Commun Transp Res

Research Article

SAFE-AD: Socially aware field-enhanced reinforcement learning for autonomous driving in interactive traffic

Zian Wang[1], Wenjie Huang[1,2], Zejian Deng[1], Yiming Shu[1], Jiahui Xu[1], Yong Wang[1], Shen Li[3], Dongpu Cao[4], Chen Sun[1,✉]

Cite this article: https://doi.org/10.26599/COMMTR.2026.9640067

ABSTRACT: An autonomous vehicle can complete its route while disrupting surrounding traffic. We introduce SAFE-AD, augmenting reinforcement learning (RL) observations with compact descriptors of a finite-horizon interaction-pressure field constructed from current traffic states. The complete objective supplements the environment's task reward with risk, comfort, and externality costs. Then, we formulate a context-conditioned physics-informed neural network (PINN) to approximate the numerical field. Evaluation combines held-out naturalistic recordings, HighwayEnv diagnostics, and closed-loop MetaDrive tests in various traffic scenarios, with a matched factorial experiment with continuous soft actor-critic (SAC) separates observation and objective effects. At unprotected intersections, the complete profile reduces collision rate from 0.296 to 0.220 and mean absolute jerk by 0.98 m/s³ relative to baseline SAC. Besides, the mean absolute jerk in roundabout decreases by 1.54 m/s³. The results show scenario-specific gains and sensitivity to objective calibration.

KEYWORDS: Autonomous driving; socially aware reinforcement learning; physics-informed neural networks; spatiotemporal risk field; tactical behavior planning; externality-aware decision making.

1 Introduction

A vehicle can complete a merge while forcing a follower to brake sharply. At other scenarios like roundabouts and unprotected intersections, the timing of an ego maneuver likewise affects nearby traffic as well as the ego vehicle (Deng et al., 2024). Socially compatible driving requires attention to these effects alongside route progress, collision risk, and passenger comfort. A useful decision-making representation should make the spatial structure of emerging interactions available earlier to guide the prompt maneuver.

Risk-field and corridor-based methods express interaction costs across spatiotemporal domain and incorporate them into decision-making and trajectory planning (Han et al., 2023; Ning et al., 2024; Li, Z. et al., 2026). While risk-aware reinforcement learning (RL) then optimizes agents' decisions under uncertainty (Huang and Haskell, 2017, 2020). When a spatial field enters learning only through a scalar penalty, the policy must obtain information about its spatial variation from other inputs. Social-preference models introduce further tradeoffs between ego benefits and reactions of surrounding traffic (Deng et al., 2024; Hu, W. et al., 2025). Introducing field features and changing these costs together, however, makes it difficult to identify which change accounts for an observed improvement.

We propose **SAFE-AD**, a framework for **socially aware field-enhanced reinforcement learning** for **autonomous driving** in interactive traffic. Building on the risk-field formulation of Wang et al. (2026), the framework (Fig. 1) constructs finite-horizon interaction pressure from current vehicle states and road geometry. Each query holds the observed scene fixed while propagating its sources over the prospective horizon. Compact field descriptors augment the policy observation. Separate objective profiles retain the environment's task reward and introduce risk, comfort, and externality costs. A context-conditioned physics-informed neural network (PINN) approximates the numerical field as an alternative querying backend.

The evaluation distinguishes prospective field relevance, surrogate fidelity, and closed-loop policy behavior. HighwayEnv (Leurent, 2018) supports field interpretation across highway scenarios, while MetaDrive (Li et al., 2022) provides matched comparisons with high-fidelity qualitative result presentations.

The main contributions of this work include:

(1) We develop a field-based policy interface and a context-conditioned PINN surrogate. The descriptors expose ego and look-ahead pressure, adjacent-lane corridor risk, and local gradients, while backward pressure provides a proxy for potential follower disturbance.

(2) We use a matched observation–objective design with continuous SAC to separate field-representation and risk-cost effects. A further comparison evaluates the complete profile as a joint change to the representation and objective.

(3) We combine held-out policy comparisons with reward-independent traffic measures and a separate merging calibration. The evaluation documents scenario-specific gains and examines how jerk regularization affects the

[1] Department of Data and Systems Engineering, The University of Hong Kong, Hong Kong 999077, China. [2] Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong 999077, China. [3]Department of Civil Engineering, Tsinghua University, Beijing 100084, China. [4]State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University, Beijing 100084, China.

✉ Corresponding author. E-mail: c87sun@hku.hk

Received: June 2, 2026; Revised: August 23, 2026; Accepted: October 30, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, http://creativecommons.org/licenses/by/4.0/).

Under Accepted

balance between progress, safety, and comfort.

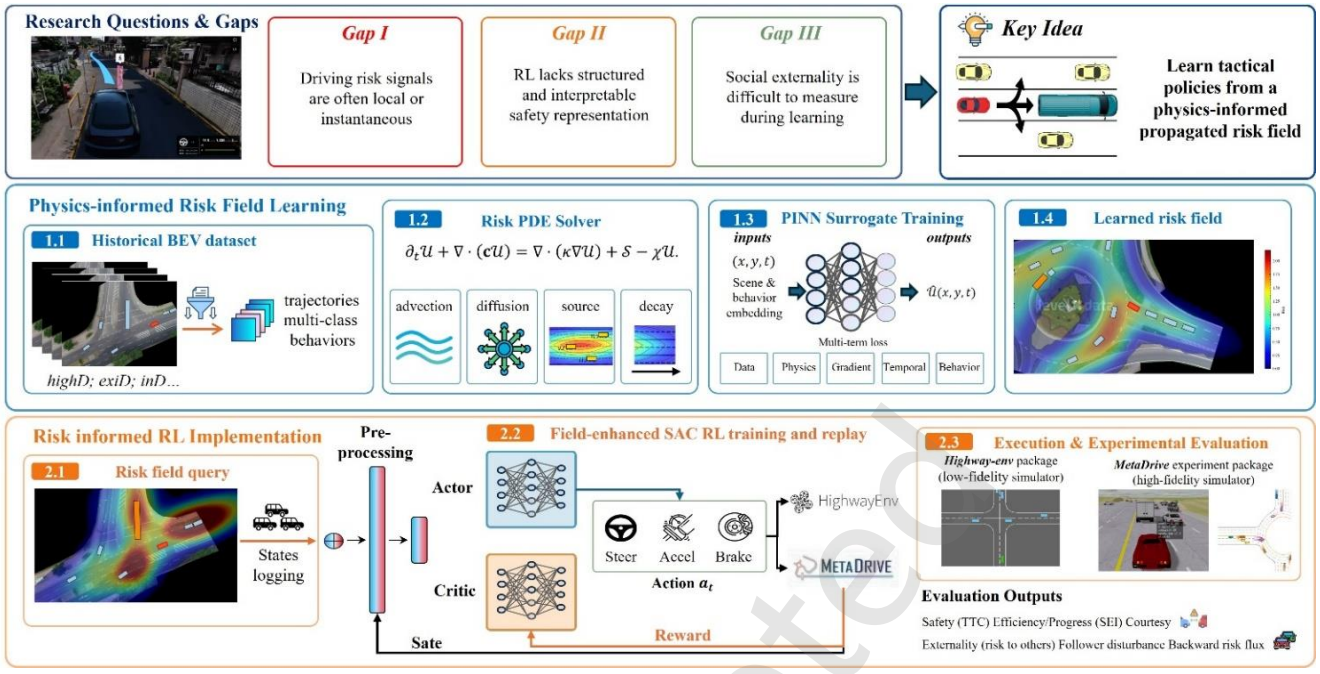


Fig. 1. SAFE-AD framework.

2 Related work

2.1. Spatiotemporal risk-field modeling

Risk fields have been widely adopted for representing driving danger as a continuous spatial quantity in the literature. Artificial Potential Field (APF) methods have been adopted in modeling potential costs for road users in order to avoid collisions and plan trajectories (Kilic et al., 2026; Xing et al., 2024; Han et al., 2023). The costs modeled by APF methods are typically incorporated into a downstream layer such as MPC (Ning et al., 2024), with Control barrier functions (CBFs) enforce explicit safety conditions as hard constraints (Liu and Wang, 2026). Most existing driving-risk fields are handcrafted and mostly instantaneous. Even though continuum and mean-field models successfully capture traffic-flow propagation, there's little connection to learned tactical control policies (Zhao and Yu, 2025). Instead, using PINNs allows for differentiable approximations of field operators (Guo et al., 2025), which contributes to a learnable risk assessment from real driving data and fast online querying for safety-critical autonomous driving systems.

2.2. Risk-aware RL for tactical driving

Typical RL algorithms have been applied to support decision-making for autonomous driving in facing uncertainty for long-horizon planning (Sheng et al., 2024; Jia et al., 2026), including proximal policy optimization (PPO), deep Q-networks (DQN), SAC, deep deterministic policy gradient (DDPG), and twin delayed deep deterministic policy gradient (TD3) (Huang et al., 2020; Jiang et al., 2026; Jin et al., 2025; Hu, C. et al., 2025). Discrete control methods typically specify actions as categorical values for lane changes or as constant speed, whereas continuous actor-critic methods, such as those that model steering, throttle/brake controls, use function approximators to represent these. Other studies also focus on behavior-constrained multi-agent policy that combines recurrent interaction encoding formal safety verification for cooperative intersection control (Gao et al., 2025). However, pure progress and efficiency-centric rewards can encourage aggressive maneuvering, whereas strong safety penalties may induce passive and over-conservative policies. Meanwhile, scalar risk terms also convey limited information about evolving merging pressure or upstream disturbance.

2.3. Social compatibility, courtesy, and cooperative planning

Social compatibility between drivers can affect both their ego and other road users. The driving style can be scored based on how socially friendly it is to allow other traffic participants to complete their maneuvers by using models such as those that incorporate courtesy and social-value-orientation (SVO) (Deng et al., 2024; Hu, W. et al., 2025). Cooperation-aware planners model other traffic in the vicinity by means of interactive agents between which the planner balances aggressiveness and conservativeness (Hu, W. et al., 2025; Hu, C. et al., 2025). Human-emulation-based proactive longitudinal control further demonstrates how anticipating neighboring drivers' lane changes can reduce traffic oscillations while facilitating safer interactions (Liu et al., 2026). Other human-like car-following studies assess anticipation and comfort against naturalistic behavior (Chen et al., 2025). The pairwise costs and post-maneuver outcomes quantify social effects but do not describe their spatial development before a safety event.

Therefore, SAFE-AD is designed to address issues by connecting a physics-informed spatiotemporal field with RL and integrates environment understanding that a driving

Just Accepted

policy can use before potential conflicts unfold.

