

RISK-CONSTRAINED SPATIO-TEMPORAL ATTENTION PPO FOR LANE-CHANGING DECISIONS IN MIXED TRAFFIC

ZeYu Chen

China Fire and Rescue Institute, Beijing 102202, China.

Abstract: Lane-changing decisions in mixed traffic must reconcile task completion, surrounding-vehicle interaction, and safety under partial observability. A vanilla Proximal Policy Optimization (PPO) policy that flattens neighboring-vehicle features tends to obscure which vehicles are relevant to the current maneuver, while directly adding a risk penalty to the reward does not provide an explicit budget for cumulative risk. This paper develops a reproducible three-lane Eclipse SUMO environment and proposes a risk-constrained spatio-temporal attention PPO method. The method represents the ego vehicle and up to eight neighboring vehicles as traffic-agent tokens, applies multi-head attention within each frame, aggregates four historical frames with a gated recurrent unit (GRU), and introduces a separate cost critic with a Lagrangian risk budget. The experimental protocol compares vanilla PPO, attention-enabled PPO, risk-constrained PPO, and the full model under identical traffic, reward, and optimizer settings across three training seeds and 20 evaluation episodes per seed. Across the resulting 240 evaluation episodes no policy produced a recorded collision, so the policies were separated by task completion and surrogate risk rather than by collision counts. The full model reached the target lane in 76.7% of its episodes with a mean episode risk cost of 0.040, compared with 46.7% and 0.247 for vanilla PPO, at a mean speed 0.42 m/s lower; attention and the risk constraint each improved completion modestly on their own, and only their combination produced a gain whose direction was consistent across all three seeds. The results support the modular design of the method, and they also show that the simulator safety backstop, rather than the collision count, sets the limit on what this evaluation can claim about safety.

Keywords: Mixed traffic; Lane changing; Proximal policy optimization; Risk-constrained reinforcement learning; Spatio-temporal attention; SUMO; Autonomous driving

1 INTRODUCTION

Lane changing is a compact but safety-critical decision problem in automated driving. A vehicle must decide whether a target lane offers an acceptable gap while accounting for the longitudinal motion of vehicles ahead and behind, the progress of the driving task, and the uncertainty of human-driven traffic. In mixed traffic, heterogeneous car-following and lane-changing behavior makes the transition dynamics non-stationary from the ego vehicle's perspective. An overly conservative policy delays the maneuver and reduces throughput; an aggressive policy can create near-conflict states even when no physical collision occurs.

Deep reinforcement learning provides a natural formulation for this problem because the policy can optimize a sequence of interactions with a microscopic traffic simulator. PPO is attractive for a baseline because its clipped policy objective limits destructive updates while keeping the implementation relatively simple [1]. Nevertheless, a standard multilayer perceptron (MLP) receives a flattened vector and has no explicit mechanism for selecting the neighboring vehicles that matter for the current lane change. A single observation frame also makes it difficult to distinguish a stable gap from a gap that is rapidly closing. Safety constraints are often represented as reward penalties, but reward shaping alone does not guarantee that a policy respects a cumulative risk budget.

This study addresses the following research question: can a PPO lane-changing policy become more interpretable and safety-aware when spatial interaction, temporal evolution, and cumulative risk are represented by separate but coupled modules? The proposed answer is a token-based encoder with neighbor attention and a short temporal GRU, combined with a cost critic and a Lagrangian constraint. The design keeps the action semantics and SUMO interface unchanged, so that the effect of each module can be evaluated by controlled ablations.

The contributions are fourfold:

1. A reproducible SUMO/TraCI mixed-traffic lane-changing environment with explicit observation, action, reward, risk, and safety-backstop definitions.
2. A spatio-temporal encoder that assigns attention weights to neighboring vehicles within each frame and retains four-frame risk evolution with a GRU.
3. A PPO objective with an independent cost critic and a Lagrangian multiplier that limits cumulative surrogate risk rather than only penalizing individual steps.
4. A four-way ablation protocol that isolates the effects of attention, risk constraint, and their combination under matched seeds and training budgets.

2 RELATED WORK

2.1 Reinforcement Learning for Lane-Changing Decisions

Early autonomous-driving decision systems used hand-crafted rules, finite-state machines, or optimization-based behavior planning. These approaches can encode explicit safety margins, but the number of interacting vehicles and the variety of traffic situations lead to increasingly complicated rule sets. Deep reinforcement learning replaces a fixed rule table with a policy learned from state transitions. In lane-changing research, the action is commonly a tactical maneuver such as keep lane, change left, or change right, while a low-level controller executes the lateral and longitudinal motion. This separation is also adopted here: the learned policy issues a high-level lane-change command and SUMO advances the microscopic traffic state.

