Safety-Critical Multi-Agent MCTS for Mixed Traffic Coordination at Unsignalized Intersections

Zhihao Lin[1], Jianglin Lan[1,§], Christos Anagnostopoulos[2], Zhen Tian[1], and David Flynn[1]

Abstract—**Decision making at unsignalized intersections presents significant challenges for autonomous vehicles (AVs), particularly in mixed traffic scenarios where both AVs and human-driven vehicles (HDVs) must safely coordinate their movements. This paper proposes a safety-critical multi-agent Monte Carlo tree search (MCTS) framework that integrates deterministic and probabilistic predictions to enable cooperative decision making in complex intersection scenarios. The framework incorporates three main innovations: 1) a safety assessment mechanism that systematically handles AV-to-AV (V2V), AV-to-HDV (V2H), and Vehicle-to-Road (V2R) interactions using dynamic safety thresholds and spatiotemporal risk metrics, 2) an adaptive HDV behavior awareness by combining the Intelligent Driver Model (IDM) with probabilistic distributions, and 3) a multi-objective reward function optimization approach that balances safety, efficiency, and cooperation. Extensive simulations demonstrate our framework's efficacy and superior capability in ensuring safe and efficient intersection navigation across the fully-autonomous scenario (100% AVs) and challenging mixed traffic scenario (50% AVs + 50% HDVs). Compared to benchmarks, our method reduces trajectory deviations by up to 37.56% in the fully-autonomous scenario and 62.43% in the mixed traffic scenario, while maintaining significantly lower Post-Encroachment Time (PET) violations (0% and 2.8%, respectively).**

Index Terms—**Autonomous vehicles, decision making, mixed traffic, Monte Carlo tree search, risk assessment**

I. INTRODUCTION

DECISION making at unsignalized intersections presents significant challenges for autonomous vehicles (AVs) [1], particularly in mixed traffic environments where both AVs and human-driven vehicles (HDVs) must safely coordinate their movements without traffic signal guidance [2]. The complexity arises from the need to handle multiple types of critical interactions simultaneously while ensuring both safety and efficiency at intersections, a highly dynamic environment [3]. This challenge becomes more pronounced as the interaction patterns among vehicles become more intricate, thus requiring a comprehensive understanding of both deterministic AV behaviors and uncertain HDV behaviors [4]. This work targets unsignalized intersections where traffic signals are impractical due to cost, infrastructure, or geographical constraints. Such intersections are widespread globally, particularly in rural or developing areas, highlighting the need for infrastructure-light algorithmic coordination solutions.

While many unsignalized intersections operate with explicit priority rules (e.g., stop signs and yield signs), several scenarios exist where such rules are absent, ambiguous, or insufficient for optimal traffic flow. These include unmarked rural intersections, temporary construction zones, low-volume residential areas, parking lots, and emergency situations [5]. In developing regions, intersections with unclear or poorly maintained signage are common. Even at intersections with established priority rules, human drivers' inconsistent interpretation creates ambiguity that AVs must safely navigate [6]. As vehicle autonomy advances, interest grows in cooperative intersection management approaches that may replace traditional priority rules with more efficient negotiation-based protocols, particularly with increasing AV penetration rates [7]. Our focus on unsignalized intersections without rigid priority assignments provides a challenging testing ground for autonomous decision-making algorithms, requiring vehicles to negotiate passage dynamically through implicit communication and behavior prediction rather than relying on predefined rules, thus ensuring robustness across diverse real-world scenarios.

Traditional approaches to intersection management often rely on rule-based methods, which attempt to generate conflict-free passage sequences through preset regulations [8]. Although computationally efficient, these methods struggle to capture the complex and diverse decision-making behaviors of human drivers [9]. The First In, First Out strategies [10] ensure safety by allowing only one vehicle to pass at a time, but significantly reduce the traffic efficiency [11]. More sophisticated rule-based approaches incorporating virtual rotation projection and conflict-free passage sequence trees have been proposed, but their effectiveness diminishes in scenarios involving HDVs with varying abilities and driving styles [12].

Recent advances in machine learning have revolutionized the approach to autonomous intersection management [13], [14]. Deep learning techniques, particularly those incorporating recurrent neural networks and graph neural networks, have demonstrated remarkable success in modeling the complex interdependencies between vehicles at intersections. Deep reinforcement learning (DRL) [15], [16] has shown particular promise in handling mixed traffic scenarios, with approaches such as multi-agent deep deterministic policy gradient [17] achieving notable improvements in safety and efficiency. However, these learning-based methods face challenges such as

This work was supported in part by the China Scholarship Council Ph.D. Scholarship for 2023-2027 (No.202206170011), in part by the Leverhulme Trust Early Career Fellowship (ECF-2021-517), in part by the EPSRC Digital Twinning Research Hub for Decarbonising Transport (TransiT, Ref: EP/Z533221/1), and in part by the EPSRC Research Hub for Decarbonised Adaptable and Resilient Transport Infrastructures (DARe, Ref: EP/Y024257/1).

[1] Zhihao Lin, Jianglin Lan, Zhen Tian and David Flynn are with the James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, United Kingdom.

[2] Christos Anagnostopoulos is with the School of Computing Science, University of Glasgow, Glasgow G12 8QQ, United Kingdom.

[§] Corresponding author: Jianglin Lan (e-mail: jianglin.lan@glasgow.ac.uk)

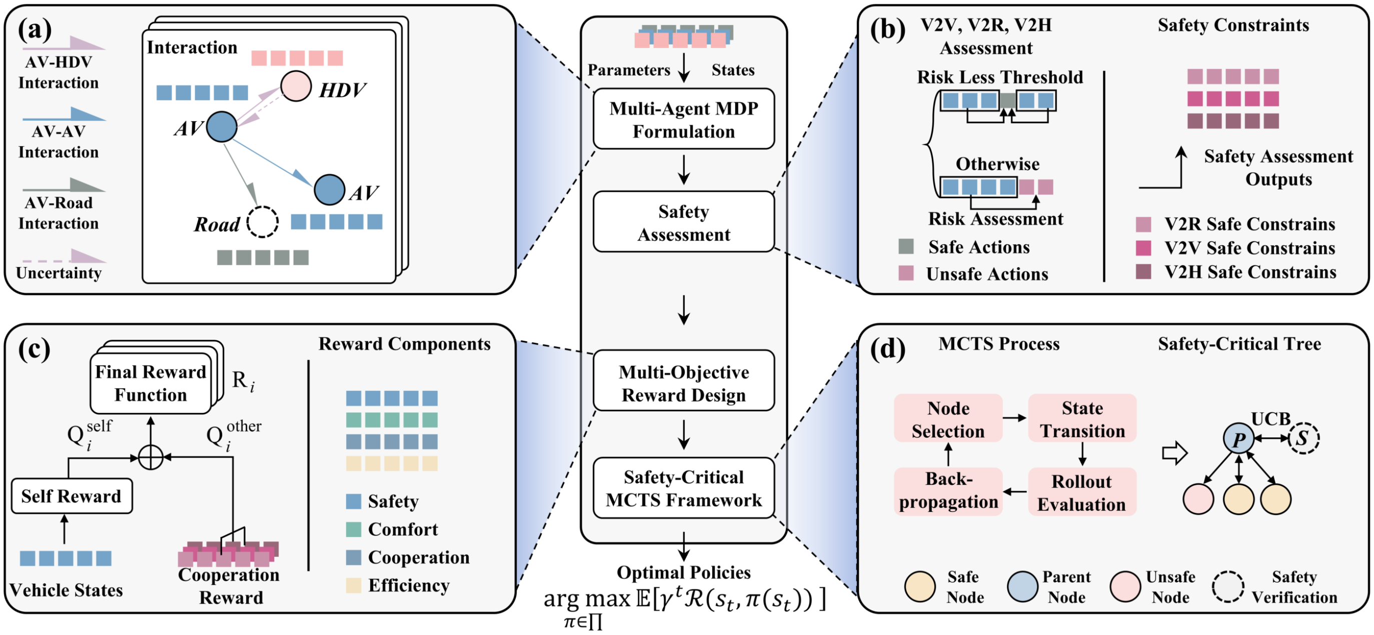


Fig. 1: Overview of the safety-critical decision-making framework based on MCTS. The framework includes: (a) interaction modeling among AVs, HDVs, and roads with uncertainty; (b) safety assessment based on V2V, V2R, and V2H constraints; (c) multi-objective reward composition incorporating safety, comfort, cooperation, and efficiency; and (d) safety-critical MCTS rollout with node filtering and backpropagation. Optimal policies are derived under risk-aware constraints via tree search.

limited interpretability, a reliance on extensive training data, and difficulties in ensuring consistent safety guarantees in novel situations [18], [19]. Recent advancements include decision-making models with distributional reinforcement learning for perceptual uncertainty [20] and spatiotemporal-restricted A∗ algorithms for lane-free intersection coordination [21], [22]. However, these approaches typically address specific aspects of intersection management rather than providing comprehensive frameworks for mixed traffic coordination. Moreover, deploying these methods in real world remains a challenge, as ensuring consistent and robust performance across diverse traffic conditions is inherently difficult.

Game-theoretic methods [23]–[25] have emerged as an alternative paradigm for modeling strategic interactions among vehicles at unsignalized intersections [26]. Approaches such as Nash equilibrium and Stackelberg games [27] can effectively capture the competitive and cooperative behaviors between vehicles. However, these methods often rely on assumptions of perfect rationality and complete information, which rarely hold in real-world scenarios where HDVs exhibit varying levels of uncertainty and inconsistency. Additionally, the challenge of selecting among multiple Nash equilibria [28] can undermine cooperative consistency and lead to potential safety violations, especially in dynamic mixed traffic environments where rapid decision-making is critical [29].

Monte Carlo Tree Search (MCTS) [30] has emerged as a promising approach by marrying the learning-based and game-theoretic methods [31]. Unlike traditional DRL which requires extensive offline training, MCTS can efficiently explore the action space through online planning [32], [33]. The algorithm's inherent ability to balance exploration and exploitation makes it suitable for handling the uncertainties in mixed traffic environments [34]. However, current MCTS implementations such as [35], [36] often fall short in addressing comprehensive safety considerations and face significant scalability challenges in multi-agent scenarios.

This paper introduces a safety-critical multi-agent MCTS framework for coordinating mixed traffic at unsignalized intersections. The main contributions are summarized as follows:

- We propose a safety-critical multi-agent MCTS framework that integrates deterministic and probabilistic vehicle behavior predictions, enabling cooperative decision making among AVs and HDVs at unsignalized intersections.
- We develop a safety assessment mechanism that systematically handles three critical interaction types, AV-to-AV (V2V) [37], AV-to-HDV (V2H) [38] and Vehicle-to-Road (V2R) [39], by using dynamic safety thresholds and spatiotemporal risk metrics, providing comprehensive safety guarantees.
- We design a multi-objective reward function optimization approach that balances driving safety, efficiency, and cooperation, enabling efficient intersection navigation under safety constraints.
- We adopt an adaptive human driving behavior awareness framework that combines the deterministic Intelligent Driver Model (IDM) with probabilistic distributions to effectively capture human driving uncertainties, ensuring robot decision making in mixed traffic environments.

The rest of this paper is organized as follows: Section II describes the overview of proposed framework, Section III and IV illustrates the methodology, Section V presents simulation results, and Section VI concludes the paper.

II. Safety-critical decision making framework Overview

We consider the unsignalized intersection scenario where AVs need to coordinate with other vehicles (a mix of AVs and HDVs) without traffic signal guidance, with consideration of both deterministic autonomous behaviors and uncertain human driving patterns. The core challenge lies in handling three critical interaction types: V2V, V2H, and V2R interactions.

The proposed safety-critical decision-making framework consists of four main components, as illustrated in Fig. 1. First, we formulate the unsignalized intersection problem as a multi-agent Markov Decision Process (MDP), which involves defining the state and action spaces, safety constraints, and system dynamics model (Sec. III-A). Then, we develop a safety-critical decision-making mechanism by considering the safety interactions, establishing dynamic safety thresholds, and integrating HDVs prediction with risk assessment (Sec. III-B). Based on these foundations, we propose a safety-critical multi-agent MCTS framework that incorporates node structure and policy space design (Sec. IV-A). Additionally, we design a comprehensive reward function that combines multi-objective reward considerations—including V2V, V2R, and V2H safety aspects—as well as safety and dynamic constraints, leading to a safety-critical optimization formulation (Sec. IV-C). The proposed framework systematically addresses the challenges of safety-critical decision-making at unsignalized intersections, particularly focusing on the complex interactions between AVs and HDVs. The detailed design of each component and their interactions will be presented in the following sections.

III. The Proposed System Formulation

A. Multi-Agent MDP Formulation

We propose a centralized decision-making framework for multiple AVs at unsignalized intersections to better coordinate the behaviors of multiple AVs and achieve system-level optimal performance. The problem involving N AVs and M HDVs is formulated as a multi-agent MDP: $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}: \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the state transition function, $\mathcal{R}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines the reward function, and $\gamma \in (0,1]$ is the discount factor.