3 Preliminaries and problem formulation

3.1. Problem formulation

The test scenario consists of an autonomous ego vehicle which is driving on the highway and performs cruising, merging, driving through a roundabout and an unprotected intersection. All other vehicles on the scene are driving by means of pre-defined planners within the closed-loop simulation paradigm (details formulated in Section. 5). At each decision step, local kinematics and lane topology define the scene, and a compact field descriptor supplies prospective interaction information to the RL policy. At step t, the scene comprises the ego state and selected-neighbor set as:

$$\mathbf{e}_t = [\mathbf{p}_t^e, \psi_t^e, v_t^e, a_t^e, \ell_t^e]^\top.$$
$$\mathcal{N}_t = \{\mathbf{z}_{j,t}\}_{j=1}^{n_t}, \qquad n_t = |\mathcal{N}_t|. \tag{1}$$

For vehicle j, the normalized vector $\mathbf{z}_{j,t}$ contains position and velocity on both axes heading difference, acceleration, lane relation and topology flags. Table 1 summarizes the notation used in the methodology.

Table 1. Methodological notation.

Symbol	Meaning
$\mathbf{e}_t, \mathcal{N}_t, n_t, \mathcal{G}_t$	Ego state, selected-neighbor set/count, and local geometry
$\boldsymbol{\xi} = (\xi_x, \xi_y)^\top, \Omega_t, M_t^{\text{road}}$	Ego-local coordinate, drivable domain, and road mask
$Q_t, \mathbf{c}_t, D_t, \bar{D}_t$	Source, transport, and diffusivity fields
$U_{H,t}^*, \hat{U}_{\theta,t}, U_{H,t}^b, \bar{U}_t$	Numerical teacher, PINN prediction, and selected/clipped backend field
$\boldsymbol{\eta}_t, \mathbf{o}_t, \mathbf{a}_t, \mathbf{u}_t = (\delta_t, p_t)^\top$	Field descriptor, observation, action, and continuous command
π_ϕ, γ	Policy and discount factor
$C_t^{\text{risk}}, C_t^{\text{comf}}, C_t^{\text{soc}}, B_t, C_{\text{int}}, D_v$	Training costs, backward pressure, and independent social metrics

3.2. Dataset processing and behavior extraction

BEV traffic datasets provide comprehensive historical driving scenarios with various situations for guiding the filed modeling. Partial inD recordings are used for surrogate calibration, and the remaining recordings are held out for validation (Bock et al., 2020). While rounD and exiD provide external qualitative replays (Krajewski et al., 2020; Moers et al., 2022). Positions, velocities, headings, vehicle classes, and track identities define each ego-centered sample:

$$\mathbb{s}_t = (\mathbf{e}_t, \mathcal{N}_t, \mathcal{G}_t, \Omega_t).$$
$$\mathbf{b}_t^{\text{src}} = \left[\kappa_{j,t}, d_{j,t}, \Delta\mathbf{v}_{j,t}, \mathbf{a}_{j,t}, \text{TTC}_{j,t}, \text{THW}_{j,t}\right]_{j\in\mathcal{N}_t}.$$
$$\mathbf{y}_{t:t+H}^{\text{val}} = \left[Y_{t:t+H}^{\text{occ}}, \Delta\text{TTC}_{t:t+H}^{\text{rear}}, a_{t:t+H}^{\text{fol}}\right]^\top. \tag{2}$$

3.3. Risk-aware Markov decision process

The tactical task is a partially observable Markov decision process (POMDP), as defined in Eq. (4). The sets $\mathcal{S}$ and $\mathcal{A}$ denote latent states and actions, P is the traffic transition kernel, and r specifies the objective. The remaining terms are discount factor γ, initial-state distribution ρ_0, and observation map $\mathcal{O}$:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \rho_0, \mathcal{O}). \tag{4}$$

Eq. (5) specifies the latent state and field-augmented observation:

$$\mathbf{r}_t^{\text{look}} = [\bar{U}_t(0,0), \bar{U}_t(5,0), \bar{U}_t(10,0), \bar{U}_t(20,0)]^\top.$$
$$\boldsymbol{\eta}_t = \left[\left(\mathbf{r}_t^{\text{look}}\right)^\top, \rho_t^L, \rho_t^R, \partial_x\bar{U}_t(0,0), \partial_y\bar{U}_t(0,0)\right]^\top.$$

3.4. PDE-governed interaction-risk field

Let $\boldsymbol{\xi}$ denote the ego-centered coordinate over the visible drivable domain. Eq. (7) defines the field supplied by backend b, with identical clipping for numerical and PINN outputs before descriptor extraction. The field represents finite-horizon interaction pressure:

In Eq. (2), geometry comprises lane centerlines, boundaries, travel direction, and conflict topology, while the spatial domain defines the ego-centered field window. Source conditioning uses current class, distance, relative motion, time to collision (TTC), and time headway (THW):

$$(\mathbb{s}_t, \mathbf{b}_t^{\text{src}}) \to \left(Q_t, \mathbf{c}_t, D_t, M_t^{\text{road}}\right) \to U_{H,t}^*.$$
$$U_{H,t}^b \in \{U_{H,t}^*, \hat{U}_{\theta,t}\} \to \bar{U}_t \to \boldsymbol{\eta}_t \to a_t. \tag{3}$$

Current scene variables determine the source and transport-diffusion context in Eq. (3). The numerical backend or frozen PINN returns the clipped field from which the policy descriptor is extracted.

$$\mathbf{x}_t = (\mathbf{e}_t, \mathcal{N}_t, \mathcal{G}_t, \bar{U}_t).$$
$$\mathbf{o}_t = [\mathbf{o}_t^{\text{env}}; \hat{\boldsymbol{\eta}}_t], \qquad \hat{\boldsymbol{\eta}}_t = \text{clip}\left(\boldsymbol{\eta}_t \oslash \boldsymbol{\sigma}_\eta, -5, 5\right). \tag{5}$$

The simulator state is concatenated with the eight-channel field descriptor in Eq. (5). The positive scale vector σ_η normalizes the channels, $\oslash$ denotes elementwise division, and clipping to $[-5,5]$ limits the influence of rare peaks. The native-observation ablation removes this descriptor. Eq. (6) defines four pressure samples at the ego and 5, 10, and 20 m ahead, left/right corridor risk, and the local field gradient. A lower corridor value m_t^q indicates safer lane availability:

$$m_t^q = \max_{\xi_x \in [0, L_q]} \bar{U}_t\left(\xi_x, y_q\right), \qquad q \in \{L, R\}, \tag{6}$$

$$U_{H,t}^b : \Omega_t \to \mathbb{R}_{\geq 0}, \qquad \bar{U}_t(\boldsymbol{\xi}) = \text{clip}\left(U_{H,t}^b(\boldsymbol{\xi}), 0, U_{\max}\right). \tag{7}$$

The numerical teacher holds the observed scene fixed during each query and integrates the current source over the prospective horizon in Eq. (8):

$$U_{H,t}^*(\boldsymbol{\xi}) = \frac{1}{Z_H}\int_0^H e^{-\lambda s}\left[G_{2\bar{D}_t s} * Q_t\right](\boldsymbol{\xi} - \mathbf{c}_t s)\, \mathrm{d}s. \tag{8}$$

Not Accepted

Here Q_t is the current-state source, c_t transports it in the ego frame, and D_t is local diffusivity. Its positive median supplies constant diffusivity in Eq. (8). G is the two-dimensional Gaussian kernel. The discount rate is λ, and $Z_H = (1 - e^{-\lambda H})/\lambda$, $\lambda = 0.25\ s^{-1}$ normalizes the integral. The road mask removes off-road support. We use horizon $H = 3\ s$ and step size 0.25 s:

$$\mathcal{L}_t U + \lambda U = \frac{Q_t - e^{-\lambda H} Q_{t,H}}{Z_H}.$$

$$\mathcal{L}_t U = \nabla \cdot (\mathbf{c}_t U) - \bar{D}_t \nabla^2 U. \tag{9}$$

The PINN residual follows the finite-horizon transport identity in Eq. (9), where $Q_{t,H}$ is the terminal transported source. At each quadrature point, the semi-Lagrangian teacher backtraces the source, applies Gaussian diffusion, and removes off-road support. Each query is thus a deterministic function of the current scene.

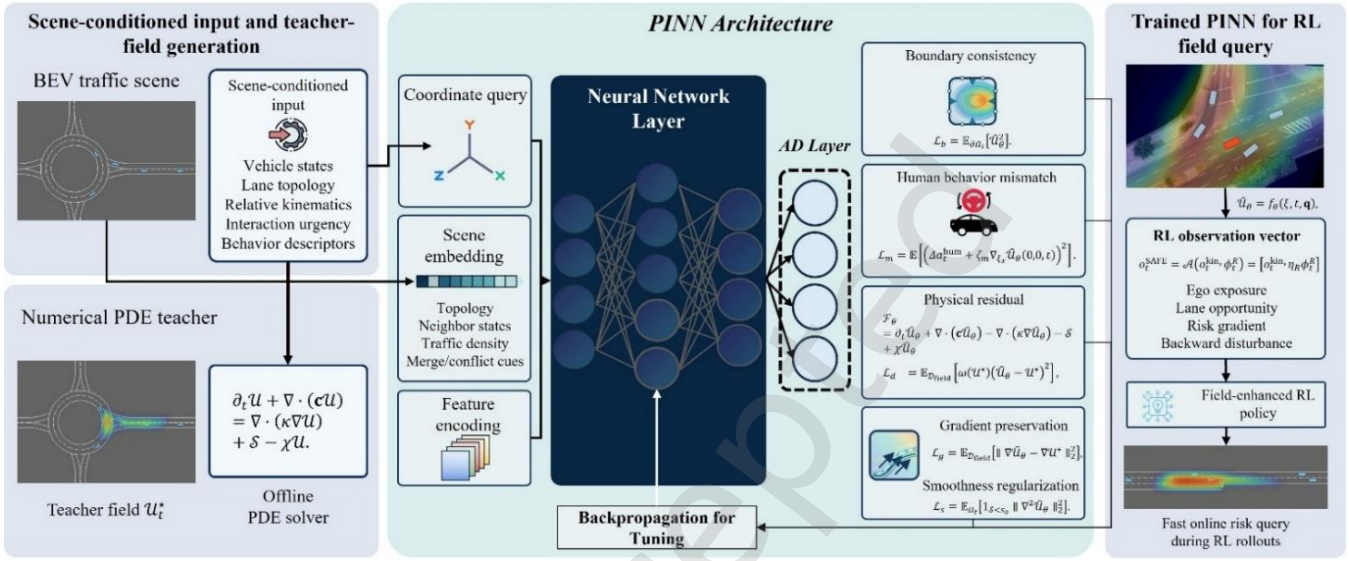


Fig. 2. PINN field learning and querying.