Shi et al. developed a deep-reinforcement-learning approach for autonomous lane-change decision and control, illustrating the value of learning maneuver policies from traffic interaction rather than prescribing every gap-acceptance rule [2]. Li et al. later emphasized risk awareness in autonomous-vehicle lane-change decision making and showed why safety-related variables should be part of the decision model [3].

Work published since then has extended the formulation in several directions. Cooperative lane changing has been learned for mixed traffic, where the response of human-driven vehicles is modelled explicitly and different leader-follower compositions are enumerated rather than assumed [4]. Hierarchical and human-like policies have been used to decompose the tactical decision into interpretable sub-decisions [5], and planning has been combined with value-based learning so that speed and lane-change decisions are optimized jointly [6]. Robustness to aggressive neighbors has been addressed with safety-aware multi-agent adversarial training [7], while uncertainty about the social preferences of surrounding drivers has been represented as an incomplete-information game [8] and as an observation-robustness problem [9]. Multi-agent formulations have also been used to coordinate longitudinal and lateral control in mandatory lane-change scenarios [10]. These studies raise the behavioural realism of lane-change learning and show that interaction structure matters. Their observation encoders, however, mostly treat the surrounding vehicles as a fixed-length vector in a prescribed order, so the selection of the vehicles that actually constrain the current maneuver remains implicit and is not separately measurable.

2.2 PPO and Constrained Policy Optimization

PPO uses a clipped probability ratio to prevent a new policy from moving too far from the behavior policy in one update [1]. The method is well suited to discrete maneuver actions because a categorical policy can be optimized directly from on-policy rollouts. However, the original PPO objective optimizes a scalar reward and does not distinguish task progress from safety cost. Constrained Policy Optimization (CPO) formulates safe reinforcement learning as a constrained policy-improvement problem [11]. CPO provides a useful conceptual reference, but its trust-region computation is more involved than the small-batch PPO update used in this study. We therefore use a Lagrangian relaxation: a separate cost value function estimates risk returns and a non-negative multiplier adjusts the policy gradient when the episode cost exceeds a fixed budget.

Lagrangian and augmented-Lagrangian variants have since been refined to make constrained policy updates more stable in practice. Dai et al. augmented the Lagrangian function of PPO with a quadratic deviation term to reduce training oscillation and constraint violation [12], and augmented-Lagrangian policies have been applied to cooperative intersection management, where unsafe exploration is the main obstacle to deployment [13]. Related work on intelligent vehicles has handled time-varying state and control constraints with a model-based safe reinforcement learning controller and has imposed a fixed planning horizon so that the constrained return stays tractable in complex scenes [14-15]. Constraint-driven formulations that separate safety from efficiency have also been evaluated for autonomous driving decision making [16]. Surveys of state-wise constraints and of constraint formulations show that instantaneous state constraints and cumulative cost budgets are now treated as distinct design choices, and that the choice between them changes both the learning dynamics and the guarantees that can be stated [17-18].

Risk-aware lane-changing work has also explored probabilistic threat assessment, collision-risk metrics, and scenario-dependent safety policies. The central methodological issue is the distinction between a safety proxy and an observed collision. Time-to-collision (TTC) and gap thresholds can expose near-conflict behavior earlier than collision labels, but they remain proxies whose thresholds must be reported and stress-tested. The present study treats the risk quantity as a surrogate cost, reports it separately from collisions, and does not claim that a zero collision count alone proves the learned constraint is effective. The constrained object is therefore the cumulative episode cost, and the cost definition, budget, and multiplier update are kept explicit so that they can be audited independently of the reward.

2.3 Attention and Interaction-Aware Representation

Self-attention computes a weighted interaction between queries, keys, and values and was introduced as a general sequence-modeling mechanism in the Transformer architecture [19]. For a lane-changing vehicle, the ego state is a natural query and each nearby vehicle can be represented as a key-value token. This design is more structured than flattening all neighbors because it allows the policy to assign different weights to front, rear, current-lane, and target-lane vehicles. Guo et al. used a multi-agent Transformer reinforcement-learning approach to model coupled driving behavior during lane changes [20], while Wang et al. proposed an attention-based deep-reinforcement-learning model for highway lane-change decisions [21]. These studies demonstrate the relevance of interaction-aware representations in this domain.