For the i-th vehicle, the state vector s_i consists of its position coordinates (x_i, y_i), velocity v_i, and heading angle θ_i. The control input a_i includes the acceleration command acc_i and steering rate $\dot{\theta}_i$. Then the joint state and action spaces are

$$\mathcal{S} = \prod_{i=1}^{N+M} \{ s_i = [x_i, y_i, v_i, \theta_i]^\top \in \mathbb{R}^4 \mid v_i \in [0, v_{\max}], |\theta_i| \leq \pi \},$$
$$\mathcal{A}_i = \{ a_i = [acc_i, \dot{\theta}_i]^\top \in \mathbb{R}^2 \mid |acc_i| \leq a_{\max}, |\dot{\theta}_i| \leq \dot{\theta}_{\max} \},$$

where $v_{\max}$ is maximum velocity, $a_{\max}$ is maximum acceleration/deceleration, and $\dot{\theta}_{\max}$ is maximum steering rate.

The intersection navigation decisions must meet basic safety constraints on state changes and inter-vehicle distances:

$$\begin{aligned} \mathcal{S}_{\text{safe}} &= \{ s \in \mathcal{S} \mid d_{i,j} \geq d_{\text{safe}}, \forall i, j \in \mathcal{V} \cup \mathcal{H} \}, \\ d_{i,j} &= \min_{p_i, p_j \in \mathbf{P}(s_i)} \| p_i - p_j \|_2, \end{aligned} \tag{1}$$

where $s = [s_1, s_2, ..., s_{N+M}]$, $d_{i,j}$ is the minimum distance between the i-th and j-th vehicles, d_{safe} is the minimum safe distance, $\mathbf{P}(s_i)$ is the four vertices of the i-th vehicle, $\mathcal{V}$ is the set of AVs and $\mathcal{H}$ is the set of HDVs.

Let $s_t = [s_{1,t}, s_{2,t}, ..., s_{N+M,t}]$ be the joint state of all vehicles at time step t, with $s_{i,t} = [x_{i,t}, y_{i,t}, v_{i,t}, \theta_{i,t}]^\top$ being the state vector of vehicle $i \in \mathcal{V} \cup \mathcal{H}$. The intersection navigation decisions must also satisfy constraints over T:

$$s_{t+1} \in \mathcal{S}_{\text{safe}}, \ \forall t \in [0, T], \ |v_i| \leq v_{\max}, \ d_{\text{v2r}}(s_i) \geq d_{\min}, \tag{2}$$

where $d_{\text{v2r}}(s_i)$ is the minimum distance to road boundaries, with the minimum allowable value $d_{\min}$.

The AV dynamics are governed by the kinematic model:

$$s_{t+1} = \Phi(s_t, \pi_t), \tag{3}$$

with the transition function $\Phi(s_t, \pi_t) = \{ f_i(s_{i,t}, \pi_{i,t}) \}_{i=1}^N$, and

$$f_i(s_{i,t}, \pi_{i,t}) = \begin{bmatrix} x_{i,t} + v_{i,t} \cos(\theta_{i,t}) \Delta t \\ y_{i,t} + v_{i,t} \sin(\theta_{i,t}) \Delta t \\ \text{sat}_{[0, v_{\max}]}(v_{i,t} + acc_{i,t} \Delta t) \\ \text{wrap}_{[-\pi, \pi]}(\theta_{i,t} + \dot{\theta}_{i,t} \Delta t) \end{bmatrix}, \tag{4}$$

where $\pi_t = [\pi_{1,t}, \pi_{2,t}, \cdots, \pi_{N,t}]$, with $\pi_{i,t} = (acc_i, \dot{\theta}_i)$, contains the control inputs for all AVs at time t. Δt is the time step, $\text{sat}_{[0, v_{\max}]}(\cdot)$ keeps $v_{i,t}$ within $[0, v_{\max}]$, and $\text{wrap}_{[-\pi, \pi]}(\cdot)$ handles angle continuity by wrapping θ to the interval $[-\pi, \pi]$.

While AV states are updated by (3), HDV behavior involves inherent uncertainties that grow over time, which must be properly characterized for reliable safety assessment.

B. Safety Assessment

To ensure safe navigation at unsignalized intersections, we develop a safety assessment framework that improves upon existing approaches through three key innovations: 1) Comprehensive analysis across V2V, V2H, and V2R interactions; 2) Dynamic context-aware safety thresholds that adapt to changing traffic conditions; and 3) Integration of both instantaneous and predictive risk metrics with explicit modeling of human driving uncertainties. This enables more realistic safety evaluations in mixed traffic environments.

1) V2R Safety Assessment: This assessment focuses on spatial constraints by partitioning the environment into the intersection area Ω_{int} and its approach area Ω_{app}. We define a Cartesian coordinate system centered at the intersection:

$$\begin{aligned} \Omega_{\text{int}} &= \{ s \in \mathcal{S} \mid |x| \leq R_{EX} \wedge |y| \leq R_{EX} \}, \\ \Omega_{\text{app}} &= \{ s \in \mathcal{S} \mid |x| \leq 2R_{EX} \vee |y| \leq 2R_{EX} \}, \end{aligned} \tag{5}$$

where R_{EX} denotes the half-length of the intersection area. $\wedge$ and $\vee$ represent the logical AND and OR operations, respectively. The approach area Ω_{app} extends beyond Ω_{int} to facilitate early risk assessment as vehicles approach the intersection. The safety level is evaluated through the minimum distance

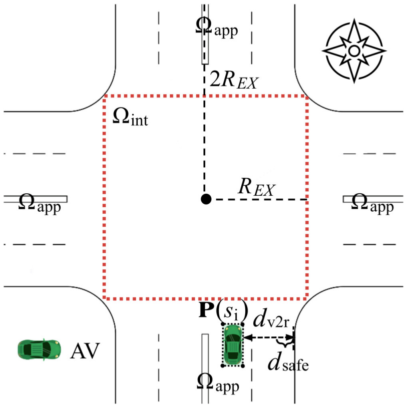


Fig. 2: Illustration of the intersection regions. The origin is set at the intersection center. The red dashed square denotes the intersection area Ω_{int}. The approach area Ω_{app} extends outward along the roads. The minimum vehicle-to-road distance (d_{v2r}) and the safe distance threshold (d_{safe}) are also shown.

to road boundaries $d_{\text{v2r}}(s_i)$ and its corresponding penalty function $\phi_{\text{v2r}}(d)$, defined as

$$\begin{aligned} d_{\text{v2r}}(s_i) &= \min_{p \in \mathbf{P}(s_i)} \text{distance}(p, \partial\Omega_{\text{int}} \cup \partial\Omega_{\text{app}}), \\ \phi_{\text{v2r}}(d) &= \begin{cases} -\infty, & d \leq d_{\min} \\ -\beta(d_{\min}/d)^2, & d_{\min} < d \leq d_{\text{safe}} \\ 0, & d > d_{\text{safe}} \end{cases}, \end{aligned} \tag{6}$$

where the symbol ∂ denotes the area boundary, d_{safe} is the safe threshold, and β is a scaling factor, as illustrated in Fig. 2.

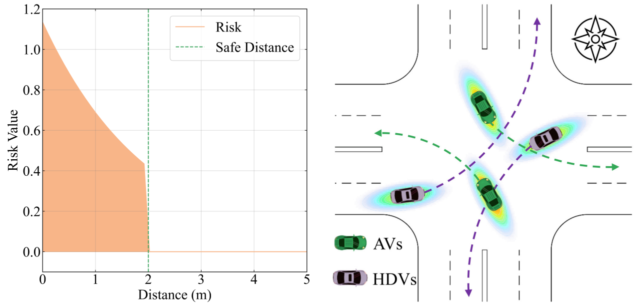
2) V2V and V2H Safety Assessment: The safety assessment incorporates both temporal and instantaneous risk evaluations. It utilizes an adaptive safety threshold that accounts for dynamic interaction conditions, defined as

$$d_{\text{safe}} = \max\{d_{\text{base}}, \kappa_v |\Delta v_{ij}|\} \cdot \prod_{k=1}^{3} \alpha_k(s_i, s_j) \tag{7}$$

where d_{base} is the minimum safe base distance, κ_v is a scaling factor, and $\Delta v_{ij} = v_i - v_j$ is the relative velocity between vehicles i and j. The first term ensures that safety distance increases with $|\Delta v_{ij}|$, consistent with safe driving practices. The adjustment factors $\{\alpha_k\}_{k=1}^{3}$ then modify this baseline considering specific interaction characteristics as follows:

$$\begin{aligned} \alpha_k(s_i, s_j) &= 1 + \beta_k \cdot f_k(s_i, s_j)/g_k, \ \ k = 1, 2, \\ f_1(s_i, s_j) &= |\Delta v_{ij}|, \ \ g_1 = v_{\text{ref}}, \\ f_2(s_i, s_j) &= |\Delta \theta_{ij}|, \ \ g_2 = \pi, \\ \alpha_3(s_i, s_j) &= 1 + \mathbb{1}_{\Omega_{\text{app}}}(s_i, s_j) + \mathbb{1}_{\Omega_{\text{int}}}(s_i, s_j), \end{aligned} \tag{8}$$

where the parameter β_k controlling the influence of the relative speed ($k = 1$) and relative heading angle ($k = 2$) on the safety distance, f_1 captures the velocity-dependent risk with $g_1 = v_{\text{ref}}$ (a reference velocity) serving as the normalization factor, and f_2 accounts for heading angle difference $\Delta\theta_{ij} = \theta_i - \theta_j$ that is normalized to $[0, 1]$ via $g_2 = \pi$. α_3 is the spatial risk factor to increase the safety margins in these conflict zones, where the indicator functions $\mathbb{1}\Omega_{\text{app}}$ and $\mathbb{1}\Omega_{\text{int}}$ are 1 when vehicles are in the zones Ω_{app} or Ω_{int}, respectively.



(a) (b)

Fig. 3: Safety-critical risk assessment. (a) Distance-based risk. (b) Mixed traffic interaction scenario at intersection.

Using (1) and (7), we define the instantaneous risk function

$$r(s_i, s_j) = \exp\left(-\frac{d_{i,j,\min}}{d_{\text{safe}}}\right) \cdot \left(1 + \lambda_v \frac{\|\Delta v_{ij}\|}{v_{\max}}\right), \tag{9}$$

where $d_{i,j,\min}$ is the minimum distance between vehicles i and j and λ_v is a scaling factor. The exponential term captures rising risk as distance nears the safety threshold, while the velocity term reflects added risk from higher relative speeds.

We further formulate the following temporal risk function concerning near-term safety over a prediction horizon T_p:

$$R_{T_p}(s_i, s_j) = \frac{1}{T_p} \sum_{t=1}^{T_p} \frac{1}{1+t} \cdot \rho(d_{i,j}, d_{\text{safe}}), \tag{10}$$

where $1/(1+t)$ is used to prioritize immediate risks by assigning higher weights to near-term predictions. This discounting reflects the greater certainty and importance of imminent events compared to those further in the future. The function $\rho(\cdot)$ evaluates the proximity to safety boundaries according to the current states s_i and s_j and is defined as

$$\rho(d_{i,j}, d_{\text{safe}}) = \begin{cases} 0, & d_{i,j} \geq d_{\text{safe}} \\ (1 - \frac{d_{i,j}}{d_{\text{safe}}})^2, & \text{otherwise} \end{cases}. \tag{11}$$

The distance-based risk function $\rho(d_{i,j}, d_{\text{safe}})$ and a typical mixed traffic interaction scenario at intersection are illustrated in Fig. 3(a) and Fig. 3(b), respectively. This piecewise function sets risk to zero at safe distances and applies a quadratic penalty as vehicles breach the safety threshold, capturing the sharply rising danger at close range.

The safety assessment methodology differs between V2V and V2H interactions based on their inherent characteristics. For V2V interactions, vehicle states evolve deterministically according to control inputs and the dynamic model (4). Using (9) and (10), the overall safety level of V2V interaction is quantified by the risk assessment function

$$Q_{\text{risk}}^{\text{v2v}}(s_i, s_j) = w_1^{\text{v2v}} r(s_i, s_j) + w_2^{\text{v2v}} R_{T_p}(s_i, s_j), i, j \in \mathcal{V}, \tag{12}$$

where w_1^{v2v} and w_2^{v2v} are the given weights that balance immediate and predictive risk components. This approach offers several advantages over existing methods: 1) The dynamic safety thresholds adapt to specific interaction contexts rather than using fixed distances, 2) the combination of instantaneous and predictive risk enables both reactive and proactive safety

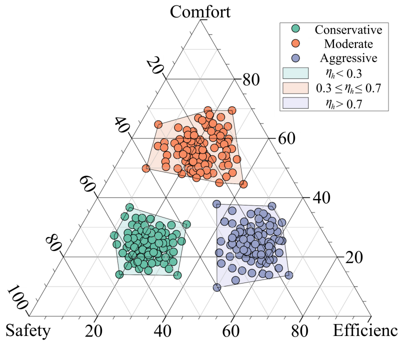


Fig. 4: Driving style classification based on preference distribution among safety, efficiency, and comfort. The parameter η_h distinguishes conservative ($\eta_h < 0.3$), moderate ($0.3 \le \eta_h \le 0.7$), and aggressive ($\eta_h > 0.7$) driving behaviors. Each point represents a possible driving behavior where the three preference values sum to 100%.

evaluations, and 3) the incorporation of spatial risk factors accounts for location-specific variations at intersections.