4 Methodology

4.1. BEV trajectory processing and teacher-field generation

Recordings and simulator rollouts are mapped to a common ego-centered 1 m grid. Field construction follows the field-transmission formulation in Wang et al. (2027), with a context-conditioned surrogate and descriptor interface for policy evaluation. Eq. (10) combines visible-vehicle evidence with occlusion and topology terms and defines each vehicle kernel's heading-aligned covariance and nonnegative amplitude:

$$Q_t(\boldsymbol{\xi}) = S_{\mathrm{veh},t}(\boldsymbol{\xi}) + S_{\mathrm{occ},t}(\boldsymbol{\xi}) + S_{\mathrm{topo},t}(\boldsymbol{\xi}).$$

$$S_{\mathrm{veh},t}(\boldsymbol{\xi}) = \sum_{j=1}^{n_t} A_{j,t} \exp\left(-\frac{1}{2}\mathbf{d}_{j,t}^{\top} \Sigma_{j,t}^{-1} \mathbf{d}_{j,t}\right).$$

$$\mathbf{d}_{j,t} = \boldsymbol{\xi} - \mathbf{p}_{j,t}, \qquad \Sigma_{j,t} = R(\psi_{j,t})\,\mathrm{diag}(\sigma_{\parallel,j,t}^2, \sigma_{\perp,j,t}^2) R^{\top}(\psi_{j,t}).$$

$$A_{j,t} = \omega_{\mathrm{cls},j}\omega_{\mathrm{dist},j}\omega_{\mathrm{rel},j}\omega_{\mathrm{brake},j}. \tag{10}$$

Vehicle heading rotates the covariance through $R(\psi_{j,t})$.

4.2. PINN-based risk field surrogate

The surrogate is a stateless context-conditioned operator that maps the full ego-local tensor to a field in one pass (Eq. 12). A sigmoid enforces nonnegativity, the road mask removes off-road support, and a frozen domain scale restores magnitude. Fig. 2 summarizes the fidelity and policy-sensitivity tests:

$$\hat{U}_{\theta,t} = s_d M_t^{\mathrm{road}}\, \sigma\big(f_\theta(\bar{\mathbf{C}}_t)\big).$$

$$\mathbf{C}_t^{\mathrm{src}} = \left[Q_t, \{G_\sigma * Q_t\}_{\sigma\in\{2,6,15\}}\right], \qquad \mathbf{C}_t^{\mathrm{dist}} = \left[e^{-d_{\min,t}/\rho}\right]_{\rho\in\{2,6,15\}}.$$

$$\mathbf{C}_t = \left[\mathbf{C}_t^{\mathrm{src}}, c_{x,t}, c_{y,t}, D_t, M_t^{\mathrm{occ}}, M_t^{\mathrm{road}}, \mathbf{C}_t^{\mathrm{dist}}, \hat{x}, \hat{y}, \dot{\psi}_t^e, n_t, b_d\right].$$

$$\bar{C}_{t,k} = \mathrm{clip}\left(\frac{C_{t,k}}{s_{d,k}}, -4, 4\right), \qquad s_{d,k} = \mathrm{stat}\big(\mathcal{D}_{\mathrm{cal},d,k}\big). \tag{12}$$

Closing speed and braking expand the longitudinal scale $\sigma_{\parallel}$, while lane occupancy sets $\sigma_{\perp}$, producing an upstream wake. The amplitude combines four current cues. The class multiplier is 2.0 for cars and 2.5 for heavy vehicles. The braking multiplier defaults to 3.0 when braking evidence is unavailable and takes 2.5, 4.0, or 6.0 as observed deceleration increases. Transport and discounting follow Eq. (8).

Quadrature yields the cached implementation in Eq. (11). The transport operator combines semi-Lagrangian backtracking, diffusion, and road masking at locations $s_k = k\Delta s$.

$$U_{H,t}^{*} = \frac{\sum_{k=0}^{K} w_k e^{-\lambda s_k} \mathcal{T}_{s_k}\left(Q_t; \mathbf{c}_t, D_t, M_t^{\mathrm{road}}\right)}{\sum_{k=0}^{K} w_k e^{-\lambda s_k}}, \qquad s_k = k\Delta s.$$

$$\mathcal{D}_{\mathrm{cal}} = \{(\mathbf{C}_t^{(r)}, U_{H,t}^{*,(r)}, Q_{t,H}^{(r)}) : r \in \mathcal{R}_{\mathrm{cal}},\ t \in \mathcal{T}_r\}. \tag{11}$$

The 17-channel input in Eq. (12) comprises four multiscale source maps, transport and diffusivity, occupancy and road masks, nearest-agent radial bases, normalized coordinates, ego yaw rate, agent count, and a domain indicator. A 1 × 1 stem precedes six depthwise-separable residual blocks with dilations 1, 2, 4, 8, 16, and 1. The network has width 16, group normalization, and SiLU activations (Elfwing et al., 2018).

Centered spatial differences evaluate the finite-horizon physics residual in Eq. (13). The residual is computed in physical units and scaled by the calibration source magnitude:

Just Accepted

$$\mathcal{F}_\theta = \mathcal{L}_t\widehat{U}_\theta + \lambda\widehat{U}_\theta - \frac{Q_t - e^{-\lambda H}Q_{t,H}}{Z_H}. \quad (13)$$

The objective in Eq. (14) combines teacher-field matching, gradient-magnitude and direction matching, and a masked physics residual. Hotspot weighting limits the contribution of background cells, while the active-gradient quantile and stabilizing constant suppress numerically negligible directions.

$$\mathcal{L}_{\text{PINN}} = \mathcal{L}_{\text{field}} + 0.30\mathcal{L}_{\text{grad}} + 0.002\mathcal{L}_{\text{phys}}.$$
$$\mathcal{L}_{\text{field}} = \frac{\sum_{i=1}^{N} w_i \text{SL1}(\widehat{U}_{\theta,i} - U_i^*)}{\sum_{i=1}^{N} w_i}.$$
$$m_{\theta,i} = \| \nabla\widehat{U}_{\theta,i} \|_2, \qquad m_i^* = \| \nabla U_i^* \|_2.$$
$$\mathcal{L}_{\text{grad}} = \text{SL1}\left(\frac{\nabla\widehat{U}_\theta - \nabla U^*}{g_q + \varepsilon_g}\right) + \lambda_{\text{dir}}\mathcal{L}_{\text{dir}}.$$
$$\mathcal{L}_{\text{dir}} = 1 - \frac{\nabla\widehat{U}_\theta \cdot \nabla U^*}{(m_\theta + \varepsilon_g)(m^* + \varepsilon_g)}.$$
$$\mathcal{L}_{\text{phys}} = \frac{\sum_{i=1}^{N} M_{t,i}^{\text{road}} \text{SL1}(\mathcal{F}_{\theta,i}/s_{d,Q})}{\sum_{i=1}^{N} M_{t,i}^{\text{road}}}.$$
$$w_i = 1 + 4\text{clip}(U_i^*/0.05, 0, 1).$$
$$g_q = \text{quantile}(\{m_i^*\}_{i=1}^N, 0.85), \qquad \varepsilon_g = 10^{-6}. \quad (14)$$

4.3. PINN validation

Validation measures raw-field root mean square error (RMSE), active-gradient angular error, equal-mass hotspot intersection over union (IoU), and policy action deviation (Eq. 15). Active cells exceed the framewise 75th percentile of teacher-gradient magnitude. The hotspot threshold is the larger of the teacher field's framewise 90th percentile and 0.05 times its maximum. The predicted hotspot set has equal cardinality:

$$\varepsilon_{\text{RMSE}} = \left[\frac{1}{N}\sum_{i=1}^{N}\left(\widehat{U}_{\theta,i} - U_i^*\right)^2\right]^{1/2}.$$
$$\varepsilon_\angle = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \cos^{-1}\left(\frac{\nabla\widehat{U}_{\theta,i} \cdot \nabla U_i^*}{(m_{\theta,i} + \varepsilon_g)(m_i^* + \varepsilon_g)}\right).$$
$$\text{IoU}_{\text{hot}} = \frac{|\mathcal{H}_\theta \cap \mathcal{H}^*|}{|\mathcal{H}_\theta \cup \mathcal{H}^*|}.$$
$$\Delta\mathbf{u}_t = \pi_\phi([\mathbf{o}_t^{\text{env}}; \widehat{\boldsymbol{\eta}}_t]) - \pi_\phi([\mathbf{o}_t^{\text{env}}; \boldsymbol{\eta}_t^*]). \quad (15)$$

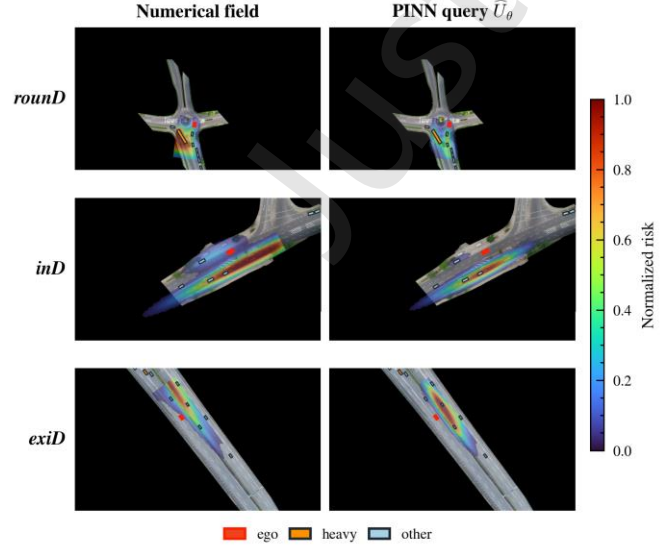


Fig. 3. Naturalistic rounD, inD, and exiD replays with field overlay: numerical field (left column) and context-conditioned PINN (right column).

Prospective relevance is assessed by comparing current-state fields with future occupancy in inD recordings (Fig. 3). The instantaneous source achieves an area under the receiver operating characteristic curve (AUROC) of 0.8249 and an area under the precision-recall curve (AUPRC) of 0.4167. Propagation raises these values to 0.9096/0.5836 for the numerical field and 0.8657/0.5143 for the frozen PINN. Relative to the source, the PINN gains are 0.0408 (95% confidence interval (CI) [0.0190,0.0576]) and 0.0976 (95% CI [0.0720,0.1207]).

Changing the field backend in the final term of Eq. (15) tests downstream sensitivity. Ten synchronized merging replays yield descriptor RMSE 0.0047 with zero collisions for both backends. Cumulative progress differs by -0.061 m (95% CI $[-0.076, -0.046]$) over approximately 586 m. Including context construction and device transfer, latency is 6.97 ms for the PINN and 3.88 ms for the numerical backend. These results characterize backend agreement in merging. Table 2 reports fidelity across the other geometries.

