.24**	**0.22**

captures behavioural patterns that hand-crafted social forces cannot represent. Compared with BC+MF, the substantially lower collision rate (0.22×10^{-3} vs. 1.21×10^{-3}) demonstrates that adversarial reward recovery yields a qualitatively better policy than direct imitation. Over MF-AIRL (Chen et al., 2025), the continuous formulation substantially reduces both ADE and collision rate, reflecting better tracking of curved crosswalk paths where discrete grids impose angular biases.

Ablation Study

Table 2: **Ablation study results.** ADE and FDE are in metres, collision rate is scaled by $\times10^{-3}$, and best values are shown in bold.

Model variant	ADE (m) ↓	FDE (m) ↓	Coll. Rate ($\times 10^{-3}$) ↓
MF-AIRL-C+K (full)	**1.38**	**2.24**	**0.22**
w/o MF	2.17	3.60	0.22
w/o Veh.	1.68	2.66	0.86

Table 2 shows that removing μ entirely (w/o MF) raises ADE from 1.38 m to 2.17 m and FDE from 2.24 m to 3.60 m, confirming that density-field interaction is the primary driver of trajectory accuracy. Without μ, the policy falls back on destination-conditioned kinematics alone, producing systematic errors in congested scenes where crowd-induced route adjustments dominate observed trajectory curvature.
Removing only the vehicle component (w/o Veh.) causes a moderate accuracy decline, but the collision rate rises nearly fourfold to 0.86×10^{-3}, confirming that vehicle context is the critical signal for safety-aware avoidance.

Qualitative Analysis

In the low-density scene (a), the continuous covariance Σ_θ expands freely, yielding wide, smooth Gaussian ellipses that envelope the observed positions at t+1 s without the angular quantization error inherent in discrete action grids.
In the high-pedestrian-density scene (b), the confidence ellipses rotate away from the crowd, confirming that the mean-field input μ shapes the predicted distributions. In the high-vehicle-density scene (c), vehicle presence constrains the admissible heading, and the learned Σ_θ reproduces the resulting elongation as an ellipse with aspect ratio approximately 3 to 1, a resolution unachievable with discrete action grids.
The even-symmetry constraint $k_\theta(-z) = k_\theta(z)$ on the learned kernel (Figure 2) guarantees $H(x, y) = H(y, x)$, satisfying the symmetry condition of potential MFGs. The spatial pattern of H is broadly consistent with known pedestrian behaviour. Within the crosswalk, H takes small positive values reflecting weak following behaviour, while at the footpath–crosswalk boundary H turns weakly negative, consistent with the congestion avoidance kernel identified by Lachapelle & Wolfram (2011). In the roadway area beyond the crosswalk,