The V2H interactions require explicit consideration of human behavioral uncertainties. To better capture the diversity in human driving behaviors, we classify HDVs into different driving styles based on observable driving characteristics from the the Next Generation Simulation (NGSIM) datasets [40]. Specifically, we define a driving style parameter $\eta_h \in [0, 1]$ for each HDV $h \in \mathcal{H}$, where values closer to 0 indicate more conservative driving (prioritizing safety), and values closer to 1 indicate more aggressive driving (prioritizing efficiency), as illustrated in Fig. 4. The style parameter affects both the deterministic predictions and the uncertainty estimates.

For HDVs, we use the style-aware probabilistic prediction

$$P(\hat{s}_h|s_h, \eta_h) = \mathcal{N}(f_{\text{IDM}}(s_h, \eta_h), \Sigma_h(t, \eta_h)), \ \ h \in \mathcal{H}, \quad (13)$$

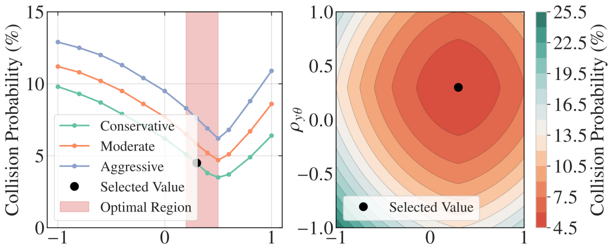
where s_h represents the current HDV state, $\hat{s}_h$ denotes its predicted future state, and $f_{\text{IDM}}(s_h, \eta_h)$ captures the nominal human driving behavior adjusted for the specific driving style. The IDM parameters are adapted based on the driving style:

$$\begin{aligned} a_{\text{max}}(\eta_h) &= a_{\text{max,base}} \cdot (1 + \alpha_a \cdot \eta_h), \\ T_{\text{headway}}(\eta_h) &= T_{\text{headway,base}} \cdot (1 - \alpha_T \cdot \eta_h), \\ d_{\text{safe}}(\eta_h) &= d_{\text{safe,base}} \cdot (1 - \alpha_d \cdot \eta_h), \end{aligned} \quad (14)$$

where $a_{\text{max,base}}$, $T_{\text{headway,base}}$, and $d_{\text{safe,base}}$ are the baseline IDM parameters, while α_a, α_T, and α_d are scaling factors that control how driving style affects each parameter. Aggressive drivers (higher η_h) tend to have higher maximum acceleration, shorter time headway, and reduced safety distances. $\Sigma_h(t, \eta_h)$ is a time-varying covariance matrix capturing the prediction uncertainty and given by

$$\Sigma_h(t, \eta_h) = \begin{bmatrix} \Sigma_{xx} & 0 & \Sigma_{xv} & 0 \\ 0 & \Sigma_{yy} & 0 & \Sigma_{y\theta} \\ \Sigma_{xv} & 0 & \Sigma_{vv} & 0 \\ 0 & \Sigma_{y\theta} & 0 & \Sigma_{\theta\theta} \end{bmatrix}, \quad (15)$$

where $\Sigma_{xx} = \sigma_x^2(\eta_h)t + \epsilon_x^2(\eta_h)t^2$, $\Sigma_{yy} = \sigma_y^2(\eta_h)t + \epsilon_y^2(\eta_h)t^2$, $\Sigma_{vv} = \sigma_v^2(\eta_h)$, $\Sigma_{\theta\theta} = \sigma_\theta^2(\eta_h)$, $\Sigma_{xv} = \rho_{xv}\sigma_x(\eta_h)\sigma_v(\eta_h)t$, and $\Sigma_{y\theta} = \rho_{y\theta}\sigma_y(\eta_h)\sigma_\theta(\eta_h)t$. The uncertainty parameters are also adjusted based on driving style:



(a) ρ_{xv} (b) ρ_{xv}

Fig. 5: Sensitivity analysis of correlation parameters in the HDV prediction model. (a) Impact of ρ_{xv} on collision probability. (b) Combined effect of correlation parameters ρ_{xv} and $\rho_{y\theta}$ on collision probability in intersection turning.

$$\begin{aligned} &\sigma_x(\eta_h) = \sigma_{x,\text{base}}(1 + \beta_x \eta_h), \ \sigma_y(\eta_h) = \sigma_{y,\text{base}}(1 + \beta_y \eta_h), \\ &\sigma_v(\eta_h) = \sigma_{v,\text{base}}(1 + \beta_v \eta_h), \ \sigma_\theta(\eta_h) = \sigma_{\theta,\text{base}}(1 + \beta_\theta \eta_h), \end{aligned} \quad (16)$$

where the baseline uncertainty parameters $\sigma_{x,\text{base}}$, $\sigma_{y,\text{base}}$, $\sigma_{v,\text{base}}$, and $\sigma_{\theta,\text{base}}$ are scaled according to driving style, with β parameters controlling the effect magnitude; aggressive drivers typically have higher uncertainty, shown by larger β.

The covariance matrix $\Sigma_h(t, \eta_h)$ is designed to capture key correlations observed in human driving behavior. Specifically, we model two critical correlations: 1) the correlation between longitudinal position and velocity (ρ_{xv}), and 2) the correlation between lateral position and heading angle ($\rho_{y\theta}$). These correlations reflect the physical coupling in vehicle dynamics - velocity directly influences position change over time, while heading angle determines the direction of lateral movement.

To validate the sensitivity of our model to these correlation parameters, we conducted a systematic analysis varying ρ_{xv} and $\rho_{y\theta}$ within the range $[-1, 1]$ at intervals of 0.2. Fig. 5 shows the impact on collision probability estimation with different correlation values for two representative driving scenarios: longitudinal movement and intersection turning. The analysis reveals that ρ_{xv} has the most significant impact in longitudinal movement scenarios, with higher positive correlations (0.3-0.5) providing the most accurate predictions compared to the NGSIM dataset. In turning scenarios at intersections, as shown in Fig. 5(b), both correlations contribute significantly to accurate risk assessment, with their combined effect creating a clear optimal region in the parameter space.

Based on this sensitivity analysis and validation against the NGSIM datasets [41], we selected $\rho_{xv} = 0.3$ and $\rho_{y\theta} = 0.3$ as our default correlation parameters, as they provide a balanced representation across different driving scenarios while matching the observed correlations in human driving patterns. These values also align with previous findings in driver behavior modeling [42], [43], which reported correlation coefficients in similar ranges for natural driving tasks.

The zero elements in the covariance matrix $\Sigma_h(t, \eta_h)$ in (15) reflect our modeling assumption that certain state variables have negligible direct correlation (e.g., lateral position and longitudinal velocity). This simplification is supported by vehicle dynamics principles and empirical observations in human

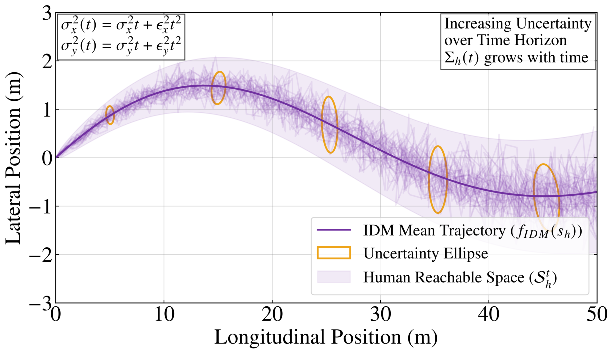


Fig. 6: Visualization of HDV’s trajectory evolution uncertainty.

driving data, where such cross-correlations are typically weak compared to the primary correlations we model [44].

To obtain physically feasible prediction of HDVs, we also bound HDVs’ reachable state space as the set

$$\begin{aligned}\mathcal{S}_h^t = \{\hat{s}_h \in \mathbb{R}^4 \; || \|p_h - p_h(t)\| \leq (v_{\max} + \sigma_v)t, \\ |v_h| \leq v_{\max} + 2\sigma_v, \, |\theta_h| \leq \pi\},\end{aligned} \tag{17}$$

where σ_v reflects the behavioral uncertainty. The evolution of the prediction uncertainty is visualized in Fig. 6.

We quantify the collision probability of V2H using

$$C_{ih} = \int_{\hat{s}_h \in \mathcal{S}_h^t(\eta_h)} \psi(\hat{s}_i, \hat{s}_h) \cdot \mathcal{N}(f_{\text{IDM}}(s_h, \eta_h), \Sigma_h(t, \eta_h)) \mathrm{d}\hat{s}_h, \tag{18}$$

where $\psi(\hat{s}_i, \hat{s}_h) = \mathbb{I}(d(\hat{s}_i, \hat{s}_h) < d_{\text{safe}})$ is the collision indicator function comparing AV’s predicted state $\hat{s}_i$ and HDV’s future states $\hat{s}_h$ against a style-dependent safety threshold d_{safe}. $\hat{s}_i$ is predicted using a constant acceleration model, while $\hat{s}_h$ follows (13) with the driving style η_h.

By using (9), (10), (11) and (18), the overall safety level is quantified by the risk assessment function

$$\begin{aligned}Q_{\text{risk}}^{\text{v2h}}(s_i, s_h) = w_1^{\text{v2h}} r(s_i, \hat{s}_h) + w_2^{\text{v2h}} R_{T_p}(s_i, s_h) + w_3^{\text{v2h}} C_{ih}, \\ i \in \mathcal{V}, \;\; h \in \mathcal{H}, \;\; \hat{s}_h \in \mathcal{S}_h^t,\end{aligned} \tag{19}$$

where $w_k^{\text{v2h}}, k \in [1, 3]$, are given weights. The first term evaluates instantaneous risk with predicted HDV states, the second term considers temporal risk evolution, and the third term means collision probability under prediction uncertainties.

This comprehensive safety assessment framework offers several advantages over existing approaches: 1) It handles both deterministic and probabilistic interactions through a unified mathematical formulation; 2) It quantifies safety in terms of both immediate risk and future collision probability, providing a more complete risk assessment; 3) It explicitly accounts for human driving uncertainty through principled probabilistic models rather than simplistic assumptions; and 4) It seamlessly integrates with the decision-making framework to enable safety-critical planning. These characteristics make our approach particularly suitable for complex mixed traffic scenarios at unsignalized intersections, where existing methods often fail to balance safety and efficiency due to their limited consideration of interaction complexity and uncertainty.

Algorithm 1 Safety-Critical Multi-Agent MCTS Algorithm

```
Input: Initial state $s_0$, iteration limit $K$, and $d_{\max}$
Initialize root node $n_0$ with state $s_0$, $N_{n_0} \leftarrow 0$, $Q_{n_0} \leftarrow 0$
for iteration $k = 1$ to $K$ do
    $n \leftarrow n_0$                                        ▷ Current node in tree
    while $d_n < d_{\max}$ and $\xi_n =$ o” do               ▷ Safe node check
        if $N_n = 0$ then                                       ▷ Unvisited node
            $\Delta \leftarrow$ Rollout($n$)
            Backpropagate($n, \Delta$)
            break
        else if $C_n = \emptyset$ then                               ▷ Leaf node
            Expand($n$)                                  ▷ Generate child nodes
            for $n_{\text{child}} \in C_n$ do
                Evaluate safety using $Q_{\text{risk}}^{\text{v2v}}$, $Q_{\text{risk}}^{\text{v2h}}$, $d_{\text{v2r}}$
                if $Q_{\text{risk}}^{\text{v2v}} \leq Q_{\text{th}}^{\text{v2v}}$ and
              $Q_{\text{risk}}^{\text{v2h}} \leq Q_{\text{th}}^{\text{v2h}}$ and $d_{\text{v2r}} \geq d_{\min}$ then
                    $\xi_{n_{\text{child}}} \leftarrow$ “o”                        ▷ Mark as safe
                else
                    $\xi_{n_{\text{child}}} \leftarrow$ “x”                      ▷ Mark as unsafe
                end if
            end for
            $\Delta \leftarrow$ Rollout($n$)
            Backpropagate($n, \Delta$)
            break
        else
            $n \leftarrow$ SelectChild($n$)                 ▷ Using UCB in (25)
        end if
    end while
end for
return ExtractPolicy($n_0$)
```

IV. The Multi-Agent MCTS Framework

The proposed safety-critical multi-agent MCTS framework is summarized in Algorithm 1, whose design is detailed in Sections IV-A, IV-B, and IV-C. Its computational complexity is also analyzed in Section IV-D.

A. Safety-Critical Tree Search Design

1) Node Structure and Policy Space: We propose a structured tree search framework where the risk assessment functions in (12) and (19) are used to evaluate the safety of each node (state-action pair) and prune unsafe nodes that exceed predefined safety thresholds ($Q_{\text{th}}^{\text{v2v}}$ for V2V interactions and $Q_{\text{th}}^{\text{v2h}}$ for V2H interactions). This ensures the generated policies optimize objectives while maintaining safety.