4.4. Field-enhanced RL objective formulation

The field-enhanced RL formulation is discussed through the following hierarchy. Firstly, Eq. (16) defines total, mean, and maximum field exposure over neighboring-vehicle footprints, together with pairwise dispersion. These quantities diagnose exposure and its allocation across nearby vehicles:

$$E_t^{\text{sv}} = \sum_{j=1}^{n_t} \int_{\mathcal{B}_{j,t}} \bar{U}_t(\boldsymbol{\xi})\, d\boldsymbol{\xi}.$$
$$E_{j,t} = \int_{\mathcal{B}_{j,t}} \bar{U}_t(\boldsymbol{\xi})\, d\boldsymbol{\xi}.$$
$$E_t^{\text{mean}} = \frac{1}{n_t + \varepsilon}\sum_{j=1}^{n_t} E_{j,t}, \qquad E_t^{\text{max}} = \max_j E_{j,t}.$$
$$D_t^{\text{sv}} = \frac{1}{n_t(n_t - 1) + \varepsilon}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\left|E_{i,t} - E_{j,t}\right|. \quad (16)$$

Upstream disturbance is measured in Eq. (17) by weighting the clipped field behind the ego by positive follower closing speed. Backward pressure increases as a faster follower approaches a high-pressure ego configuration and is zero when no follower approaches:

$$B_t = \sum_{j:\xi_{x,j}<0} \bar{U}_t(\xi_{x,j}, \xi_{y,j})\left[v_{j,t} - v_t^e\right]_+. \quad (17)$$

Policy optimization in Eq. (18) retains the environment's route-progress reward and termination logic. Risk, comfort, and social costs vary across profiles while traffic dynamics remain fixed:

$$J(\pi_\phi) = \mathbb{E}_{\pi_\phi}[\sum_{t=0}^{T-1}\gamma^t r_t].$$
$$r_t = r_t^{\text{env}} - C_t^{\text{risk}} - C_t^{\text{comf}} - C_t^{\text{soc}}. \quad (18)$$

The cost formulation given the field querying is formulated in Eq. (19). We combine ego risk, comfort, and social-externality penalties, including bounded braking and rear-TTC terms. Look-ahead and adjacent-lane queries supply observation features as:

$$C_t^{\text{risk}} = \lambda_R \text{clip}(\bar{U}_t(0,0), 0, U_{\max}).$$
$$C_t^{\text{comf}} = \lambda_u \| \Delta\mathbf{u}_t \|_1 + \lambda_j|j_t| + \lambda_\delta|\delta_t| + \lambda_{\Delta\delta}|\Delta\delta_t| + \lambda_{\Delta p}|\Delta p_t|.$$
$$C_t^{\text{soc}} = w_h\varphi(d_t^e) + w_c\varphi(d_t^f) + w_T\ell_t^{\text{rear}} + w_B\frac{B_t}{B_t + b_0}.$$
$$\varphi(d; d_0, d_1) = \text{clip}\left(\frac{d - d_0}{d_1 - d_0}, 0, 1\right)^2.$$
$$\ell_t^{\text{rear}} = \begin{cases}\text{clip}\left((T_{\text{safe}} - T_t^{\text{rear}})/T_{\text{safe}}, 0, 1\right), & T_t^{\text{rear}} < \infty, \\ 0, & \text{otherwise.}\end{cases} \quad (19)$$

Besides, deceleration is measured for the ego and nearest

Just Accepted

in-lane follower. A separate constant controls backward-pressure saturation, and all weights are nonnegative. The complete profile sets the common command-variation weight to zero.

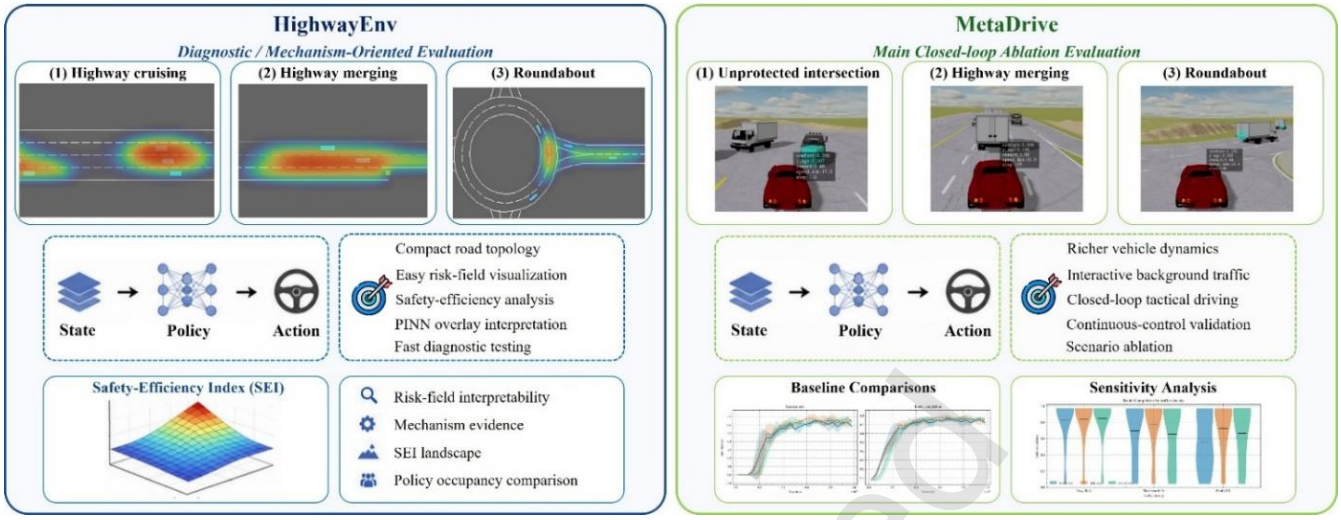


Fig. 4. Experimental design

The factorial in Eq. (20) crosses observation type with objective profile and defines three contrasts before evaluation. The complete profile jointly changes the risk coefficient and introduces comfort and externality costs, so its contrast measures the combined system effect:

$$O \in \{N, F\}, \qquad R \in \{N, R, S\}, \qquad \pi_\phi^{(O,R)}.$$

$$\Delta_{\text{repr}} = M(F,N) - M(N,N), \qquad \Delta_{\text{risk}} = M(F,R) - M(F,N).$$

$$\Delta_{\text{complete}} = M(F,S) - M(N,N).$$

$$r_t^{(q)} = \begin{cases} r_t^{\text{env}}, & q = N, \\ r_t^{\text{env}} - C_t^{\text{risk}}, & q = R, \\ r_t^{\text{env}} - C_t^{\text{risk}} - C_t^{\text{comf}} - C_t^{\text{soc}}, & q = S. \end{cases} \tag{20}$$

Each decision begins with a field query, whose descriptor is appended in field-observation cells. The primary MetaDrive factorial uses the numerical field, while PINN is evaluated through HighwayEnv queries with different highway scenarios (discussed in Section. 5).

5 Experiments and results

5.1. Implementation details

Two closed-loop simulators serve complementary roles (Fig. 4). Experiments use Python, Gymnasium, and Stable-Baselines3 v2.8.0 (Raffin et al., 2021). HighwayEnv provides highway cruising, ramp merging, unprotected-intersection, and roundabout scenarios. Cruising uses computationally accelerated highway environment; other policies use their native environments. Unless stated otherwise, medium traffic has 20 surrounding vehicles, density coefficient 1.0, initial ego-spacing factor 1.5, and a 30-s horizon. Target speeds are sampled uniformly from 18–30 m/s with Gaussian noise of standard deviation 1.5 m/s. Background traffic uses the intelligent driver model (IDM; Treiber et al., 2000) and Minimizing Overall Braking Induced by Lane Change (MOBIL; Kesting et al., 2007). MetaDrive supplies procedurally generated merging, intersection, and roundabout maps with continuous vehicle dynamics. Its reactive IDM traffic combines route following, lane selection, proportional-integral-derivative (PID) steering, and car-following acceleration. Tests use three maps at density 0.3 and a 500-step horizon. The primary 2 × 3 SAC factorial uses the numerical field, with architecture, action interface, training budget, protocol, and seeds matched across conditions (Section 4.4). In both simulators, ego actions affect subsequent traffic states, observations, rewards, and termination.

All 90 MetaDrive runs use a [256, 256] multilayer perceptron (MLP), discount factor 0.99, learning rate 3×10^{-4}, replay capacity 2×10^5, batch size 256, and target-update coefficient 0.005. Learning begins after 1,000 transitions. Comfort and hard-deceleration thresholds are 2 and 4 m/s² for the ego and 2 and 3 m/s² for the follower; rear TTC is 3 s, and b_0 = 1. Before confirmation, the merging screen selects the smallest tested positive comfort multiplier meeting a success non-inferiority margin of 0.10 against baseline SAC and a route-completion floor of 0.70.

5.2. Evaluation metrics

Success rate and ***route completion*** measure parts of progress. The first term shows whether the task was completed while the later indicator keeps track of episodes that ended before reaching the destination. ***Collision*** shows how safe driving is and ***mean absolute jerk*** shows how smooth the movement is. We also report exposure to the modeled interaction-pressure field. For the time-normalized exposure in Table 3 the clipped field value at the position is averaged for each episode and that average is divided by how long the episode lasted. At the vehicle-pair level the efficiency index rewards when the ego and leader speeds match and when spacing is balanced. The social-efficiency index adds to this by considering closing conditions: when the ego gets close to a leader a smaller TTC makes the score lower. Eq. (21) defines both indexes and averages the social-efficiency index, over pairs, time and episodes.

Just Accepted

$$\mathrm{EI}_{EL}(t) = \left[1 - \left(\frac{s_E}{s_L} - 1\right)^2\right] e^{-|k_{EL} - k_{FE}|}.$$

$$k_{EL} = \frac{d_{EL}}{d_{EL}+d_{FE}}, \qquad k_{FE} = \frac{d_{FE}}{d_{EL}+d_{FE}}.$$

$$\mathrm{TTC}_{EL}(t) = \frac{d_{EL}}{s_E - s_L}, \qquad s_E > s_L.$$

$$\mathrm{SEI}_{EL}(t) = \begin{cases} \mathrm{EI}_{EL}(t)\left[1 - e^{-\mathrm{TTC}_{EL}(t)/\tau_{\mathrm{TTC}}}\right], & s_E > s_L, \\ \mathrm{EI}_{EL}(t), & s_E \le s_L, \end{cases} \quad \tau_{\mathrm{TTC}} = 1\ \mathrm{s}.$$

$$\mathrm{SEI} = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{T_i}\sum_{t=1}^{T_i} \frac{1}{|\mathcal{P}_{i,t}|}\sum_{(E,L)\in\mathcal{P}_{i,t}} \mathrm{SEI}_{EL}(t). \tag{21}$$

External effects are assessed separately from the quantities used to diagnose the reward. Field exposure, follower deceleration, backward pressure, and the social composite describe mechanisms within the formulation. The reward-independent outcomes are computed from rollout kinematics and are not directly included as the specified training costs. Eq. (22) defines anisotropic interaction exposure, normalized traffic-speed deficit, and the paired SVO vector (Schwarting et al., 2019).

$$C_{\mathrm{int}} = \frac{1}{K}\sum_{t=1}^{K} \frac{1}{n_t}\sum_{j=1}^{n_t} c_{j,t}.$$

$$c_{j,t} = e^{-d_{j,t}^{\mathrm{clr}}/\lambda_d}\left(1 + \frac{v_{j,t}^{\mathrm{cl}}}{v_0}\right)\left[\alpha + (1-\alpha)\frac{1+\cos\beta_{j,t}}{2}\right].$$

$$D_v = \frac{\sum_{t=1}^{K}\sum_{j=1}^{n_t} \max(0, v_f - v_{j,t}) v_f^{-1} \Delta t}{\sum_{t=1}^{K} n_t \Delta t}.$$

$$\Delta U_{\mathrm{ego}} = \mathrm{RC}_c - \mathrm{RC}_b.$$

$$\Delta U_{\mathrm{oth}} = -\frac{D_v^c - D_v^b}{\max(|D_v^c|, |D_v^b|, \varepsilon)}.$$

$$\psi_{\mathrm{SVO}} = \mathrm{atan2}\left(\Delta U_{\mathrm{oth}}, \Delta U_{\mathrm{ego}}\right). \tag{22}$$

The SVO vector summarizes changes in ego route completion and surrounding-traffic speed deficit relative to baseline. Baseline SAC uses native observations and the native objective.