Subsequent work has used richer scene encoders. Graph convolutions over multi-view agent relations have been combined with value-based learning for cooperative control in highway mixed traffic [22], and transformer scene encoders have been embedded in reinforcement-learning decision agents to improve sample efficiency and adaptability in urban driving [23]. Transformer networks have also been used for multimodal manoeuvre and trajectory prediction on highways, where a distribution over future behavior supports downstream risk assessment [24], and dual-attention recurrent models have been applied to infer lane-change manoeuvres from highway trajectory data [25]. These designs confirm that structured interaction encodings carry information that flat vectors lose, but they are usually evaluated as prediction or representation modules. The incremental value of the attention mechanism inside a constrained control policy is less often isolated.

Attention is not sufficient by itself because a safe gap is dynamic. The same instantaneous distance can represent either a stable situation or a rapidly closing conflict. Temporal models address this ambiguity by retaining recent states. A GRU is a compact gated recurrent architecture that can preserve relevant sequence information with fewer parameters than a full recurrent stack [26]. In this paper, attention summarizes each frame and a GRU aggregates four such summaries. The separation makes the spatial and temporal contributions measurable in the ablation study.

2.4 Surrogate Risk and Safety Measurement

Lane-change safety is normally evaluated with surrogate measures, because physical collisions are rare in simulation and expensive to observe in the field. Time-to-collision and gap-based indicators remain the most common descriptors, and two-dimensional formulations have been used to describe driver evasive behavior in safety-critical lane changes [27]. Their limitation is that they depend on the instantaneous configuration of one or two vehicles: distance and TTC surrogates can miss interactions that unfold across several vehicles, which motivates predictive risk measures that anticipate them [28]. A complementary direction treats risk itself as a learning target, with collision-risk models estimated from large-scale naturalistic driving data instead of hand-set thresholds [29].

Thresholds therefore remain a modelling choice that changes the reported safety level. Simulation studies of mixed and transition-period traffic make the same point: surrogate-based safety conclusions depend on the assumed behaviour models and on the penetration level of automated vehicles [30], and robust control formulations for highway lane changing treat the uncertainty of other drivers explicitly rather than through a tuned threshold [31]. The present study follows the surrogate tradition but reports the cost definition, the gap and TTC thresholds, and the resulting episode cost separately from the collision count, and it treats the surrogate as a relative rather than an absolute safety measure.

2.5 Microscopic Simulation and Mixed Traffic

Microscopic traffic simulators make it possible to repeat controlled scenarios while preserving vehicle-level interactions. SUMO exposes an online Traffic Control Interface (TraCI) that lets an external controller read vehicle states and issue actions at each simulation step [32]. The vehicle-state documentation further specifies how lane-change requests interact with collision avoidance and safety-gap enforcement [33]. These interfaces are important for reproducibility because the simulator, rather than an unreported custom kinematic model, determines the surrounding traffic response.

SUMO has become a common platform for reinforcement-learning studies of mixed traffic, ranging from eco-driving strategies evaluated near signalized intersections to federated multi-agent travel-time minimization [34-35], and simulation-based studies have used it to quantify how the penetration of automated vehicles changes surrogate safety indicators [30]. Proactive assistance strategies for lane changes in mixed traffic have likewise been evaluated in simulation, with the aim of preventing hazardous events rather than reacting to them [36]. These studies share the environmental assumptions used here, which makes them a fair reference point for the present results, but they also show that reported safety benefits depend on the behaviour model assigned to the human-driven vehicles.

The use of a simulator also creates a validity boundary. SUMO safety modes can prevent unsafe requests from becoming collisions, which is desirable for stable training but can mask differences between policies. A rigorous experiment should therefore report collision counts together with surrogate risk, completion, speed, and intervention measures. The environment used here follows that principle and keeps the safety backstop identical across all methods.