The joint policy space $\Pi_{\text{joint}} = \bigotimes_{i=1}^{N+M} \mathcal{P}_i$ represents the Cartesian product of the basic policy sets $\mathcal{P}_i$ of all $N + M$ vehicles. For each vehicle, we define a discrete action space:

$$\mathcal{P}_i = \left\{ \begin{bmatrix} acc_i \\ \dot{\theta}_i \end{bmatrix} \middle| \begin{array}{l} acc_i \in \{-a_{\max}, -a_{\text{med}}, 0, a_{\text{med}}, a_{\max}\} \\ \dot{\theta}_i \in \{-\dot{\theta}_{\max}, -\dot{\theta}_{\text{med}}, 0, \dot{\theta}_{\text{med}}, \dot{\theta}_{\max}\} \end{array} \right\}, \tag{20}$$

with a medium steering rate $\dot{\theta}_{\text{med}}$ for finer control precision.

Let $\mathbb{T}$ be the search tree whose node $n \in \mathbb{T}$ defined as:

$$\begin{aligned} n = (d_n, p_n, C_n, N_n, Q_n, \pi_n, \xi_n) \in \\ \mathbb{N} \times \mathbb{N} \times 2^{\mathbb{N}} \times \mathbb{N} \times \mathbb{R} \times \Pi_{\text{joint}} \times \{“o”, “x”\}. \end{aligned} \tag{21}$$

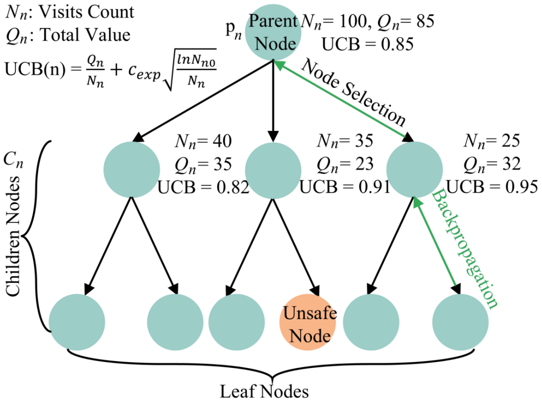


Fig. 7: Illustration of the safety-critical MCTS framework. Each node maintains N_n, Q_n, and UCB scores, with unsafe nodes (orange) eliminated during safety validation. The green path shows the backpropagation.

The status "o" indicates *safe* while "x" indicates *unsafe*. Each node maintains essential information including depth d_n, parent node p_n, child nodes set C_n, visit count N_n, value estimates Q_n, policy π_n, and safety status ξ_n. The safety status is determined through a comprehensive evaluation:

$$\xi_n = \begin{cases} \text{“o”}, & \text{if } Q_{\text{risk}}^{\text{v2v}} \leq Q_{\text{th}}^{\text{v2v}} \text{ and} \\ & Q_{\text{risk}}^{\text{v2h}} \leq Q_{\text{th}}^{\text{v2h}} \text{ and } d_{\text{v2r}} \geq d_{\min} \\ \text{“x”}, & \text{otherwise} \end{cases}, \tag{22}$$

where $Q_{\text{risk}}^{\text{v2v}}$ is the vehicle-to-vehicle risk measure, $Q_{\text{th}}^{\text{v2v}}$ is its threshold, $Q_{\text{risk}}^{\text{v2h}}$ is the vehicle-to-human risk measure, $Q_{\text{th}}^{\text{v2h}}$ is its threshold, d_{v2r} is the distance to road boundary, and $d_{\min}$ is the minimum safe distance.

The safety-critical multi-agent MCTS operates over the discretized action space Π_{joint} using depth-first search with safety checks at each node. It comprises four stages: selection, transition, rollout, and backpropagation (Fig. 7).

2) Safety-Critical Tree Search Process: Let n_t be the node at time t. The search process evaluates nodes recursively based on both their potential value and safety constraints:

$$\text{Search}(n_t) = \begin{cases} \text{Expand}(n_t) \cup \text{Rollout}(n_t), & \textit{if new\&safe} \\ \text{Search}(\text{UCB}(n_t)), & \textit{if visited} \\ \text{Terminate}, & \textit{if unsafe or max depth} \end{cases} \tag{23}$$

During expansion, we generate child nodes by applying all possible actions from Π_{joint}. The safety evaluation follows:

$$\begin{aligned} \text{SafetyCheck}(n) = & \left(\wedge_{i,j \in \mathcal{V}, i \neq j} \text{V2V}(s_i, s_j)\right) \\ & \wedge \left(\wedge_{i \in \mathcal{V}, h \in \mathcal{H}} \text{V2H}(s_i, s_h)\right) \wedge \left(\wedge_{i \in \mathcal{V}} \text{V2R}(s_i)\right) \end{aligned} \tag{24}$$

The node selection balances exploration and exploitation using

$$\text{UCB}(n) = \begin{cases} \frac{Q_n}{N_n} + c_{\text{exp}}\sqrt{(\ln N_{n_0})/N_n}, & \text{if } N_n > 0 \\ +\infty, & \text{if } N_n = 0 \end{cases} \tag{25}$$

where $c_{\text{exp}} \geq \frac{1}{2\sqrt{2}}\Delta_{\max}$ with $\Delta_{\max} = \max_{n,n' \in \mathcal{T}} |Q_n - Q_{n'}|$ ensuring sufficient exploration as shown in Fig. 7. After node selection, the system states are evolved following (4).

3) Rollout Strategy and Value Backpropagation: For unvisited nodes, we employ a hybrid rollout strategy combining model-based prediction and random sampling:

$$\begin{aligned} \pi_{\text{rollout}} &= \alpha \pi_{\text{model}} + (1 - \alpha) \pi_{\text{random}}, \\ \pi_{\text{model}} &= \arg\max_{\pi \in \Pi_{\text{safe}}} Q_{\text{pred}}(s, \pi), \ \pi_{\text{random}} \sim \mathcal{U}(\Pi_{\text{safe}}), \end{aligned} \tag{26}$$

where α is a mixing parameter, Π_{safe} is the set of safe actions, Q_{pred} is a learned value predictor, s is the current state, π_{rollout} is the final rollout policy, π_{model} is the model-based policy component, π_{random} is randomly sampled from a uniform distribution $\mathcal{U}$ over the safe action space, and Π_{safe} is the set of safe actions.

The rollout estimates node values via forward simulation:

$$\begin{aligned} \mathcal{P}_n &= \{\pi_k\}_{k=1}^{l_n}, \quad \pi_k = \pi_{p^{l_n - k}(n)}, \\ \mathcal{P}_{n,\text{ext}} &= \mathcal{P}_n \cup \{\pi_k\}_{k=l_n+1}^{d_{\max}}, \quad \pi_k \sim \mathcal{U}(\Pi_{\text{joint}}), \\ Q_{\text{eval}}(n) &= \sum_{t=0}^{d_{\max}-1} \gamma_r^t \mathcal{R}(s_t, \pi_t), \end{aligned} \tag{27}$$

where $\mathcal{P}_n$ is the action sequence from root to node n, l_n is the depth of node n, $p^k(n)$ is the k-th parent of node n, $d_{\max}$ is the maximum simulation depth, $\gamma_r \in (0, 1)$ is the reward-specific discount factor prioritizing immediate safety, $\mathcal{R}$ is the reward function, and s_t and π_t are the state and action at time t.

The value backpropagation follows

$$N_n \leftarrow N_n + 1, \quad Q_n \leftarrow Q_n + \frac{1}{N_n^{\beta_{\text{dec}}}}[Q_{\text{eval}}(n_{\text{eval}}) - Q_n], \tag{28}$$

where N_n is the visit count of node n, Q_n is the value estimate of node n, n_{eval} is the evaluated leaf node, and β_{dec} controls the learning rate decay of value updates.

B. Multi-Objective Reward Design

The multi-objective reward function is designed to address three key aspects of autonomous driving at unsignalized intersections: safety in various interaction scenarios, motion efficiency, and driving comfort.

1) Safety Component: The safety component Q_{safety}^i is defined as

$$Q_{\text{safety}}^i = w_{\text{v2v}} Q_{\text{v2v}}^i + w_{\text{v2r}} Q_{\text{v2r}}^i + w_{\text{v2h}} Q_{\text{v2h}}^i, \tag{29}$$

with $Q_{\text{v2v}}^i = \sum_{j \in \mathcal{V} \backslash \{i\}} (\phi_{\text{v2v}}(d_{i,j}, \Delta v_{ij}, \Delta\theta) + \lambda_T R_{T_p}(s_i, s_j))$, $Q_{\text{v2r}}^i = \phi_{\text{v2r}}(d_{\text{v2r}}(s_i))$, and $Q_{\text{v2h}}^i = \sum_{h \in \mathcal{H}} (\phi_{\text{v2h}}(d_{i,h}, \Delta v_{ih}, \Delta\theta, \xi) + \lambda_T R_{T_p}(s_i, s_h) + \lambda_c C_{ih})$, where $\xi \in \Omega$ denotes the scenario type, w_{v2v}, w_{v2r}, and w_{v2h} are weighting factors determining the relative importance of different safety components: w_{v2v} emphasizes V2V interactions, w_{v2r} reflects V2R safety measures, and w_{v2h} prioritizes V2H considerations. These weights can be tuned based on specific driving scenarios, such as urban environments or highways.

The V2V safety penalty function $\phi_{\text{v2v}}(d_{i,j}, \Delta v_{ij}, \Delta\theta)$ is defined as

$$\phi_{\text{v2v}}(\cdot) = \begin{cases} -1, & d_{i,j} \leq d_{\text{safe}} \\ 0, & \text{otherwise} \end{cases}. \tag{30}$$

The V2H safety penalty function $\phi_{\text{v2h}}(d_{i,h}, \Delta v_{ih}, \Delta\theta, \xi)$ is defined as

$$\phi_{\text{v2h}}(\cdot) = \begin{cases} -\mu_{\text{s}}\psi_1(d_f, \Delta v_{ih}, \Delta\theta), & d_{i,h} \leq d_{th} \\ -\mu_{\text{s}}\psi_2(d_f, \Delta v_{ih}, \Delta\theta), & d_{th} < d_{i,h} \leq d_{\text{safe}} \\ 0, & d_{i,h} > d_{\text{safe}} \end{cases}$$

where $\psi_1(d_f, \Delta v_{ih}, \Delta\theta) = \kappa_1 + \kappa_2 d_f + \kappa_3 \Delta v_{ih} + \kappa_4 \Delta\theta$ and $\psi_2(d_f, \Delta v_{ih}, \Delta\theta) = d_f(1 + \eta_1 \Delta v_{ih}/v_{\text{ref}} + \eta_2 \Delta\theta/\pi)$ are the penalty functions for different distance ranges, $d_f = (d_{\text{safe}} - d_{i,h})/(d_{\text{safe}} - d_{th})$ is the distance factor, $d_{th} = 0.5 d_{\text{safe}}$ is the intermediate threshold, and $\mu_{\text{s}}(\xi)$ adapts to different scenarios.

2) Efficiency Component: The efficiency component Q^i_{eff} evaluates motion quality through velocity tracking (Q^i_{vel}), acceleration smoothness (Q^i_{acc}), and reference path following (Q^i_{ref}) defined as

$$\begin{aligned} Q^i_{\text{acc}}(a_i) &= -w_{\text{acc}}(acc_i - a^i_{\text{des}})^2, \\ Q^i_{\text{ref}}(p_i) &= -w_{\text{ref}}\|pos_i - p^i_{\text{ref}}\|^2, \\ Q^i_{\text{vel}}(v_i) &= \begin{cases} 0, & |v_i - v^i_{\text{des}}| \leq v_{\text{tol}} \\ -\alpha_v(v_i - v^i_{\text{des}} - v_{\text{tol}}), & v_i - v^i_{\text{des}} > v_{\text{tol}} \\ -\beta_v|v_i - v^i_{\text{des}}|, & \text{otherwise} \end{cases}, \end{aligned}$$

where acc_i, v_i, and pos_i are the current acceleration, velocity, and position, respectively. a^i_{des}, v^i_{des}, and p^i_{ref} are the desired acceleration, velocity, and reference position, respectively. v_{tol} is the velocity tolerance threshold. w_{vel}, w_{acc}, and w_{ref} are weight parameters. α_v and β_v are velocity deviation penalties for overspeeding and underspeeding, respectively.

3) Comfort and Cooperation Components: The comfort evaluation considers both longitudinal and lateral dynamics:

$$Q^i_{\text{comfort}} = -w_{\text{jerk}}|\dot{a}_i(s_t)|^2 - w_{\text{yaw}}|\ddot{\theta}_i(s_t)|^2, \tag{31}$$

where w_{jerk} and w_{yaw} are user-specified weights, $\dot{a}_i$ is the jerk (rate of acceleration change), and $\ddot{\theta}_i$ is the yaw acceleration of vehicle i. The cooperation component captures interactions with other vehicles in both sets $\mathcal{V}$ and $\mathcal{H}$:

$$Q^{\text{other}}_i = \sum_{j \in (\mathcal{V} \cup \mathcal{H}) \setminus \{i\}} w_{ij}(Q^j_{\text{safety}} + Q^j_{\text{eff}}), \tag{32}$$

where w_{ij} represents the weight between vehicles i and j.