5.3. Field structure and qualitative interpretation

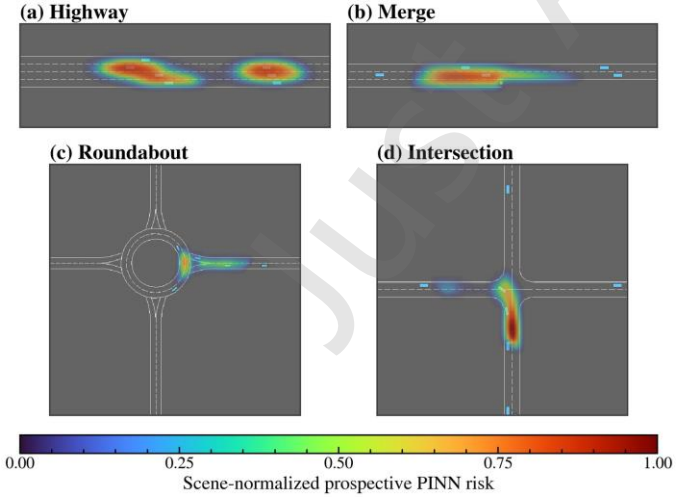


Fig. 5. Prospective PINN field in HighwayEnv replays: (a) highway, (b) merging, (c) roundabout, and (d) intersection. Blue-to-red denotes scene-normalized pressure.

At ramp entries, the frozen context-conditioned PINN places pressure where the mainline and merging streams approach one another, so their modeled influence overlaps before geometric contact (Fig. 5(b)). Highway fields instead show elongated lobes along the travel direction. At roundabouts, pressure follows the entry conflict, whereas intersection fields concentrate along crossing paths. These patterns describe how the model represents the current scene over its prospective horizon. The panels in Fig. 5 use separate color scales, with road masking and interpolation applied for display.

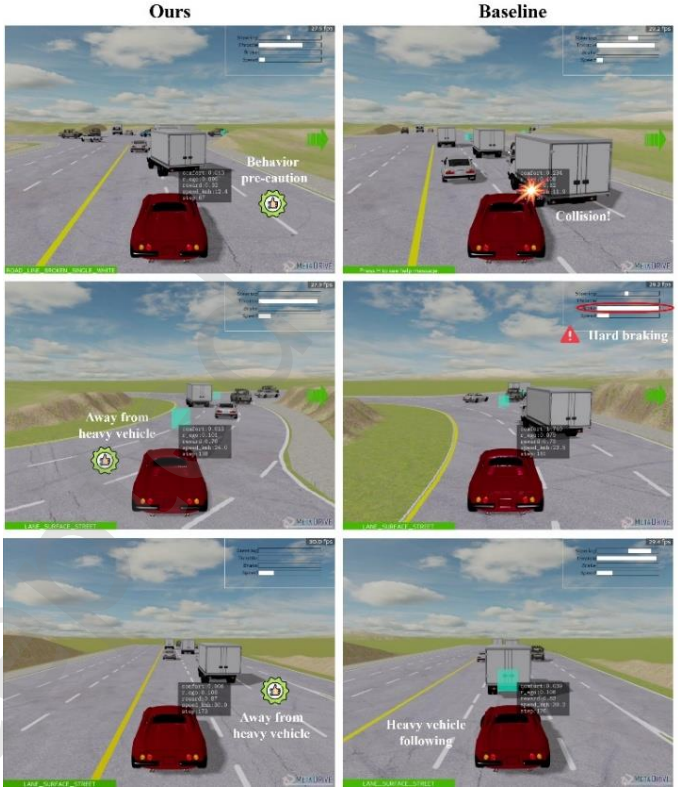


Fig. 7. MetaDrive replays of complete-profile SAC (social-SAC) and baseline SAC. White labels mark behavior.

PINN followed the spatial variation of the numerical reference closely in highway and merging scenes, with correlations of 0.969 and 0.977, respectively (Table 2). Agreement was slightly weaker at intersections (0.851) and roundabouts (0.760). Although the pooled correlation reached 0.925 across 271 simulator frames, the local-field errors differed considerably between geometries. Active-gradient angular error was 7.4° in merging but 33.5° at the roundabout, where hotspot IoU was only 0.165. The roundabout results therefore leave substantial uncertainty about the accuracy of the local pressure features supplied to the policy, despite the high pooled correlation.

The selected MetaDrive replays showed differences in gap selection and following behavior (Fig. 7). Across the illustrated highway-cruising, roundabout, and late-stage merging episodes (rows 1–3), complete-profile SAC opened an approach gap before potential conflict and maintained greater separation from heavy vehicles.

Just Accepted

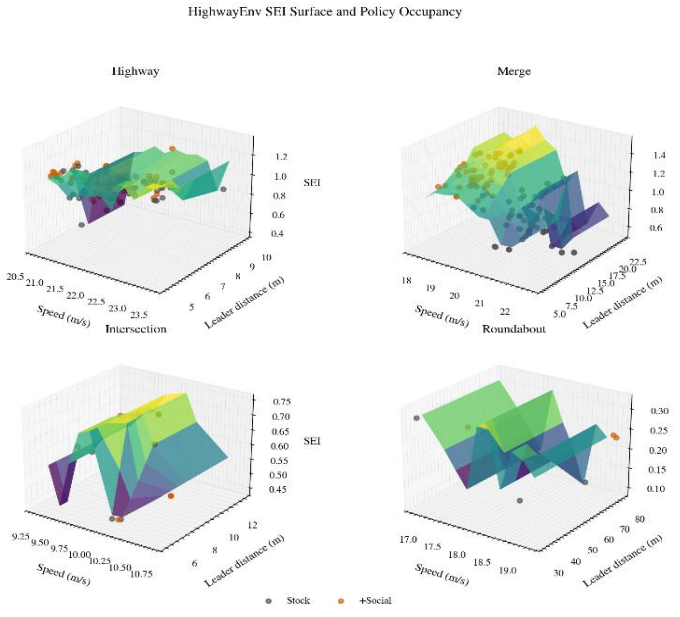


Fig. 6. Safety-efficiency index landscape and policy occupancy in HighwayEnv.

The corresponding baseline episodes included a collision, hard braking, or close following. These observations describe the displayed episodes; the held-out comparisons in Sections 5.4–5.5 provide the evidence for aggregate changes in safety, comfort, and surrounding traffic.

Table 2. Teacher-surrogate fidelity and isolated forward/propagation latency.

Domain (n)	RMSE	Corr.	Grad. angle	Hotspot IoU	PINN / Num. (ms)
inD (300)	2.225	0.800	34.9°	0.564	2.44 / 2.65
Highway (93)	0.029	0.969	9.1°	0.833	2.28 / 2.80
Merging (56)	0.022	0.977	7.4°	0.854	2.28 / 2.85
Intersection (29)	0.039	0.851	21.7°	0.609	2.70 / 2.61
Roundabout (93)	0.009	0.760	33.5°	0.165	2.38 / 2.51

Table 3. Paired effects for controlled continuous-SAC contrasts (candidate minus reference).

Scenarios	Paired effects	ΔSR	ΔRC	ΔCR	ΔE_R	$\Delta\lvert j\rvert$	ΔC_{int}	ΔD_v
Merging	Field representation	-0.072 [-0.180,0.012]	-0.037 [-0.105,0.020]	0.016 [-0.028,0.064]	-0.0003 [-0.0022,0.0015]	-0.79 [-2.82,0.94]	0.0008 [-0.0026,0.0044]	0.0032 [-0.0059,0.0117]
	Risk objective	-0.012 [-0.176,0.104]	0.012 [-0.074,0.088]	0.056 [-0.036,0.200]	0.0013 [-0.0002,0.0027]	-0.40 [-0.83,0.10]	0.0006 [-0.0010,0.0024]	0.0016 [-0.0099,0.0175]
	Complete profile (calibrated)	-0.020 [-0.084, 0.036]	-0.012 [-0.053, 0.027]	+0.016 [-0.028, 0.068]	+0.0003 [-0.0012, 0.0018]	-0.69 [-2.63, 1.12]	+0.0002 [-0.0023, 0.0026]	+0.0022 [-0.0067, 0.0112]
Intersection	Field representation	-0.028 [-0.116,0.056]	-0.008 [-0.050,0.030]	-0.028 [-0.092,0.040]	0.0001 [-0.0010,0.0012]	-0.32 [-1.02,0.36]	-0.0011 [-0.0039,0.0011]	-0.0031 [-0.0092,0.0033]
	Risk objective	0.080 [-0.068,0.200]	0.036 [-0.047,0.102]	-0.016 [-0.124,0.120]	-0.0007 [-0.0017,0.0003]	0.20 [-0.56,1.04]	0.0008 [-0.0025,0.0051]	0.0014 [-0.0038,0.0061]
	Complete profile	0.048 [-0.052,0.148]	0.021 [-0.041,0.083]	**-0.076 [-0.140,-0.016]**	**-0.0024 [-0.0040,-0.0008]**	**-0.98 [-1.76,-0.28]**	-0.0017 [-0.0048,0.0011]	**-0.0040 [-0.0101,-0.0002]**
Roundabout	Field representation	0.012 [-0.100,0.140]	0.020 [-0.031,0.074]	-0.064 [-0.148,0.008]	**-0.0027 [-0.0051,-0.0001]**	**-0.96 [-1.62,-0.41]**	**-0.0016 [-0.0028,-0.0006]**	-0.0052 [-0.0140,0.0002]
	Risk objective	0.064 [-0.008,0.148]	0.023 [-0.005,0.051]	0.000 [-0.032,0.028]	0.0015 [-0.0009,0.0043]	0.46 [-0.10,1.17]	0.0008 [-0.0003,0.0019]	-0.0012 [-0.0055,0.0010]
	Complete profile	0.004 [-0.144,0.152]	0.008 [-0.064,0.079]	-0.032 [-0.120,0.036]	-0.0008 [-0.0038,0.0025]	**-1.54 [-2.41,-0.54]**	-0.0002 [-0.0013,0.0008]	-0.0047 [-0.0150,0.0014]

Paired changes are reported with hierarchical 95% confidence intervals for success rate (SR), route completion (RC), collision rate (CR), time-normalized field exposure, mean absolute jerk, anisotropic interaction exposure, and traffic-speed deficit. The last two metrics are reward-independent. Each factorial cell uses five training seeds and 50 matched test seeds per policy.