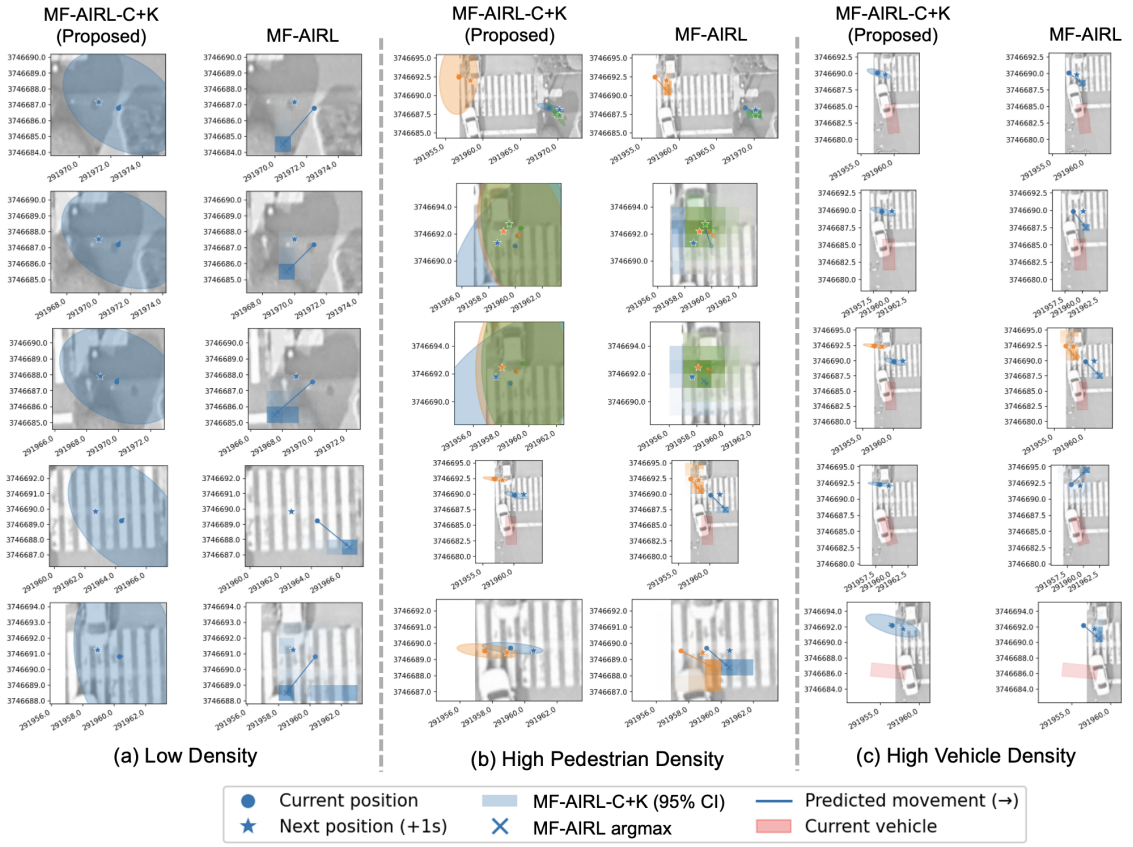


Figure 1: **Trajectory prediction comparison for three crowd-density conditions (1-step, 1 s prediction horizon).** Each panel shows the background aerial map (greyscale), observed position at t+1 s (★, colour-coded per agent), MF-AIRL-C+K (proposed) predicted distributions as Gaussian 95% confidence ellipses (translucent, matching colour), and MF-AIRL (Chen et al., 2025) predicted points (×, colour-coded per agent).

H shows strong positive values. However, as training samples in this region are sparse, this signal possibly reflects estimation instability that could propagate into a downstream vehicle planner and cause it to misjudge pedestrian crowd dynamics near the crosswalk.

4 Conclusions

We proposed MF-AIRL-C+K, which extends adversarial inverse reinforcement learning to continuous action spaces with mean-field interaction to model pedestrian reactive behavior at unsignalized intersections. Proposition 1 establishes that a negative-definite quadratic utility guarantees a unique closed-form Gaussian optimal policy, and kernel decomposition of the state reward enforces the potential MFG symmetry condition while enabling interpretable Hessian analysis (Figure 2).

Experiments at JR Matsuyama Station show that MF-AIRL-C+K outperforms all baselines on ADE, FDE, and collision rate. Ablation results confirm that mean-field interaction is the primary driver of trajectory accuracy, while vehicle occupancy context is the critical signal for safety-aware avoidance, with its removal causing a fourfold increase in collision rate. The learned Hessian map reveals kernel patterns consistent with known pedestrian behavior at crosswalks, supporting the interpretability of the recovered reward structure.

Future work will integrate the learned reactive model into an autonomous vehicle control pipeline (Bazilinskyy et al., 2021) to enable anticipation of pedestrian responses at unsignalized intersections.