To ensure fairness and prevent excessive yielding, we introduce a dynamic cooperation coefficient for AV i:

$$\lambda_i(t) = \lambda_{\text{base}} \cdot \exp(-\alpha_{\text{wait}} \cdot T^i_{\text{wait}}(t)), \tag{33}$$

where λ_{base} is the base cooperation coefficient, α_{wait} is a waiting time penalty factor, and $T^i_{\text{wait}}(t)$ represents the accumulated waiting time of AV i at time t defined as $T^i_{\text{wait}}(t) = \sum_{\tau=0}^{t} \mathbb{I}(v_i(\tau) < v_{\text{thres}})$, where $\mathbb{I}(\cdot)$ is an indicator function that equals 1 when the velocity of AV i falls below a threshold v_{thres}, indicating it is waiting or slowed down. The exponential decay form follows established principles in utility-based decision making [45], where agents exhibit diminishing sensitivity to accumulated delay. This aligns with behavioral economic models such as prospect theory, in which the subjective cost of waiting increases sub-linearly, leading to reduced willingness to cooperate as waiting time grows.

This dynamic cooperation mechanism ensures that AVs that have been waiting longer will gradually reduce their cooperation level, prioritizing their own passage through the intersection. This prevents situations where certain AVs might be persistently excluded from entering the intersection due to excessive cooperation.

To balance safety, efficiency, comfort, and cooperation in the decision-making process, the final reward function for AV i is defined as

$$\mathcal{R}_i(s_t, \pi(s_t)) = \frac{Q^{\text{self}}_i(s_t, \pi(s_t)) + \lambda_i Q^{\text{other}}_i(s_t, \pi(s_t))}{1 + \lambda_i(N-1)}, \tag{34}$$

where $Q^{\text{self}}_i = w_1 Q^i_{\text{safety}} + w_2 Q^i_{\text{eff}} + w_3 Q^i_{\text{comfort}}$, with the given weights w_1, w_2, and w_3. The denominator $1 + \lambda_i(N-1)$ ensures that the cooperative term Q^{other}_i is properly scaled relative to the individual term Q^{self}_i regardless of the number of AVs N.

The total reward of all AVs over horizon T is defined as

$$R_{\text{total}} = \sum_{t=0}^{T-1} \gamma_r^t \sum_{i \in \mathcal{V}} \mathcal{R}_i(s_t, \pi(s_t)), \tag{35}$$

where $\gamma_r \in (0,1)$ is the reward-specific discount factor that prioritizes immediate rewards over future ones

While multi-objective reward functions are common in autonomous driving literature, our proposed design introduces several distinctive features specifically tailored for mixed traffic coordination at unsignalized intersections. First, unlike standard approaches that typically focus on individual vehicle optimization, our formulation explicitly incorporates inter-vehicle cooperation through the term Q^{other}_i, enabling system-level coordination. Second, we introduce a novel normalization mechanism $1 + \lambda_i(N-1)$ that ensures consistent performance regardless of the number of vehicles, addressing the limitation in existing cooperative reward designs. Third, our hierarchical decomposition of safety components (Q^i_{v2v}, Q^i_{v2r}, and Q^i_{v2h}) with scenario-adaptive penalties (μ_{s}) provides fine-grained control over different risk sources, rather than using a monolithic safety term. Fourth, our efficiency component uses an asymmetric penalty structure that differentiates between overspeeding and underspeeding, reflecting the different risk profiles of these behaviors in intersection scenarios. These innovations collectively enable more nuanced and effective decision-making in complex intersection environments compared to standard reward formulations.

C. From Reward to Optimal Policy

The comprehensive reward function guides the search for optimal policies. The global optimization objective can be formulated as

$$\pi^* := \arg\max_{\pi \in \Pi_{\text{joint}}} \mathbb{E}\Big[\sum_{t=0}^{T-1} \gamma_r^t \sum_{i \in \mathcal{V}} \mathcal{R}_i(s_t, \pi(s_t)) \mid \pi\Big] \tag{36}$$

subject to the following constraints:

$$\begin{aligned}
&\textbf{Safety Constraints:} && \\
&Q^{\text{v2v}}_{\text{risk}}(s_i, s_j) \le Q^{\text{v2v}}_{\text{th}}, && \forall i, j \in \mathcal{V} && \text{(V2V)}\\
&Q^{\text{v2h}}_{\text{risk}}(s_i, s_h) \le Q^{\text{v2h}}_{\text{th}}, && \forall i \in \mathcal{V}, h \in \mathcal{H} && \text{(V2H)}\\
&d_{\text{v2r}}(s_i) \ge d_{\min}, && \forall i \in \mathcal{V} && \text{(V2R)}\\
&\textbf{Dynamic Constraints:} && \\
&v_i \in [0, v_{\max}], && \forall i \in \mathcal{V} && \text{(Velocity)}\\
&|acc_i| \le a_{\max}, && \forall i \in \mathcal{V} && \text{(Acceleration)}\\
&|\dot{\theta}_i| \le \dot{\theta}_{\max}, && \forall i \in \mathcal{V} && \text{(Steering)}
\end{aligned}$$

The safety constraints incorporate risk metrics $Q^{\text{v2v}}_{\text{risk}}(s_i, s_j)$ for V2V and $Q^{\text{v2h}}_{\text{risk}}(s_i, s_h)$ for V2H interactions, which are computed based on instantaneous risk (9), temporal risk (10), risk boundary condition (11), and collision probability (18). The separate thresholds $Q^{\text{v2v}}_{\text{th}}$ and $Q^{\text{v2h}}_{\text{th}}$ are designed to handle different uncertainty levels for V2V and V2H interactions.

Our reward-based optimization differs from standard methods by explicitly incorporating safety constraints alongside a structured reward function. Unlike traditional approaches that rely solely on reward penalties—often resulting in either conservative or risky behavior depending on tuning—our method enforces hard safety constraints, ensuring safety while optimizing performance within the admissible action space. This separation enables more aggressive pursuit of efficiency and comfort without compromising safety. Furthermore, our cooperative reward normalization ensures consistent optimization across varying numbers of agents, addressing a common limitation in multi-agent formulations where cooperation levels depend on agent count.

To solve this optimization problem through MCTS, at each tree node, the local policy selection is determined by maximizing the expected cumulative reward:

$$\pi^*(s) := \underset{\pi \in \Pi_{\text{safe}}}{\arg\max}\, \mathbb{E}\Big[\sum_{t=0}^{d_{\max}} \gamma_r^t \mathcal{R}_i(s_t, \pi(s_t)) \mid s_0 = s\Big]. \quad (37)$$

The node values Q_n maintained by the MCTS algorithm represent the estimated cumulative reward:

$$Q_n \approx \sum_{t=0}^{d_{\max}} \gamma_r^t \mathcal{R}_i(s_t, \pi_t). \quad (38)$$

This estimation is refined through backpropagation:

$$Q_n \leftarrow (1 - \alpha_n) Q_n + \alpha_n \Big[\mathcal{R}_i(s_n, \pi_n) + \gamma_r \max_{n' \in C_n} Q_{n'}\Big], \quad (39)$$

where $\alpha_n = 1/N_n^{\beta_{\text{dec}}}$ is the learning rate with $\beta_{\text{dec}} \in (0.5, 1]$.

The convergence of this update rule is guaranteed when $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$. With $\alpha_n = 1/N_n^{\beta_{\text{dec}}}$ and $\beta_{\text{dec}} \in (0.5, 1]$, both conditions are satisfied, ensuring that:

$$\lim_{N_n \to \infty} Q_n = \mathbb{E}\Big[\sum_{t=0}^{d_{\max}} \gamma_r^t \mathcal{R}_i(s_t, \pi_t)\Big]. \quad (40)$$

The policy execution process operates in a receding horizon manner, iteratively selecting and executing actions while maintaining safety. After completing the MCTS iterations, the execution process consists of three key steps:

1) Best Action Selection: At each planning cycle, the best child node n^* is selected via

$$n^* = \underset{n \in C_{n_0}}{\arg\max} \{N_n + \epsilon Q_n\}, \quad (41)$$

where $\epsilon \in (0, 1)$ balances visit count and node value. This selection criterion is more robust than raw Q-values, reflecting the most thoroughly explored action sequence.

2) Action Execution: The policy π_{n^*} associated with n^* provides the control commands a_t for the current timestep:

$$a_t = [acc_{n^*},\ \dot{\theta}_{n^*}]^\top, \quad (42)$$

where acc_{n^*} and $\dot{\theta}_{n^*}$ represent the acceleration and steering rate commands from the optimal policy π_{n^*}.

3) Tree Update: After executing the action, the planning tree is updated by promoting n^* as the new root node:

$$n_0^{\text{new}} = n^*, \quad s_0^{\text{new}} = f(s_t, a_t), \quad (43)$$

where $f(s_t, a_t)$ is defined in (4). The subtree rooted at n^* is preserved for warm-starting the next planning iteration, while other branches are pruned.

This receding horizon approach ensures computational efficiency through tree reuse while maintaining safety through the embedded constraint checks at each iteration.

D. Computational Complexity Analysis

The computational complexity of the proposed algorithm stems from tree expansion and rollout simulation.

The tree expansion process considers the joint actions of N AVs, each having $|\mathcal{A}|$ actions, resulting in a branching factor of $(|\mathcal{A}|)^N$. For each expanded node, safety validation checks must be performed. The V2V interactions contribute a complexity of $O(N^2)$, V2H interactions contribute $O(NM)$, and V2R boundary checks contribute $O(N)$, yielding a total validation complexity of $O(N^2 + NM + N)$.

The rollout simulation is conducted for each new node up to a depth $d_{\max}$. Each rollout step involves state transitions and reward evaluations. The state transitions for all vehicles require $O(N + M)$ operations, while evaluating pairwise interactions in the reward computation scales as $O(N^2)$. The total rollout complexity can be expressed as $O(d_{\max}(N^2 + NM + N))$.

Considering K MCTS iterations, the worst-case computational complexity can be estimated as

$$O(K \cdot d_{\max} \cdot |\mathcal{A}|^N \cdot (N^2 + NM + N)). \quad (44)$$

In practical implementations, the actual computational cost is often lower than this theoretical worst-case bound. The pruning of unsafe nodes reduces the effective branching factor, while the selective nature of UCB-based exploration ensures efficient tree expansion. Furthermore, the algorithm structure allows for potential parallel implementation, which can significantly improve computational efficiency.

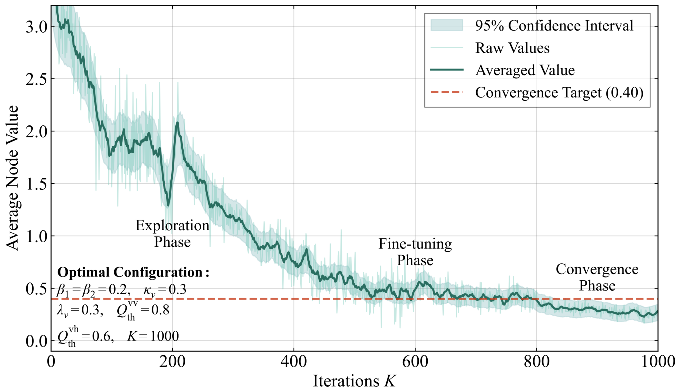


Fig. 8: MCTS algorithm convergence analysis.

E. Algorithm Convergence Analysis

Convergence analysis in Fig. 8 shows that the value function exhibits stable multi-phase behavior, typically converging within 300–600 iterations. Based on this observation, we set $K = 1000$ as the maximum number of MCTS iterations, providing a sufficient margin to ensure convergence across a broad range of scenarios. The final value standard deviation remains consistently below 0.05, further confirming the robustness of the algorithm under dynamic conditions. The key algorithmic parameters are selected through systematic sensitivity analysis and convergence validation. Adjustment factors $\beta_1 = \beta_2 = 0.2$ are used to balance the influence of relative velocity and heading angle in determining dynamic safety thresholds. The velocity scaling factor $\kappa_v = 0.3$ enables an effective trade-off between reactive safety and traffic flow efficiency. Similarly, the risk scaling parameter $\lambda_v = 0.3$ provides sufficient sensitivity to relative velocity changes while avoiding excessive conservatism. Safety thresholds are set to $Q_{\text{th}}^{\text{vv}} = 0.8$ and $Q_{\text{th}}^{\text{vh}} = 0.6$, reflecting more stringent safety considerations in scenarios involving human-driven vehicles. These design choices collectively ensure stable and reliable behavior across varying traffic densities, heterogeneous agent types, and dynamic interaction patterns.

V. Experimental Evaluation

Simulations are conducted in MATLAB 2024a to evaluate the proposed approach for safe and efficient autonomous driving at a signal-free intersection. We compare against advanced optimization algorithms, including the Stackelberg game approach [46] and the Nash equilibrium method [28]. Additionally, we include a baseline MCTS [34] implementation (referred to as "Baseline") using standard tree search with fixed thresholds, no risk adaptation, and static cooperation logic. This setup enables direct comparison and highlights our contributions. Parameter settings are listed in Table I, with reward weights empirically tuned for robust safety-efficiency trade-offs.