5.4. Controlled ablation and held-out performance

Just Accepted

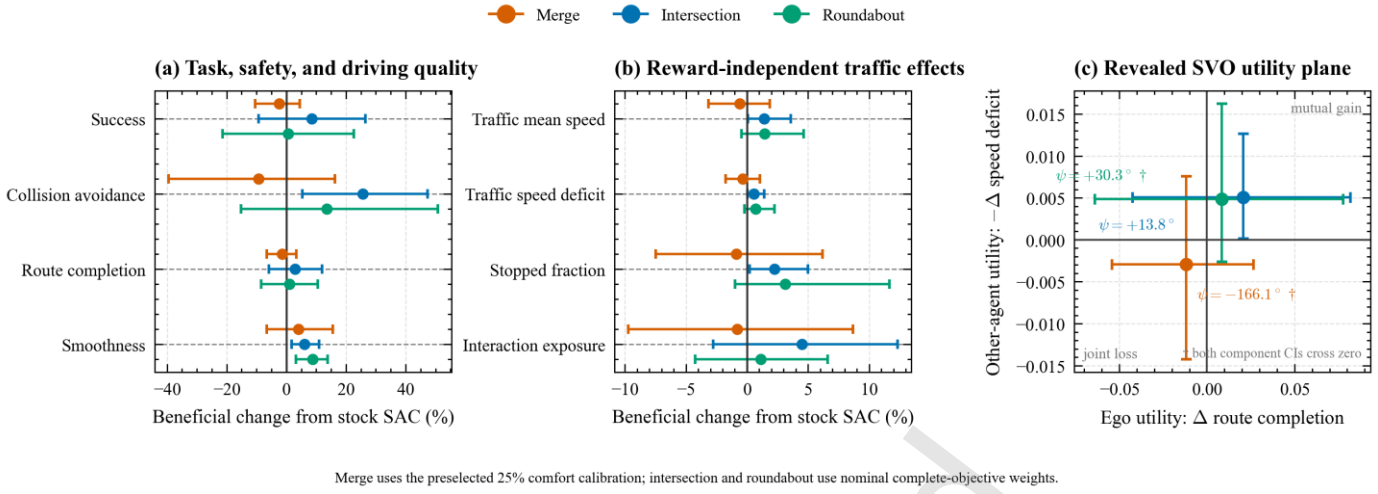


Fig. 8. Complete-profile outcomes relative to baseline SAC: (a) task, safety, route completion, and smoothness; (b) reward-independent kinematic traffic responses; and (c) social-value-orientation utility vectors. Points show paired means and hierarchical 95% confidence intervals over five training seeds and 50 matched traffic seeds.

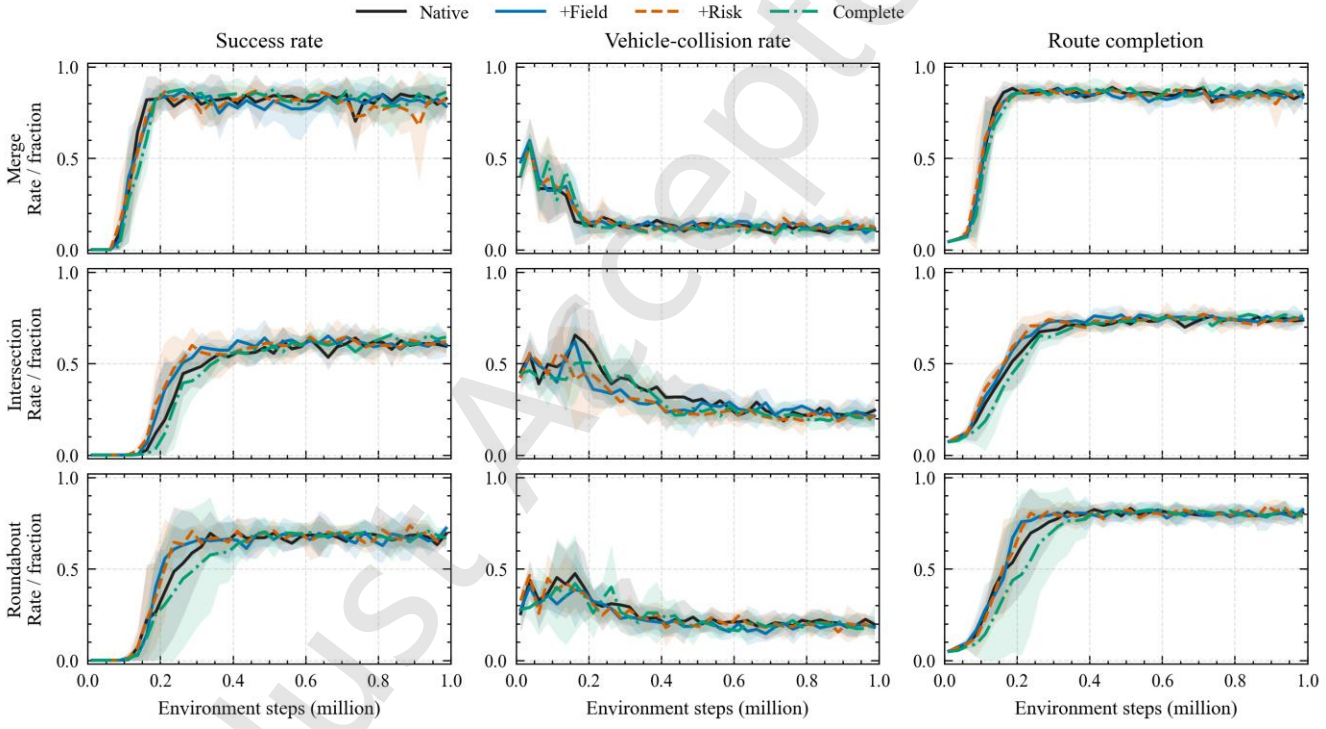


Fig. 9. Reward-independent SAC training trajectories. Rows show merging, intersection, and roundabout; columns show success, vehicle collision, and route completion. Curves are means with 95% confidence intervals across five training seeds. Merging Complete uses the preselected 25% comfort multiplier; other scenarios use nominal weights.

Statistics report that the complete profile reduced intersection collision rate from 0.296 to 0.220 relative to baseline SAC (paired $\Delta = -0.076$, 95% CI [−0.140, −0.016]). Mean absolute jerk also decreased ($\Delta = -0.98$ m/s³, 95% CI [−1.76, −0.28]). A smoothness benefit was evident at the roundabout, where the corresponding jerk change was −1.54 m/s³ (95% CI [−2.41, −0.54]). The task outcomes were less conclusive. Intersection success increased from 0.560 to 0.608, with a route-completion change of +0.021. Roundabout success was 0.676 compared with 0.672 for baseline SAC, and route completion changed by +0.008. In merging, the calibrated profile achieved success of 0.788 versus 0.808 and route completion of 0.798 versus 0.809. The success and route-completion intervals spanned zero in all three scenarios. For merging, this also held for every other reported complete-profile effect in Table 3.

To examine the contribution of the field observations, we compared policies trained with the native objective, keeping the reward unchanged. At the roundabout, field augmentation reduced mean absolute jerk by 0.96 m/s³ (95% CI [0.41, 1.62]) and kinematic interaction exposure by 0.0016 (95% CI [0.0006, 0.0028]). Time-normalized field exposure decreased by 0.0027 (95% CI [0.0001, 0.0051]). These reductions show that the representation affected behavior in this setting without the additional reward terms.

Adding the risk cost with field observations held fixed

Just Accepted

produced no resolved change in the reported outcomes: all risk-only intervals in Table 3 included zero. The complete-profile comparison estimates the joint effect of the representation and the risk, comfort, and externality costs specified in Section 4.4. Attribution of its gains to an individual cost term would require a further controlled comparison.

RL training results are plotted in Fig. 9. SAC profiles reach late-training plateaus after about 0.4 million transitions. In the final 100,000-step window, complete-profile success is 0.841, 0.636, and 0.685 for merging, intersection, and roundabout, respectively; native-policy values are 0.823, 0.604, and 0.671. Intersection and roundabout plateaus are lower across profiles.

Table 4. Late-training reward-cost decomposition in the merging calibration screen.

Setting	Risk	Jerk	Steering	Throttle	Hard brake	Externality	Jerk share (%)
Original comfort 1x	1.18	13.66	0.47	0.10	0.46	0.24	94.2
Comfort 0x	2.05	0.00	0.00	0.00	0.89	0.01	0.0
Comfort 0.25x	2.03	7.13	0.11	0.05	1.29	0.10	97.8
Comfort 0.5x	2.05	14.17	0.20	0.10	1.29	0.03	98.0
Social 0x, comfort 1x	1.64	21.96	0.32	0.16	0.00	0.00	97.2

Costs are reported in units of 10^{-3} reward per step over transitions 200,000–300,000 using three seeds. Section 5.5 reports the matched jerk-knockout intervention.