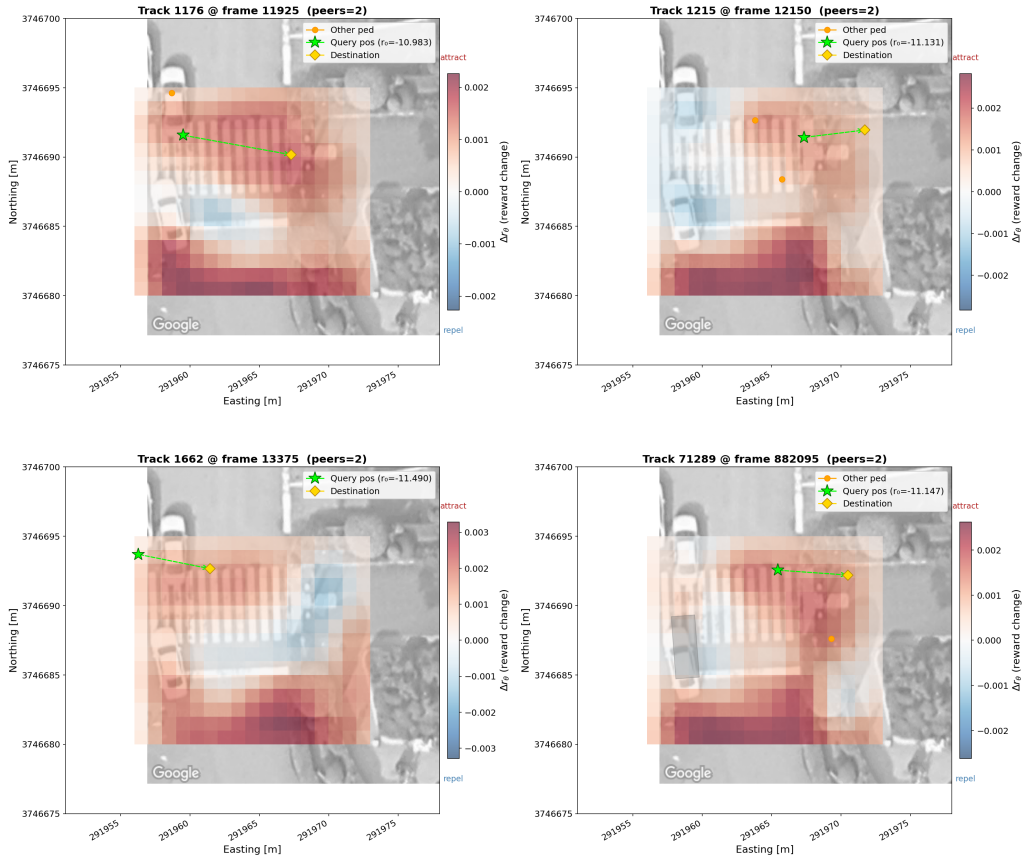


Figure 2: **Spatial interaction Hessian $H(x_0, y) = k_\theta(x_0 - y)$ learned by MF-AIRL-C+K.** The colour map (diverging, blue→white→red) is overlaid on a greyscale bird's-eye map of the survey site. The white circle marks the reference point x_0 fixed on the unsignalised crosswalk. Red (positive) regions indicate that a higher pedestrian density at y increases the reward at the query position x_0, while blue (negative) regions indicate the opposite.

Acknowledgements

The authors would like to thank Professor Pavlo Bazilinskyy at Eindhoven University of Technology for his valuable advice and guidance.
This research was supported by JSPS KAKENHI Grant Number JP25KJ0974 and the Fukushima International Research and Education Institute (JPFR 26 05 07 01).
The authors used Claude to improve the grammar and phrasing of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Aharon, N., Orfaig, R., & Bobrovsky, B.-Z. (2022). BoT-SORT: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*. Retrieved from `https://arxiv.org/abs/2206.14651`

Bazilinskyy, P., Sakuma, T., & de Winter, J. (2021, September). What driving style makes pedestrians think a passing vehicle is driving automatically? *Applied Ergonomics*, *95*, 103428. doi: 10.1016/j.apergo.2021.103428

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... Zieba, K. (2016). *End to end learning for self-driving cars.* arXiv preprint arXiv:1604.07316. Retrieved from `https://arxiv.org/abs/1604.07316`

Chen, Y., Zhang, L., Liu, J., & Witbrock, M. (2025). *Adversarial inverse reinforcement learning for mean field games.* arXiv preprint. Retrieved from http://arxiv.org/abs/2104.14654 doi: 10.48550/arXiv.2104.14654

Fu, J., Luo, K., & Levine, S. (2018). Learning robust rewards with adversarial inverse reinforcement learning. In *International conference on learning representations.* Retrieved from https://openreview.net/forum?id=rkHywl-A-

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., & Alahi, A. (2018). Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2255–2264). Retrieved from http://openaccess.thecvf.com/content_cvpr_2018/html/Gupta_Social_GAN_Socially_CVPR_2018_paper.html doi: 10.1109/CVPR.2018.00240

Helbing, D., & Molnár, P. (1995). Social force model for pedestrian dynamics. *Physical Review E*, *51*(5), 4282–4286. doi: 10.1103/PhysRevE.51.4282

Jocher, G., & Qiu, J. (2026). *Ultralytics YOLO26.* Version 26.0.0. Retrieved from https://github.com/ultralytics/ultralytics

Lachapelle, A., & Wolfram, M.-T. (2011). On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation Research Part B: Methodological*, *45*(10), 1572–1589. doi: 10.1016/j.trb.2011.07.011

Lasry, J.-M., & Lions, P.-L. (2007). Mean field games. *Japanese Journal of Mathematics*, *2*(1), 229–260. doi: 10.1007/s11537-007-0657-8

Onkhar, V., Bazilinskyy, P., Dodou, D., & de Winter, J. C. F. (2022, January). The effect of drivers' eye contact on pedestrians' perceived safety. *Transportation Research Part F: Traffic Psychology and Behaviour*, *84*, 194–210. doi: 10.1016/j.trf.2021.10.017

Trautman, P., & Krause, A. (2010). Unfreezing the robot: Navigation in dense, interacting crowds. In *2010 IEEE/RSJ international conference on intelligent robots and systems* (pp. 797–803). doi: 10.1109/IROS.2010.5654369

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, *8*(3–4), 229–256. doi: 10.1007/BF00992696

Ziebart, B. D., Maas, A. L., Bagnell, J. A., & Dey, A. K. (2008). Maximum entropy inverse reinforcement learning. In *Proceedings of the twenty-third AAAI conference on artificial intelligence (AAAI 2008)* (pp. 1433–1438). AAAI Press. Retrieved from https://cdn.aaai.org/AAAI/2008/AAAI08-227.pdf