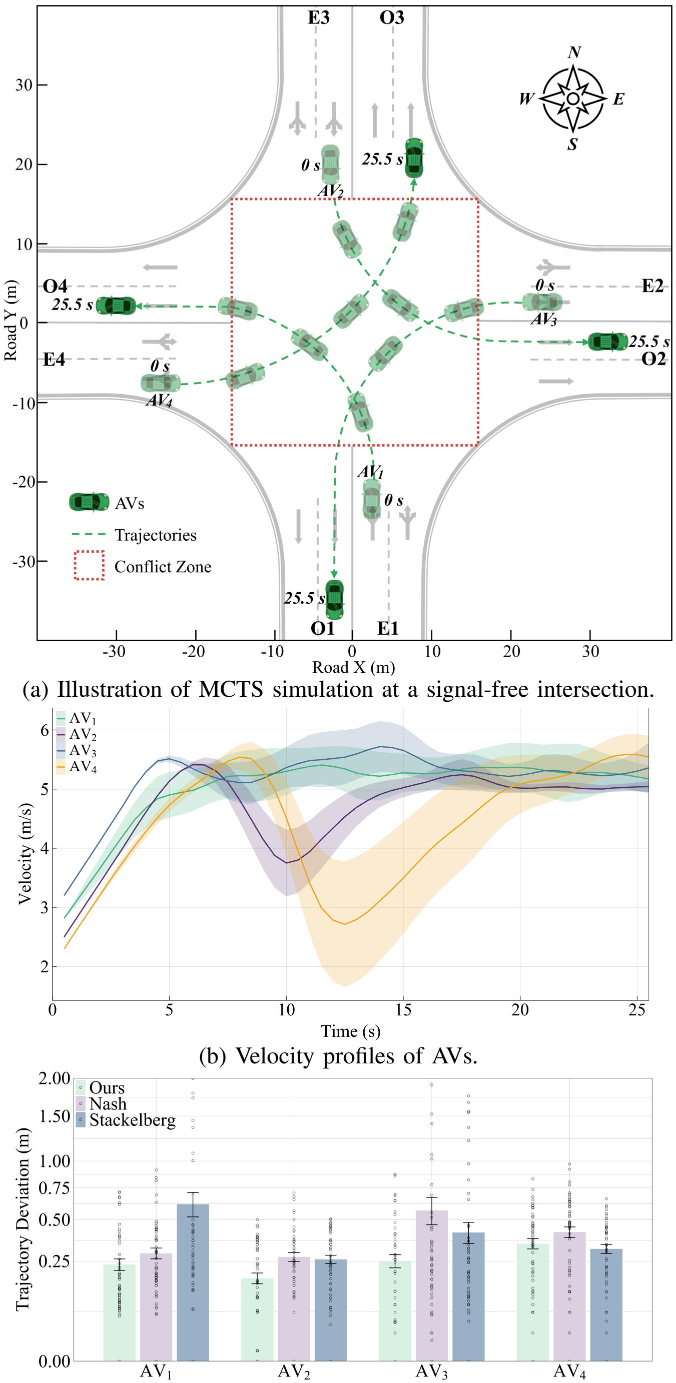
TABLE I: Algorithm Parameters Summary

Category	Parameter	Description	Value
MCTS	γ_r	MCTS reward discount factor	0.95
	β_{dec}	Value update decay	0.75
	c_{exp}	UCB exploration const	1.0
	η	Policy balance factor	0.5
	K	Max iterations	1000
	d_{max}	Max search depth	8
Risk Factors	β_1	Velocity adjustment	0.2
	β_2	Heading adjustment	0.2
	κ_v	Velocity scaling	0.3
	λ_v	Velocity risk scaling factor	0.3
	w_1^{vv}	V2V instant risk weight	0.6
	w_2^{av}	V2V temporal risk weight	0.4
	w_1^{vh}	V2H instant risk weight	0.4
	w_2^{vh}	V2H temporal risk weight	0.3
	w_3^{vh}	Collision prob weight	0.3
	λ_T	Temporal risk weight	0.4
	λ_c	Collision prob weight	0.3
State Uncertainty	σ_x	X position uncertainty	0.1
	σ_y	Y position uncertainty	0.1
	ϵ_x	X uncertainty growth	0.05
	ϵ_y	Y uncertainty growth	0.05
	σ_v	Velocity uncertainty	0.2
	σ_θ	Heading uncertainty	0.1
	ρ_{xv}	X-vel correlation	0.3
	$\rho_{y\theta}$	Y-heading correlation	0.3
	T_p	Prediction horizon	5.0
Reward Weights	w_1	Safety weight	0.5
	w_2	Efficiency weight	0.3
	w_3	Comfort weight	0.2
	w_{jerk}	Jerk penalty	0.1
	w_{yaw}	Yaw rate penalty	0.1
	λ_i	Cooperation coefficient	0.5
Motion Efficiency	w_{vel}	Velocity weight	0.4
	w_{acc}	Acceleration weight	0.3
	w_{ref}	Reference path weight	0.3
	α_v	Overspeeding penalty	1.5
	β_v	Underspeeding penalty	1.0
	v_{tol}	Velocity tolerance	2.0
Safety Thresholds	d_{base}	Base safety distance	1.5
	d_{safe}	Safe distance threshold	2
	d_{min}	Minimum road distance	1.0
	$Q_{\text{th}}^{\text{vv}}$	V2V risk threshold	0.8
	$Q_{\text{th}}^{\text{vh}}$	V2H risk threshold	0.6
Vehicle Limits	v_{max}	Maximum velocity	8.0
	a_{max}	Maximum acceleration	3.0
	a_{med}	Medium acceleration	1.5
	$\dot{\theta}_{\text{max}}$	Maximum steering rate	0.5
	$\dot{\theta}_{\text{med}}$	Medium steering rate	0.25
Driving Style	α_a	Maximum accel. scaling	0.3
	α_T	Headway time scaling	0.4
	α_d	Safety distance scaling	0.3
	β_x	X uncertainty scaling	0.5
	β_y	Y uncertainty scaling	0.5
	β_v	Velocity uncertainty scaling	0.6
	β_θ	Heading uncertainty scaling	0.4

A. Case 1: Signal-Free Intersection (ROP = 100%)

The experimental evaluation begins with a baseline scenario involving an AV rates of penetration (ROP) of 100% at a signal-free intersection. As illustrated in Fig. 9(a), the test scenario involves four AVs approaching a four-way intersection simultaneously from different directions, creating a challenging multi-agent coordination problem. The conflict zone, marked by the red dashed box, represents the critical area where vehicle trajectories intersect.

The velocity profiles shown in Fig. 9(b) demonstrate the effectiveness of our method in maintaining smooth and efficient vehicle motion. The solid lines represent the mean velocities, while the shaded areas indicate the 95% confidence intervals. The profiles reveal that vehicles maintain relatively stable speeds between 3–5 m/s, with only minor velocity



(a) Illustration of MCTS simulation at a signal-free intersection.

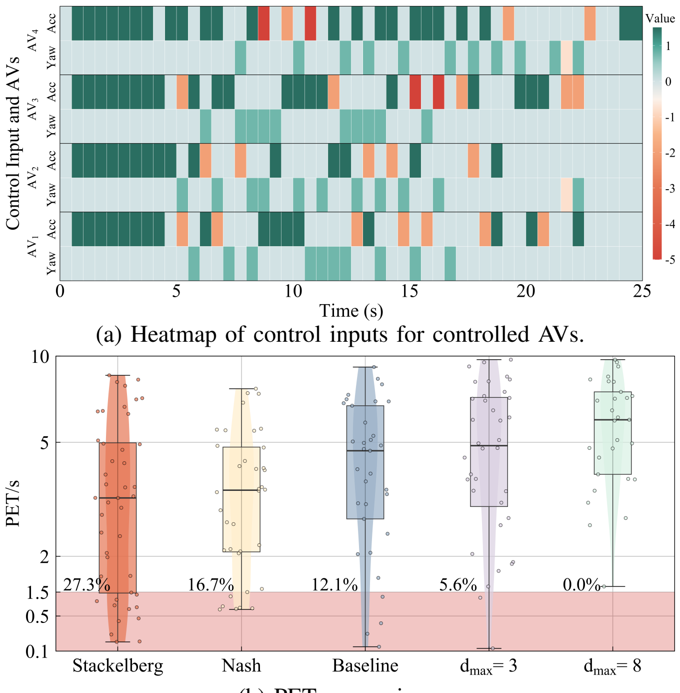
(b) Velocity profiles of AVs.

(c) Comparison of trajectory deviations with advanced approaches.

Fig. 9: Performance analysis in Case 1 (ROP = 100%).

adjustments required for safe coordination. This suggests that our MCTS-based framework effectively balances safety and efficiency without requiring excessive speed reductions.

Trajectory deviation is measured as the root mean square (RMS) distance between the actual vehicle path and a reference trajectory, defined as the shortest smooth path from the current lane to the target lane. Figure 9(c) provides a quantitative comparison of trajectory deviation between our approach and two benchmark methods (Nash and Stackelberg). Our method achieves lower trajectory deviations (mean value of approximately 0.4 m) compared to the Nash (0.6 m) and Stackelberg (0.5 m) approaches, representing significant improvements of 35.26% and 37.56% reduction in overall trajectory deviations, respectively. The marked improvement in path accuracy highlights the superior performance of our MCTS framework in maintaining trajectories under safety constraints, with smaller error bars indicating greater consistency.



(a) Heatmap of control inputs for controlled AVs.

(b) PET comparison.

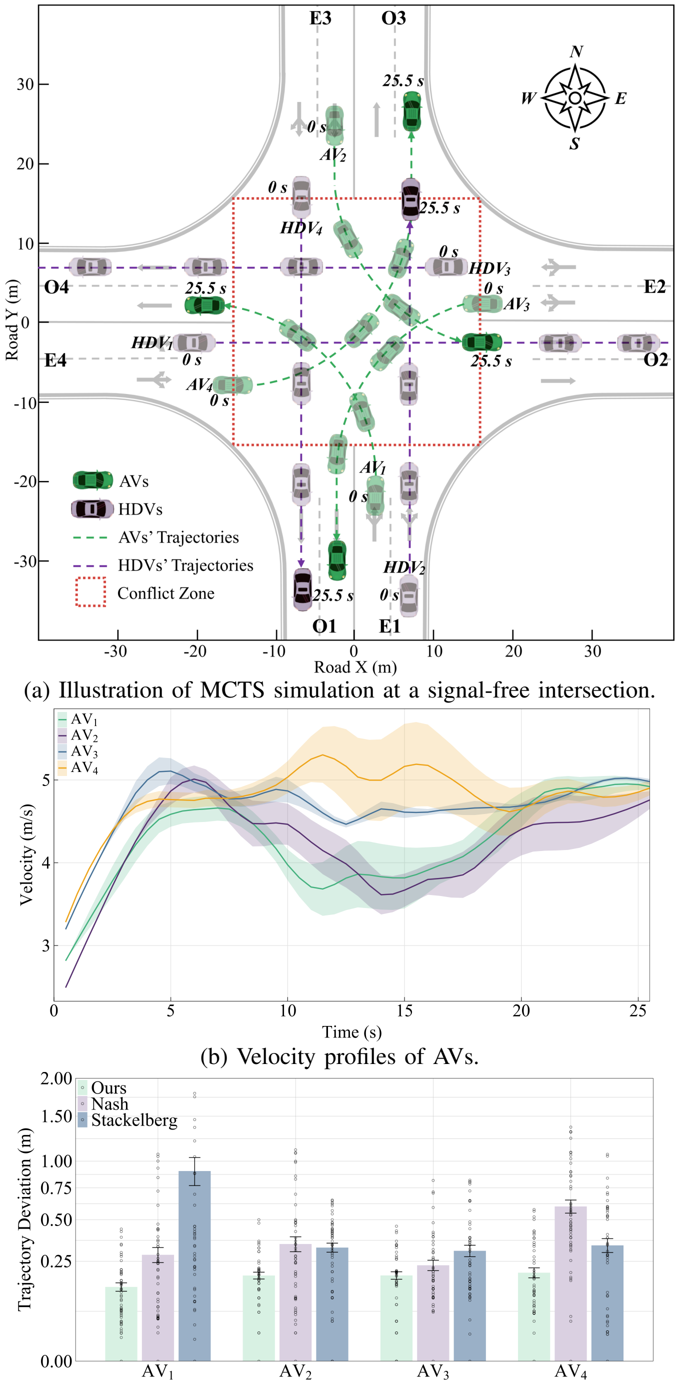
Fig. 10: Analysis of decision-making and safety performance. (a) Variations in control inputs. (b) PET distributions and violations: benchmarks, Baseline (without adaptive risk evaluation), and our methods with $d_{\max}$ of 3 and 8 (ROP = 100%).

TABLE II:
COMPARISON OF ALGORITHM PERFORMANCES IN CASE 1.

Methods	Average Arrive Rate (%)	Average Collision Rate (%)	Average Simulation Time (s)
Stackelberg	76.2 ± 4.5	16.1 ± 5.2	32.7 ± 7.3
Nash	81.6 ± 2.3	11.9 ± 3.6	46.4 ± 9.7
Baseline	83.7 ± 2.8	14.3 ± 5.4	$\mathbf{21.3} \pm 4.2$
Ours $l_{\max} = 3$	89.5 ± 3.4	3.2 ± 2.4	$25.3 \pm \mathbf{3.2}$
Ours $l_{\max} = 8$	$\mathbf{94.1} \pm \mathbf{2.1}$	$\mathbf{0}$	26.2 ± 3.9

The decision-making process is further analyzed through the control input heatmap in Fig. 10(a), which reveals the temporal evolution of acceleration and yaw rate commands for each vehicle. The predominant green coloring indicates that most control actions are moderate, with occasional stronger interventions (darker colors) occurring primarily during critical intersection crossing phases. This pattern demonstrates the framework's ability to generate comfortable trajectories while responding appropriately to dynamic interaction scenarios.