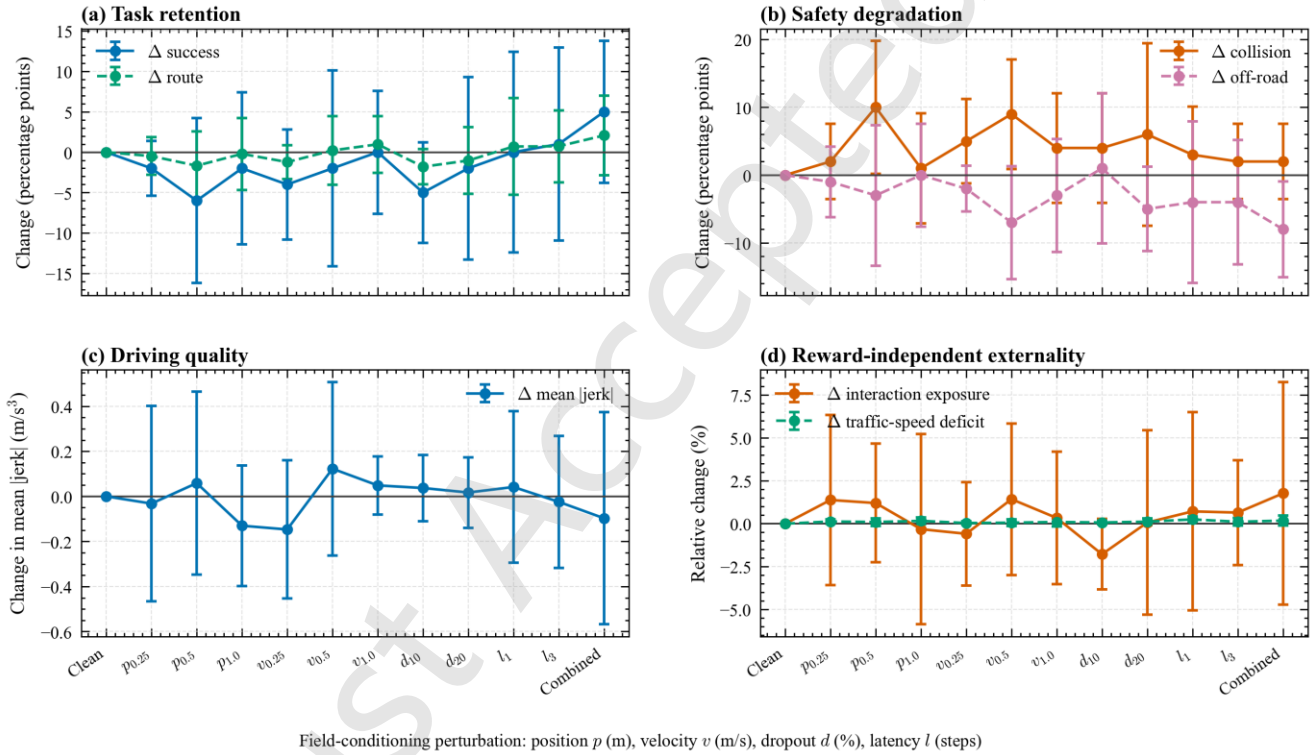


Fig. 10. Structured-state sensitivity of the field-enhanced SAC policy. Panels report paired changes relative to clean observations for task retention, safety, mean absolute jerk, and reward-independent externality under position noise p, velocity noise v, missed-detection rate d, observation latency l, and their combined moderate condition.

5.5. Social externality and reward calibration

At the intersection, the utility vector (+0.0206, +0.0051) gives an SVO angle of 13.8° (Fig. 8(c)). The positive other-agent component has an interval excluding zero, indicating a modest reduction in surrounding-traffic speed deficit relative to baseline under the normalization in Eq. (22). The other scenarios remain uncertain. Roundabout angle is 30.3°, but both component intervals include zero. Merging gives −166.1° with both components near the origin and an unstable direction. These angles summarize rollout effects rather than stable social preferences.

The nominal complete reward produced a low-motion policy in merging. Low progress persisted when the social cost group was removed at the original comfort weight, directing attention to the comfort penalties. In the validation-only sweep, comfort multipliers of 0.25 and 0.5 yielded success rates of 0.800 and 0.822 after 0.3 million transitions. Jerk accounted for 94%–98% of the logged comfort cost (Table 4). We therefore examined its contribution through a matched intervention with the jerk coefficient set to zero.

Removing this term changed both progress and driving quality. Route completion increased from 0.300 to 0.787, and success rose from 0.233 to 0.589. The same intervention increased collision rate from 0.133 to 0.311. Mean jerk rose by 5.79 m/s³, mean throttle-change rate increased from 0.279 to 1.391, and interaction exposure increased by 0.013. The two settings exposed opposite difficulties in the reward: the

Just Accepted

original jerk penalty suppressed progress, while removing it allowed more abrupt and collision-prone control. Calibration must account for both behaviors. The sweep brackets a behavioral change between comfort multipliers of 0.5 and 1.0, with the other reward coefficients fixed. Among the tested positive candidates, 0.25 is the smallest multiplier meeting the predeclared task criteria and is fixed before confirmation. Its suitability beyond this calibration setting remains untested.

5.6. Algorithm scope, computational cost, and structured-state sensitivity

The replicated comparisons support inference about continuous SAC under matched architecture, action interface, hyperparameters, training budget, and seeds. Historical cross-algorithm results are retained for descriptive context (Table 6). Those runs used unmatched tuning and mostly a single training seed, so learner family, action resolution, tuning budget, and seed variance remain confounded. Other policies can accept the descriptor, but their performance is not established by the replicated experiment.

The discretized steering-throttle interface considered here introduces a limit on action resolution. On a rectangular lattice, nearest-action projection error is bounded by half a cell diagonal, or 0.71 in normalized coordinates for a three-by-three grid. Small gradient changes can consequently leave the selected command unchanged until a quantization boundary is crossed. This bound characterizes action resolution; it does not identify the cause of the historical failures. Action-conditioned queries based on short-horizon rollouts and fail-open field-informed action masking remain possible extensions. Including context construction and device transfer, the synchronized merging comparison takes 6.97 ms with the PINN for online field modeling. In the MetaDrive evaluation path, policy inference takes 0.34–0.41 ms and field queries take 4.64–6.76 ms, giving total field-enhanced decision times of 4.98–7.19 ms (Table 5).

Additionally, we evaluated sensitivity to field-input errors using five independently trained intersection policies, twelve perturbation conditions, and 20 traffic seeds matched to the clean condition, totaling 1,200 episodes (Fig. 10). Gaussian position, velocity, and heading errors, Bernoulli detection dropout, and delayed states were introduced into the copied surrounding-agent states supplied to field construction. Position and velocity errors increased collision rate when tested individually. With 0.5 m position noise, the paired increase was 0.100 (95% CI [0.002, 0.198]); with 0.5 m/s velocity noise, it was 0.090 (95% CI [0.009, 0.171]). These results identify sensitivity to errors in two of the quantities used to construct the field. The combined moderate condition included both noise levels, along with 2° heading noise, 10% dropout and one-step latency. Success changed by 0.050 (95% CI [−0.038, 0.138]) with route completion by 0.021 (95% CI [−0.028, 0.070]). The intervals covered both improvement and deterioration, making the task-performance changes unclear. Thus, perception-in-the-loop performance and actuation uncertainty still need more robustness testing for future real-world deployment.

Table 6. Historical intersection comparison stratified by action interface.

Algorithm / action	Learning rate	Batch	Success native / full	Route native / full
DQN / D9	1×10^{-4}	64	0.00 / 0.00	0.206 / 0.180
PPO / C	3×10^{-4}	64	0.00 / 0.00	0.368 / 0.085
SAC / C	3×10^{-4}	256	0.60 / 0.55	0.692 / 0.651
TD3 / C	3×10^{-4}	256	0.60 / 0.45	0.678 / 0.690
DDPG / C	1×10^{-4}	256	0.55 / 0.60	0.678 / 0.724

C denotes continuous actions, and D9 denotes nine discrete commands (Values are reanalyzed from recorded episodes at density 0.3 with 20 common test seeds and one training seed per cell).

6 Conclusion

SAFE-AD makes finite-horizon spatial interaction pressure available to an RL policy through compact field descriptors queried by online PINN trained on historical traffic datasets. The controlled continuous-SAC comparisons distinguish changes in representation from changes in the objective function. With the propagated field querying as the risk assessment guidance, the complete profile reduces both collision rate and mean absolute jerk at intersections and reduces jerk at roundabouts. Field augmentation under the native objective also reduces field exposure, jerk, and kinematic interaction exposure at the roundabout. Success and route-completion improvements remain uncertain. Reward-independent traffic measures resolve a modest external benefit at the intersection, while effects in the other scenarios remain inconclusive.

Several shortcomings still exist. Our key findings are limited to the tested scenarios and replicated SAC setting. Meanwhile, historical comparisons cannot establish generalization across learners, and the complete-profile contrast cannot identify an independent contribution from the social penalty. Training plateaus describe empirical stability rather than guaranteed convergence, and the achieved task success does not establish deployment readiness. Besides, detector bias, sensor-level latency, and actuation uncertainty remain untested. Future work includes perception-in-the-loop validation, calibration across geometries, and end-to-end backend efficiency that connect to real-world deployment on physical platforms.

Just Accepted

Author contributions

Zian Wang: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Wenjie Huang:** Writing – review & editing, Formal analysis, Methodology. **Zejian Deng:** Validation, Formal analysis. **Yiming Shu:** Validation, Methodology. **Jiahui Xu:** Validation, Conceptualization. **Yong Wang:** Writing – review & editing, Methodology. **Shen Li:** Formal analysis, Validation. **Dongpu Cao:** Formal analysis, Methodology. **Chen Sun:** Writing – review & editing, Supervision, Formal analysis, Methodology.

Replication and data sharing

The source codes and replication package are available on ETS data at DOI: 10.26599/ETSD.2026.9190103. Supplementary materials could be accessed at: https://github.com/SAS-HKU/SAFE-AD.git.

Acknowledgements

This work was supported by the Research Grants Council of Hong Kong under the Early Career Scheme (Grant No. 27206525).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of generative AI and AI-assisted technologies in the writing process

During manuscript preparation, the authors used generative AI for grammar checking and suggestions on language consistency. The authors reviewed and edited the output and take full responsibility for the published content.

References

Al-Sharman, M., Edes, L., Sun, B., Jayakumar, V., Tahir, H., Daoud, M.A., et al., 2025. Autonomous driving at unsignalized intersections: A review of decision-making challenges and reinforcement learning-based solutions. *IEEE Trans Autom Sci Eng*, **23**, 1434-1461.

Andreotti, E. and Selpi, 2025. New microscopic indicators for evaluating traffic efficiency and safety. *Transp Res Interdiscip Perspect*, **33**, 101569.

Bock, J., Krajewski, R., Moers, T., Runde, S., Vater, L. and Eckstein, L., 2020, October. The inD dataset: A drone dataset of naturalistic road user trajectories at German intersections. In: Proceedings of the IEEE Intelligent Vehicle Symposium (IV), 1929-1934.

Chen, X., Zang, J. and Zhao, F., 2025. Human-Like Autonomous Driving Car-Following Behavior Learning Based on Adversarial Training and Uncertainty-Aware DDPG. *IEEE Robot Autom Lett*, **11**, 1754-1761.

Deng, Z., Hu, W., Sun, C., Chu, D., Huang, T., Li, W., et al., 2024. Eliminating uncertainty of driver's social preferences for lane change decision-making in realistic simulation environment. *IEEE Trans Intell Transp Syst*, **26**, 1583-1597.

Elfwing, S., Uchibe, E. and Doya, K., 2018. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw*, **107**, 3-11.

Gao, Z., Hao, H., Gao, F., Zhao, R., 2025. Behavior-constrained multi-agent proximal policy optimization for cooperative vehicle control at future intersections. *Automot Innov*, **8**, 896–912.

Gros, S. and Zanon, M., 2020. Data-driven economic NMPC using reinforcement learning. *IEEE Trans Autom Control*, **65**, 636–648.