2.6 Research Gap and Positioning

Across these strands the literature suggests three complementary requirements: an interaction-aware spatial representation, short-horizon temporal memory, and an explicit cumulative safety constraint [18, 22-23, 28]. Existing lane-change reinforcement-learning studies frequently emphasize the learned policy or the traffic objective, and constrained formulations are often evaluated on intersection management or continuous-control tasks rather than on tactical lane changing in mixed traffic. Attention-based studies, in turn, may not separate a risk budget from the reward, so the contribution of the representation and the contribution of the constraint cannot be read independently. The contribution of this paper is an integrated but deliberately modular design: attention selects the relevant vehicles; a GRU represents the evolution of their relations; and a cost critic constrains cumulative surrogate risk. The four-way evaluation protocol switches the representation module and the constraint module on and off under matched seeds, budgets, and evaluation episodes, so that each effect can be read from the same metric set.

3 PROPOSED RISK-CONSTRAINED SPATIO-TEMPORAL ATTENTION PPO

3.1 Problem Formulation

The lane-changing task is modeled as a discrete-action constrained Markov decision process $M=(S,A,P,R,C)$. The state s_t contains the ego vehicle, nearby traffic, and the target lane. The action a_t is keep lane, change to the higher-index adjacent lane, or change to the lower-index adjacent lane. SUMO defines the transition kernel P . R is the task reward and C is the safety cost. Each episode lasts at most 80 control steps with a 0.5 s step interval.

3.2 SUMO Environment and Scenario

The road is a 2 km one-way segment with three lanes indexed 0, 1, and 2. One ego vehicle starts in the middle lane and receives a randomly selected target lane from the two outer lanes. Background vehicles enter at different times and lanes under a common random seed. The ego vehicle sends high-level lane-change commands through TraCI, while SUMO advances the surrounding vehicles and enforces the lane-change safety mode. The code uses `laneChangeMode=512`, which keeps collision avoidance and safety-gap enforcement active for external requests [33].

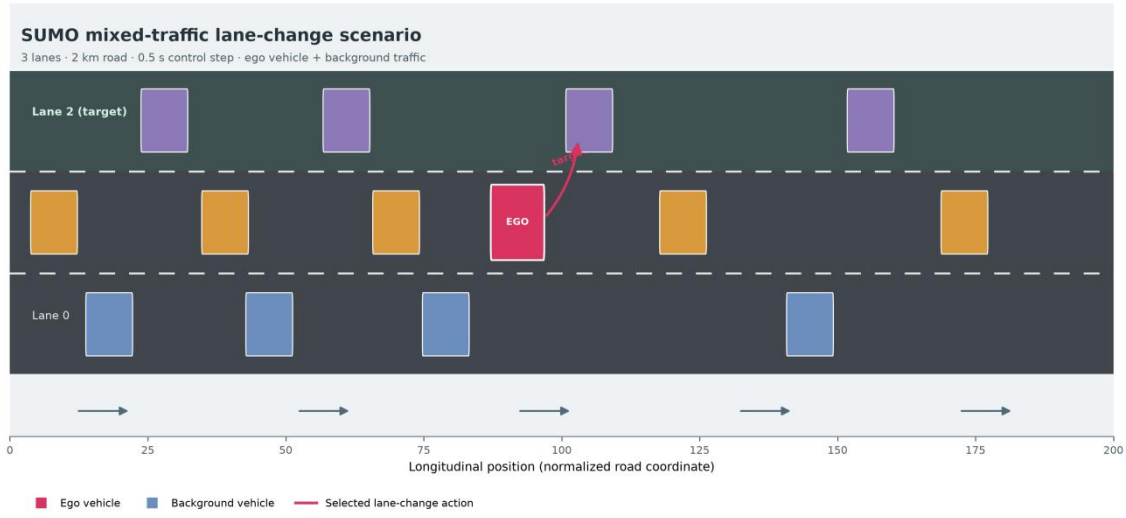


Figure 1 SUMO Three-Lane Mixed-Traffic Lane-Changing Scenario (schematic snapshot; the red vehicle is the ego vehicle and the green band marks the target lane)

Figure 1 visualizes the road and traffic configuration used by the experiment. The graphic is a reproducible scene rendering rather than a performance plot; the actual training interaction is executed by SUMO and TraCI at every control step.

3.3 Observation and Action Representation

At each step, the environment selects up to eight nearest neighboring vehicles using longitudinal distance. Each neighbor token contains relative longitudinal position, relative speed, relative lane, and an occupancy bit. The ego token contains speed, lane index, normalized longitudinal position, target-lane index, and lane difference. Four frames are stacked to form the model input:

$$O_t \in \mathbb{R}^{(4 \times 9 \times 6)} \quad (1)$$

The first dimension is the four-frame history; the second contains one ego token and eight neighbor tokens; and the last dimension is the six-feature token representation. Missing neighbors are zero-padded and remain distinguishable from valid vehicles through the occupancy bit. The action set is $A=\{0,1,2\}$: keep lane, move to the higher-index adjacent lane, and move to the lower-index adjacent lane. Boundary actions are clipped by the environment to a feasible request.