The safety performance is quantitatively evaluated through Post-Encroachment Time (PET) distributions shown in Fig. 10(b). Our method with maximum tree depth $d_{\max} = 8$ achieves the most favorable safety metrics, with zero instances of PET values below the threshold of 1.5 s. This represents



(a) Illustration of MCTS simulation at a signal-free intersection.

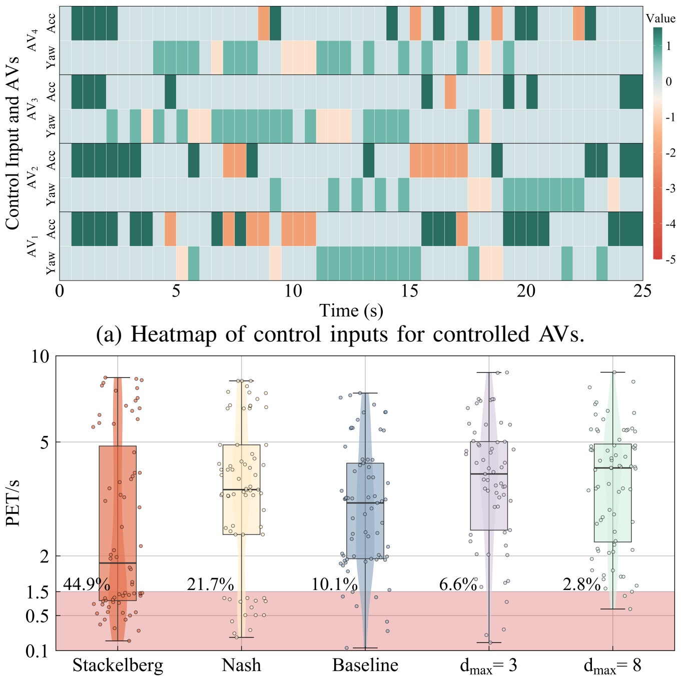
(b) Velocity profiles of AVs.

(c) Comparison of trajectory deviations with advanced approaches.

Fig. 11: Performance analysis in Case 2 (ROP = 50%).

a significant improvement over the baseline approach (12.1% violations), Stackelberg game (27.3% violations), and Nash equilibrium (16.7% violations). The superior performance can be attributed to our risk assessment framework and adaptive safety thresholds that explicitly consider both immediate and predicted vehicle interactions.

Table II demonstrates that our method with $d_{\max} = 8$ achieves the best overall performance, with the highest arrival rate ($94.1 \pm 2.1\%$) and zero collisions. While requiring slightly more computation time (26.2 ± 3.9s) compared to the baseline (21.3 ± 4.2s), this trade-off is justified by the safety improvements over traditional Stackelberg and Nash approaches, which exhibit collision rates of 16.1% and 11.9%.



(a) Heatmap of control inputs for controlled AVs.

(b) PET comparison.

Fig. 12: Analysis of decision-making and safety performance. (a) Variations in control inputs. (b) PET distributions and violations: benchmarks, Baseline (without adaptive risk evaluation), and our methods with $d_{\max}$ of 3 and 8 (ROP = 50%).

B. Case 2: Signal-Free Intersection (ROP = 50%)

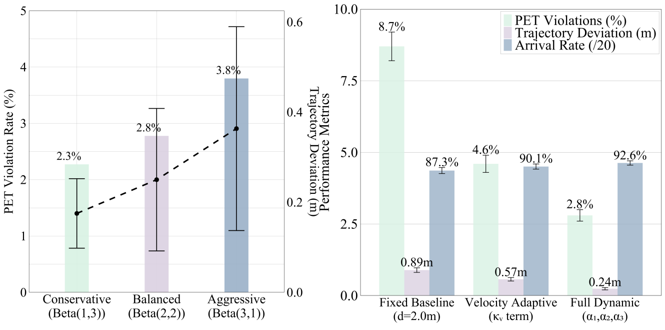
To assess robustness in mixed traffic, we test with a 50% AV penetration rate (ROP = 50%), where four AVs and four HDVs approach the intersection from different directions (Fig. 11(a)). The HDV driving styles are sampled from a $Beta(2, 2)$ distribution, yielding 25% conservative, 50% moderate, and 25% aggressive drivers. This diverse mix introduces realistic challenges by requiring adaptation to both predictable and unpredictable human behaviors.

The velocity profiles presented in Fig. 11(b) demonstrate our method's capability to handle mixed traffic interactions effectively. Compared to the full AV scenario, the velocity variations show larger fluctuations (between 3–5 $\mathrm{m/s}$) with wider confidence intervals, reflecting the increased uncertainty introduced by human drivers. Nevertheless, the profiles maintain overall smooth transitions, indicating that our framework successfully adapts to human driving behaviors while ensuring safe and efficient intersection crossing.

The trajectory deviations shown in Fig. 11(c) reveals the superior performance of our approach compared to benchmark methods. Our method achieves lower trajectory deviations, with improvements of 51.80% and 62.43% compared to the Nash and Stackelberg approaches, respectively. This demonstrates our framework's enhanced capability to handle mixed traffic scenarios through risk assessment and adaptive decision-making strategies.

TABLE III:
COMPARISON OF ALGORITHM PERFORMANCES IN CASE 2.

Methods	Average Arrive Rate (%)	Average Collision Rate (%)	Average Simulation Time (s)
Stackelberg	55.1 ± 4.9	32.4 ± 5.8	41.7 ± 7.5
Nash	59.3 ± 3.7	29.2 ± 5.5	58.4 ± 6.1
Baseline	67.9 ± 5.1	17.1 ± 8.2	$\mathbf{24.9 \pm 3.6}$
Ours $l_{\max} = 3$	85.8 ± 4.7	4.6 ± 2.9	27.3 ± 5.4
Ours $l_{\max} = 8$	$\mathbf{92.6 \pm 3.3}$	$\mathbf{1.7 \pm 0.6}$	29.7 ± 5.1



(a) (b)

Fig. 13: Comparison of PET violations and trajectory deviations of AVs under different HDV driving style distributions.

The control input heatmap in Fig. 12(a) illustrates more diverse and frequent adjustments in AV behaviors compared to the full AV scenario. The increased presence of lighter and darker color patches indicates that AVs make more dynamic adjustments to accommodate the less predictable movements of human drivers. This adaptive behavior demonstrates the framework's capability to balance assertiveness and cautiousness when interacting with HDVs. Safety performance evaluation through PET distributions, shown in Fig. 12(b), reveals the challenges of mixed traffic scenarios. While our method with $d_{\text{max}} = 8$ maintains the best safety performance with only 2.8% PET violations. The baseline approach shows 10.1% violations, while Stackelberg and Nash methods exhibit significantly higher violation rates of 44.9% and 21.7%. These results highlight the increased complexity of ensuring safety in mixed traffic environments.

Table III shows that our method with $d_{\text{max}} = 8$ maintains excellent performance even in mixed traffic scenarios, achieving a $92.6 \pm 3.3\%$ arrival rate and minimal collision rate of $1.7 \pm 0.6\%$. While requiring slightly longer computation time (29.7 ± 5.1s) than the baseline (24.9 ± 3.6s), this represents a dramatic improvement over traditional methods, as both Stackelberg and Nash approaches suffer from high collision rates (32.4% and 29.2% respectively) and low arrival rates in the presence of human drivers.

C. Component Impact Quantification

To validate the effectiveness of our HDV behavior framework with driving style consideration, we conducted a brief ablation study using three typical driving style distributions: conservative-dominated ($\eta_h \sim Beta(1,3)$), balanced ($\eta_h \sim Beta(2,2)$), and aggressive-dominated ($\eta_h \sim Beta(3,1)$). We compared the PET violations and trajectory deviations under these distributions in the 50% AV penetration rate scenario.

1) HDV Driving Style Classification Impact: As shown in Fig. 13(a), our driving style-based HDV behavior classification demonstrates robust performance across different traffic compositions. The aggressive-dominated scenario presented the highest challenges with PET violations of 3.8%, compared to 2.8% in the balanced scenario and 2.3% in the conservative-dominated scenario. These results confirm that an aggressive driving style is associated with a higher safety risk. It can be observed that the PET violation rate exhibited minimal changes under different driving styles, indicating the stability of our framework in ensuring safety. Notably, when the surrounding HDVs adopted a more aggressive driving style, the trajectory deviation of the AV increased. This response illustrates our framework's adaptive risk assessment mechanism, which dynamically adjusts safety thresholds based on detected driving behaviors. When encountering aggressive maneuvers (characterized by higher accelerations and shorter time headways), the framework dynamically expands safety margins and uncertainty bounds as defined in (7) and (15), resulting in more dynamically adjusted planning trajectories that prioritize collision avoidance over accurate reference path following. Such adaptive behavior ensures consistent safety performance despite variations in surrounding HDVs driving styles within highly heterogeneous traffic environments.

2) Dynamic Safety Threshold Impact: Figure 13(b) quantifies the substantial benefits of our adaptive safety threshold mechanism. Compared to fixed baseline thresholds ($d_{\text{safe}} = 2.0$ m), our full dynamic approach reduces PET violations by 67.8% (from 8.7% to 2.8%) and trajectory deviations by 33.9% (from 0.89 m to 0.24 m), while improving arrival rates from 87.3% to 92.6%. The velocity-adaptive configuration achieves intermediate performance with 4.6% PET violations, confirming that the complete integration of relative velocity ($\kappa_v|\Delta v_{ij}|$), heading differences (α_2), and spatial factors (α_3) provides superior safety guarantees. This validates that our context-aware safety margins effectively handle the complex multi-directional conflicts at intersections, where traditional fixed-distance approaches fail to capture the dynamic nature of vehicle interactions. While not explicitly shown in Fig. 13, our adaptive cooperation mechanism $\lambda_i(t)$ contributes significantly to system fairness. The dynamic waiting-time awareness prevents persistent vehicle exclusion while maintaining efficiency, ensuring that no AV is indefinitely delayed due to excessive cooperation with other vehicles. These results demonstrate the effectiveness of our adaptive safety and cooperation mechanisms in balancing multiple objectives across diverse traffic scenarios. Future work will focus on developing meta-learning approaches for automatic parameter tuning based on real-time traffic patterns.

D. Case 3: Signal-Free Intersection (ROP 20% - 100%)

To evaluate safety under varying AV penetration rates, experiments were conducted with ROP ranging from 20% to 100%. Fig. 14 shows the PET distributions, revealing improved safety as AV penetration increases. At low ROPs (20%-

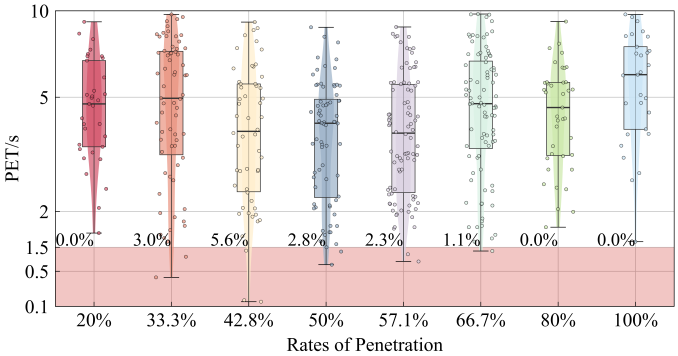


Fig. 14: PET distributions and violations across various ROPs.

33.3%), PET distributions are more variable, with low violation rates (0.0%-3.0%) due to the unpredictability of HDVs. In the medium range (42.8%–57.1%), violation rates initially rise to 5.6% at 42.8%, then stabilize around 2%, reflecting the complexity of mixed traffic. At high ROPs (66.7%-100%), violation rates drop to 0%, and PET distributions narrow, demonstrating consistent safety margins.

VI. Conclusion

This paper presents a safety-critical decision-making framework for AVs at unsignalized intersections, integrating MCTS with risk assessment. The proposed framework demonstrates significant advantages through three key innovations: a multi-agent MCTS structure for efficient action space exploration, a safety assessment mechanism for comprehensive risk evaluation, and an adaptive reward function balancing safety and efficiency. Experimental results validate the framework's effectiveness across different AV penetration rates. In homogeneous scenarios (100% AVs), our approach reduces trajectory deviations by 37.56% compared to benchmark methods while maintaining zero PET violations. The framework shows even more substantial improvements in mixed traffic scenarios (50% AVs), reducing trajectory deviations by 62.43% while effectively handling uncertainties from human drivers. The demonstrated balance between safety and efficiency suggests strong potential for real-world autonomous driving applications. Future work will focus on improving computational efficiency to address the exponential growth in action space with increasing agents, including developing efficient pruning strategies and parallel computation techniques. Additionally, extending the framework to diverse intersection geometries could further enhance its practical applicability. Furthermore, we plan to explore automatic tuning mechanisms for the multi-objective reward function weights to enhance adaptability across diverse traffic conditions. While our fixed-weight implementation has demonstrated good performance, integrating Bayesian optimization or evolutionary algorithms to adaptively adjust these weights based on real-world traffic data could yield more nuanced decision-making that better balances safety, efficiency, and comfort. We also plan to implement a structured hierarchical safety zone concept that triggers differentiated responses based on distinct risk levels, making the system's collision avoidance behavior more interpretable and human-like.