Guo, J., Zhu, H., Yang, Y. and Guo, C., 2025. Advances in physics-informed neural networks for solving complex partial differential equations and their engineering applications: A systematic review. *Eng Appl Artif Intell*, **161**, 112044.

Han, J., Zhao, J., Zhu, B. and Song, D., 2023. Spatial-temporal risk field for intelligent connected vehicle in dynamic traffic and application in trajectory planning. *IEEE Trans Intell Transp Syst*, **24**, 2963–2975.

Hu, C., Liu, D., Wang, Z., Li, D., Wang, J. and Tang, X., 2025. Risk-aware reinforcement learning for non-conservative motion planning in uncertain autonomous driving environments. *IEEE Trans Intell Transp Syst*, **26**, 18880 - 18893.

Hu, W., Deng, Z., Yang, Y., Zhang, P., Cao, K., Chu, D., et al., 2025. Socially game-theoretic lane-change for autonomous heavy vehicle based on asymmetric driving aggressiveness. *IEEE Trans Veh Technol*, **74**, 17005-17018.

Huang, W. and Haskell, W.B., 2017, December. Risk-aware Q-learning for Markov decision processes. In: Proceedings of the IEEE Conference on Decision and Control (CDC), 4928-4933.

Huang, W. and Haskell, W.B., 2020. Stochastic approximation for risk-aware Markov decision processes. *IEEE Trans Autom Control*, **66**, 1314-1320.

Huang, W., Pham, V.H. and Haskell, W.B., 2020, April. Model and reinforcement learning for markov games with risk preferences. In: Proceedings of the AAAI Annual Conference on Artificial Intelligence, **34**, 2022-2029.

Jia, J., Su, Y., Bao, Z., Hong, Y., Gao, B., Gan, Z., et al., 2026. Learning a Unified Risk Map for Autonomous Driving in Partially Observable Environments. *IEEE Robot Autom Lett*, **11**, 4457-4464.

Jiang, X., Li, M. and Liu, Q., 2026. Joint longitudinal-lateral trajectory planning for CAVs in mixed traffic at signalized intersections. *Commun Transp Res*, **6**, 9640011.

Jin, G.Z., Li, Z.R., Leng, B., Yu, R., Xiong, L., et al., 2025. Multi-timescale hierarchical reinforcement learning for unified behavior and control of autonomous driving. *IEEE Robot Autom Lett*, **10**, 12772–12779.

Kesting, A., Treiber, M. and Helbing, D., 2007. General lane-changing model MOBIL for car-following models. *Transp Res Rec*, **1999**, 86-94.

Kilic, K.I., Desoeuvres, A., Pedersen, C.B., Vasegaard, A.E. and Nielsen, P., 2026. Adaptive artificial potential field method for small autonomous vehicles. *Rob Auton Syst*, 105364.

Krajewski, R., Moers, T., Bock, J., Vater, L. and Eckstein, L., 2020, September. The rounD dataset: A drone dataset of road user trajectories at roundabouts in Germany. In: Proceedings of the IEEE 23rd International Conference on Intelligent

Just Accepted

Transportation Systems (ITSC), 1-6.
Leurent, E., 2018. An environment for autonomous driving decision-making. https://github.com/eleurent/highway-env.
Li, Q., Peng, Z., Feng, L., Zhang, Q., Xue, Z. and Zhou, B., 2022. MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Trans Pattern Anal Mach Intell*, **45**, 3461-3475.
Li, Z., Wang, Y., Zuo, Z., Zhao, R., Hu, C. and Shi, Y., 2026. Model Predictive Control-Based Trajectory Optimization for Autonomous Vehicles Using Risk-Aware Corridors. *IEEE Trans Intell Transp Syst*, **27**, 8070-8085
Liu, Q. and Wang, J., 2026. Aggressiveness-conservativeness trade-off in cooperative driving: An adaptive CBF approach under perception constraints. *Transp Res Part C Emerg Technol*, **183**, 105435.
Liu, Y., Zhou, A., Wang, Y., Peeta, S., 2026. Proactive longitudinal control to assist lane changes of human-driven vehicles in mixed traffic: Human-emulation approach. *Automot Innov*, **9**, 48–64.
Moers, T., Vater, L., Krajewski, R., Bock, J., Zlocki, A. and Eckstein, L., 2022, June. The exiD dataset: A real-world trajectory dataset of highly interactive highway scenarios in Germany. In: Proceedings of the IEEE Intelligent Vehicle Symposium (IV), 958-964.
Ning, M., Khajepour, A., Hashemi, E. and Sun, C., 2024. A novel motion planning for autonomous vehicles using point cloud based potential field. *IEEE Trans Veh Technol*, **74**, 3780-3792.
Qu, Y., Sheng, Z., Huang, Z., Chen, J., Luo, Y., Wang, T., et al., 2026. Found-RL: Foundation model-enhanced reinforcement learning via asynchronous VLM feedback for autonomous driving. *Commun Transp Res*, **6**, 9640027.
Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M. and Dormann, N., 2021. Stable-baselines3: Reliable reinforcement learning implementations. *J Mach Learn Res*, **22**, 1-8.
Schwarting, W., Pierson, A., Alonso-Mora, J., Karaman, S. and Rus, D., 2019. Social behavior for autonomous vehicles. *Proc Natl Acad Sci U S A.*, **116**, 24972-24978.
Sheng, Z., Huang, Z. and Chen, S., 2024. Traffic expertise meets residual RL: Knowledge-informed model-based residual reinforcement learning for CAV trajectory control. *Commun Transp Res*, **4**, 100142.
Sun, C., Zhang, R., Lu, Y., Cui, Y., Deng, Z., Cao, D., et al., 2023. Toward ensuring safety for autonomous driving perception: Standardization progress, research advances, and perspectives. *IEEE Trans Intell Transp Syst*, **25**, 3286-3304.
Treiber, M., Hennecke, A. and Helbing, D., 2000. Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E*, **62**, 1805.
Wang, Z., Shu, Y., Deng, Z. and Sun, C., 2026. DRIFT: Driving Risk Inference via Field Transmission for Human-like Autonomous Driving. https://doi.org/10.48550/arXiv.2605.27964.
Wang, Z., Shu, Y., Deng, Z., Cai, G., Xu, J., Tang, J., et al., 2026. DREAM: Defensive risk-aware enhanced maneuver planning for autonomous vehicles in heterogeneous traffic. *Transp Res Part C Emerg Technol*, **194**, 106062.
Xing, Y., Cai, M., Tu, Y., Wang, Y., Jiang, J., Meng, Q., et al., 2024. TREMP: Field-Based Traffic Risk Evaluation Considering Interactive Motion Prediction. In: Proceedings of the IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), 1551-1556.
Zhao, C. and Yu, H., 2025. Physics-Informed Neural Networks for Nonlocal Flow Modeling of Connected Automated Vehicles. https://doi.org/10.48550/arXiv.2506.18357.

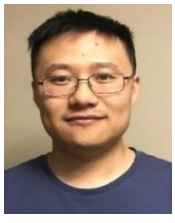
Author biography



Zian Wang received the B.Eng. degree in Electronic and Information Engineering from The Hong Kong Polytechnic University, in 2025. He is currently pursuing an M.Phil. degree in the Department of Data and Systems Engineering with The University of Hong Kong. His research interests include risk-aware planning and control in autonomous driving, robotics, and interdisciplinary topics within intelligent transportation systems.



Wenjie Huang received the Ph.D. degree in Industrial Systems Engineering and Management from the National University of Singapore in 2019, and the B.S. degree in Industrial Engineering and Management with a minor in Economics from Shanghai Jiao Tong University in 2014. He is currently a Research Assistant Professor with the Department of Data and Systems Engineering and the Musketeers Foundation Institute of Data Science at The University of Hong Kong. His research interests include risk analytics and robust optimization, sequential decision-making, and data science/AI for operations management.



Zejian Deng received the B.Sc. degree in automation and the M.A.Sc. degree in transportation engineering from Wuhan University of Technology, Wuhan, China, in 2015 and 2018, respectively, and the Ph.D. degree in mechanical and mechatronics engineering from the University of Waterloo, Waterloo, ON, Canada, in 2022. He is currently working as a Postdoctoral Fellow with the Department of Data and Systems Engineering, The University of Hong Kong. His research interests include driver model, driving safety validations, decision-making, planning, and control in automated driving.



Yiming Shu received the B.Eng. degree in Automotive Engineering from the Harbin Institute of Technology (Weihai), Weihai, China, in 2022, the M.Phil. degree in 2025 from The University of Hong Kong. She is currently pursuing her Ph.D. in the Department of Data and Systems Engineering at The University of Hong Kong. Her research interests include

Just Accepted

motion planning and decision-making of autonomous vehicles (AVs).



Jiahui Xu received her Bachelor's degree in Vehicle Engineering from Beijing Institute of Technology, China in 2022, and the Master's degree in 2025. She is currently pursuing her Ph.D. in the Department of Data and Systems Engineering at The University of Hong Kong. Her research interests include trajectory prediction, decision-making, and the safety of autonomous driving.



Yong Wang received the M.S. and Ph.D. degrees in mechanical engineering from Beijing Institute of Technology in 2021 and 2025, respectively. He was a Visiting Scholar with the Autonomous Vehicle Systems Laboratory, Technical University of Munich, and the Institute for AI Industry Research, Tsinghua University. He is currently a Post-Doctoral Researcher with The University of Hong Kong. His research focuses on intelligent electric vehicles with an emphasis on data-driven control and planning. He received the 2022 Outstanding Paper Award from IEEE Transactions on Industrial Informatics.



Shen Li received the Ph.D. degree from the University of Wisconsin–Madison in 2018. He is a Research Associate with Tsinghua University. His research focuses on intelligent transportation systems (ITS), architecture design of Connected Automated Vehicle Highway (CAVH) systems, vehicle infrastructure cooperative planning and decision methods, traffic data mining on cellular data, and traffic operations and management.



Dongpu Cao received the Ph.D. degree from Concordia University, Canada, in 2008. He is a Professor at State Key Laboratory of Intelligent Green Vehicle and Mobility, Tsinghua University. His current research focuses on driver cognition, automated driving and connected autonomous driving. He has contributed more than 200 papers and 3 books. He received the SAE Arch T. Colwell Merit Award in 2012, IEEE VTS 2020 Best Vehicular Electronics Paper Award and over 10 Best Paper Awards from international conferences.



Chen Sun received the Ph.D. degree in Mechanical & Mechatronics Engineering from University of Waterloo, ON, Canada in 2022, M.A.Sc degree in Electrical & Computer Engineering from University of Toronto, ON, Canada in 2017 and B.Eng. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. He is currently an Assistant Professor with the Department of Data and Systems Engineering, The University of Hong Kong. His research interests include field robotics, safe and trustworthy autonomous driving, and in general human-CPS autonomy.

Just Accepted