References

[1] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu, "A sufficient condition for convergences of adam and rmsprop," in *Proc. IEEE/CVF CVPR*, 2019, pp. 11 127–11 135.

[2] M. Alyamani and Y. Hassan, "Driver behavior on exit freeway ramp terminals based on the naturalistic driving study," *J. Transp. Eng. A: Syst.*, vol. 149, no. 1, p. 04022120, 2023.

[3] D. Li, J. Zhang, and G. Liu, "Autonomous driving decision algorithm for complex multi-vehicle interactions: An efficient approach based on global sorting and local gaming," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 6927–6937, 2024.

[4] C. Badue, R. Guidolini, R. V. Carneiro, P. Azevedo, *et al.*, "Self-driving cars: A survey," *Expert Syst. Appl.*, vol. 165, p. 113816, 2021.

[5] Z. Tian *et al.*, "Efficient and balanced exploration-driven decision making for autonomous racing using local information," *IEEE Trans. on Intell. Veh.*, pp. 1–17, 2024.

[6] Y. Bie, Y. Ji, and D. Ma, "Multi-agent deep reinforcement learning collaborative traffic signal control method considering intersection heterogeneity," *Transp. Res. Part C: Emerg. Technol.*, vol. 164, p. 104663, 2024.

[7] J. Liu, D. Zhou, P. Hang, Y. Ni, and J. Sun, "Towards socially responsive autonomous vehicles: A reinforcement learning framework with driving priors and coordination awareness," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 827–838, 2024.

[8] R. Vogel, F. Schmidsberger, A. Kühn, K. A. Schneider, *et al.*, "You can't drive my car-a method to fingerprint individual driving styles in a sim-racing setting," in *Proc. ICECET*, 2022, pp. 1–9.

[9] Z. Hong, Q. Lin, and B. Hu, "Knowledge distillation-based edge-decision hierarchies for interactive behavior-aware planning in autonomous driving system," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 9, pp. 11040–11057, 2024.

[10] P. Jardin, I. Moisidis, S. S. Zetina, and S. Rinderknecht, "Rule-based driving style classification using acceleration data profiles," in *Proc. IEEE ITSC*, 2020, pp. 1–6.

[11] C. Chai, X. Shi, Z. Zhou, X. Zeng, W. Yin, and M. M. Islam, *Driving style recognition based on naturalistic driving: Volatilities, decision-making, and safety performances*. Cham: Springer, 2022, pp. 359–394.

[12] X. Hu, Z. Zheng, D. Chen, and J. Sun, "Autonomous vehicle's impact on traffic: Empirical evidence from waymo open dataset and implications from modelling," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6711–6724, 2023.

[13] B. Peng, M. F. Keskin, B. Kulcsár, and H. Wymeersch, "Connected autonomous vehicles for improving mixed traffic efficiency in unsignalized intersections with deep reinforcement learning," *Commun. Transp. Res.*, vol. 1, p. 100017, 2021.

[14] X. Wang *et al.*, "Comprehensive safety evaluation of highly automated vehicles at the roundabout scenario," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 20 873–20 888, 2022.

[15] B. Peng *et al.*, "Communication scheduling by deep reinforcement learning for remote traffic state estimation with bayesian inference," *IEEE Trans. Veh. Technol.*, vol. 71, no. 4, pp. 4287–4300, 2022.

[16] Z. Lin, Z. Tian, J. Lan, Q. Zhang, Z. Ye, H. Zhuang, and X. Zhao, "A conflicts-free, speed-lossless kan-based reinforcement learning decision system for interactive driving in roundabouts," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, pp. 1–14, 2025.

[17] K. Yang *et al.*, "Towards robust decision-making for autonomous driving on highway," *IEEE Trans. Veh. Technol.*, vol. 72, no. 9, pp. 11 251–11 263, 2023.

[18] C. Zhao *et al.*, "A novel direct trajectory planning approach based on generative adversarial networks and rapidly-exploring random tree," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 17 910–17 921, 2022.

[19] Z. Tian, D. Zhao, Z. Lin, D. Flynn, W. Zhao, and D. Tian, "Balanced reward-inspired reinforcement learning for autonomous vehicle racing," in *Proc. L4DC*, 2024, pp. 628–640.

[20] G. Li, J. Yan, Y. Qiu, Q. Li, J. Li, S. E. Li, and P. Green, "Lightweight strategies for decision-making of autonomous vehicles in lane change scenarios based on deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 5, pp. 7245–7261, 2025.

[21] H. Chi, P. Cai, D. Fu, J. Zhai, Y. Zeng, and B. Shi, "Spatiotemporal-restricted a∗ algorithm as a support for lane-free traffic at intersections with mixed flows," *Green Energy Intell. Transp.*, vol. 3, no. 2, p. 100159, 2024.

[22] Z. Lin, Z. Tian, Q. Zhang, H. Zhuang, and J. Lan, "Enhanced visual slam for collision-free driving with lightweight autonomous cars," *Sensors*, vol. 24, no. 19, p. 6258, 2024.

[23] J. Zhu, K. Gao, H. Li, Z. He, and C. O. Monreal, “Bi-level ramp merging coordination for dense mixed traffic conditions,” *Fundam. Res.*, 2023.

[24] Z. e. a. Kherroubi, S. Aknine, and R. Bacha, “Novel decision-making strategy for connected and autonomous vehicles in highway on-ramp merging,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 12 490–12 502, 2022.

[25] H. Wang, H. Gao, S. Yuan, H. Zhao, *et al.*, “Interpretable decision-making for autonomous vehicles at highway on-ramps with latent space reinforcement learning,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 8707–8719, 2021.

[26] J. Zhu, S. Easa, and K. Gao, “Merging control strategies of connected and autonomous vehicles at freeway on-ramps: A comprehensive review,” *J. Intell. Connected Vehicles*, vol. 5, no. 2, pp. 99–111, 2022.

[27] M. Pourabdollah, E. Bjärkvik, F. Fürer, B. Lindenberg, and K. Burgdorf, “Calibration and evaluation of car following models using real-world driving data,” in *Proc. IEEE ITSC*, 2017, pp. 1–6.

[28] P. Hang, C. Huang, Z. Hu, and C. Lv, “Driving conflict resolution of autonomous vehicles at unsignalized intersections: A differential game approach,” *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 6, pp. 5136–5146, 2022.

[29] J. Zhang, S.-C. Chai, B.-H. Zhang, and G.-P. Liu, “Distributed model-free sliding-mode predictive control of discrete-time second-order nonlinear multiagent systems with delays,” *IEEE Trans. Cybern.*, vol. 52, no. 11, pp. 12 403–12 413, 2022.

[30] C. B. Browne, E. Powley, D. Whitehouse, S. M. Lucas, P. I. Cowling, P. Rohlfshagen, S. Tavener, D. Perez, S. Samothrakis, and S. Colton, “A survey of monte carlo tree search methods,” *IEEE Trans. Comput. Intell. AI Games*, vol. 4, no. 1, pp. 1–43, 2012.

[31] D. Lenz, T. Kessler, and A. Knoll, “Tactical cooperative planning for autonomous highway driving using monte-carlo tree search,” in *Proc. IEEE IVS*, 2016, pp. 447–453.

[32] P. Zhou, X. Sun, and T. Chai, “Enhanced nmpc for stochastic dynamic systems driven by control error compensation with entropy optimization,” *IEEE Trans. Control Syst. Technol.*, vol. 31, no. 5, pp. 2217–2230, 2023.

[33] J. Wurts, J. L. Stein, and T. Ersal, “Design for real-time nonlinear model predictive control with application to collision imminent steering,” *IEEE Trans. Control Syst. Technol.*, vol. 30, no. 6, pp. 2450–2465, 2022.

[34] C. F. Hayes, M. Reymond, D. M. Roijers, E. Howley, and P. Mannion, “Risk aware and multi-objective decision making with distributional monte carlo tree search,” *arXiv preprint arXiv:2102.00966*, 2021.

[35] P. Weingertner, M. Ho, A. Timofeev, S. Aubert, and G. Pita-Gil, “Monte carlo tree search with reinforcement learning for motion planning,” in *Proc. ITSC*, 2020, pp. 1–7.

[36] M. Wang *et al.*, “Speed planning for autonomous driving in dynamic urban driving scenarios,” in *Proc. ECCE*, 2020, pp. 1462–1468.

[37] C.-K. Ho and C.-T. King, “Lac-rrt: Constrained rapidly-exploring random tree with configuration transfer models for motion planning,” *IEEE Access*, vol. 11, pp. 97654–97663, 2023.

[38] Y. Gao, D. Li, Z. Sui, and Y. Tian, “Trajectory planning and tracking control of autonomous vehicles based on improved artificial potential field,” *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 12468–12483, 2024.

[39] R. Szczepanski, “Safe artificial potential field: Novel local path planning algorithm maintaining safe distance from obstacles,” *IEEE Robot. Autom. Lett.*, vol. 8, no. 8, pp. 4823–4830, 2023.

[40] X. Shi and X. Li, “Empirical study on car-following characteristics of commercial automated vehicles with different headway settings,” *Transp. Res. Part C: Emerg. Technol.*, vol. 128, p. 103134, 2021.

[41] T. Ravi and D. Siddharth, “Handover count based map estimation of velocity with prior distribution approximated via ngsim data-set,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4352–4361, 2022.

[42] F. Wu, Z. Cheng, H. Chen, Z. Qiu, and L. Sun, “Traffic state estimation from vehicle trajectories with anisotropic gaussian processes,” *Transp. Res. Part C: Emerg. Technol.*, vol. 163, p. 104646, 2024.

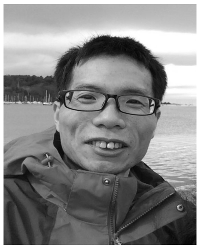
[43] X. Chen, G. Qin, T. Seo, J. Yin, Y. Tian, and J. Sun, “A macro-micro approach to reconstructing vehicle trajectories on multi-lane freeways with lane changing,” *Transp. Res. Part C: Emerg. Technol.*, vol. 160, p. 104534, 2024.

[44] X. Wang, J. Han, Y. Liu, H. Shi, L. Chen, F. Zhong, and S. Liu, “A dynamics model for driving behavior based on coupling actuation of bounded rational cognition and diverse emotions,” *Transp. Res. Part C: Emerg. Technol.*, vol. 159, p. 104479, 2024.

[45] M. Huang, M. Liu, and H. Kuang, “Vehicle routing problem for fresh products distribution considering customer satisfaction through adaptive large neighborhood search,” *Comput. Ind. Eng.*, vol. 190, p. 110022, 2024.

[46] P. Hang, C. Huang, Z. Hu, Y. Xing, and C. Lv, “Decision making of connected automated vehicles at an unsignalized roundabout considering personalized driving behaviours,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 5, pp. 4051–4064, 2021.

Zhihao Lin received the M.S. degree from the College of Electronic Science & Engineering, Jilin University, Jilin, China. He is currently pursuing a Ph.D. degree with the College of Science and Engineering, University of Glasgow, Glasgow, U.K. His main research interests focus on multi-sensor fusion SLAM systems, reinforcement learning, and hybrid control of vehicle platoons.



Jianglin Lan received the Ph.D. degree from the University of Hull in 2017. He has been a Leverhulme Early Career Fellow and Lecturer at the University of Glasgow since 2022. He was a Visiting Professor at the Robotics Institute, Carnegie Mellon University, in 2023. From 2017 to 2022, he held postdoc positions at Imperial College London, Loughborough University, and University of Sheffield. His research interests include artificial intelligence, control theory, and safe autonomy.

Christos Anagnostopoulos received the BSc, MSc, and PhD degrees in computing science from Athens University. He is an Associate Professor at University of Glasgow. His expertise is at the intersection of distributed computing and machine learning. He has received funding by EU H2020/Horizon, EPSRC and industry. He is an author of more than 200 journals/conferences. He serves as general chair of IEEE ICDCS 2025 and editor-in-chief of Open Comput. Sci. (De Gruyter).

Zhen Tian received his bachelor degree in electronic and electrical engineering from the University of Strathclyde, Glasgow, U.K. in 2020. He is currently pursuing the Ph.D. degree with the College of Science and Engineering, University of Glasgow, Glasgow, U.K. His main research interests include Interactive vehicle decision system and autonomous racing decision systems.

David Flynn received the B.Eng. degree (Hons.) in Electrical and Electronic engineering, the M.Sc. degree (Distinction) in Microsystems, and the Ph.D. degree in Microscale Magnetic Components from Heriot-Watt University, Edinburgh, UK, in 2002, 2003, and 2007, respectively. He is a Professor of Cyber Physical Systems at University of Glasgow. He is a co-founder of the UK’s EPSRC National Centre for Energy System Integration and the UK Offshore Robotics and Artificial Intelligence Hub for Offshore Energy Asset Integrity Management.