3.4 Reward and Surrogate Risk

The reward balances speed, comfort, risk, successful target-lane completion, collision, and timeout:

$$r_t = -0.02 + 0.02(v_t/30) - 0.1c_t^{\text{comfort}} - 0.5c_t^{\text{risk}} + 8I_{\text{success}} - 20I_{\text{collision}} - 5I_{\text{timeout}} \quad (2)$$

Here v_t is the ego speed in m/s. The comfort term is based on changes in longitudinal acceleration. The episode terminates when the target lane is reached, a collision is observed, or the 80-step horizon is exhausted. The risk cost is a bounded surrogate derived from closing TTC and longitudinal gaps in the current or target lane:

$$c_t^{\text{risk}} = \max(c_t^{\text{TTC}}, c_t^{\text{gap}}), \quad c_t^{\text{TTC}} = \text{clip}((3 - \text{TTC}_{\min})/3, 0, 1) \quad (3)$$

A gap below 5 m receives a cost of 0.75, a gap between 5 and 10 m receives 0.30, and larger gaps receive zero gap cost. TTC contributes only for approaching vehicles and becomes active below 3 s. The cost is intentionally reported as a surrogate risk quantity; it is not calibrated as a probability of collision.

3.5 Spatio-Temporal Attention Encoder

For every historical frame, the ego and neighbor tokens are mapped to a 64-dimensional embedding. The ego embedding supplies the query, whereas the neighboring embeddings supply keys and values:

$$Q = X_{\text{ego}}W_Q, \quad K = X_{\text{neigh}}W_K, \quad V = X_{\text{neigh}}W_V \quad (4)$$

$$\text{Attn}(Q, K, V) = \text{softmax}(QK^T/\sqrt{d_k})V \quad (5)$$

X_{ego} and X_{neigh} denote the ego and neighbor token matrices; W_Q , W_K , and W_V are learnable projections; and d_k is the key dimension per attention head. The four-head attention layer masks padded neighbors using the occupancy bit. The output is the concatenation of the ego embedding and the attention context for that frame. This operation changes the representation from “all vehicles concatenated in a fixed order” to “vehicles weighted according to their current relevance.”

3.6 Temporal Aggregation and Policy Heads

The four frame-level summaries form a sequence of length four and are passed to a GRU with hidden size 128. The last hidden state z_t feeds an actor head with three logits and a reward-value head $V_R(s_t)$. In the risk-enabled variants, a second linear head estimates the cost value $V_C(s_t)$. The full forward path is therefore: observation tensor $\rightarrow$ token embeddings $\rightarrow$ neighbor attention $\rightarrow$ four-frame GRU $\rightarrow$ policy logits, reward value, and optional cost value.

3.7 PPO Objective with a Risk Budget

The PPO probability ratio is defined by:

$$\rho_t(\theta) = \pi_{\theta}(a_t|s_t) / \pi_{\theta_{\text{old}}}(a_t|s_t) \quad (6)$$

Using the generalized-advantage estimate $\hat{A}_t$ and clip parameter $\epsilon=0.2$, the reward objective is:

$$L_{\text{clip}}(\theta) = E_t[\min(\rho_t \hat{A}_t, \text{clip}(\rho_t, 1 - \epsilon, 1 + \epsilon) \hat{A}_t)] \quad (7)$$

Risk-enabled policies estimate a cost advantage $\hat{C}_t$ from the independent cost critic. The policy term becomes:

$$L_{\text{policy}} = -E_t[\min(\rho_t \hat{A}_t, \text{clip}(\rho_t) \hat{A}_t) - \lambda \min(\rho_t \hat{C}_t, \text{clip}(\rho_t) \hat{C}_t)] \quad (8)$$

The multiplier λ is non-negative and is updated after each episode according to the difference between the observed cumulative cost C_{episode} and the budget δ :

$$\lambda \leftarrow \max(0, \lambda + \eta(C_{\text{episode}} - \delta)), \quad \delta = 0.05, \eta = 0.05 \quad (9)$$

When the episode cost exceeds the budget, λ increases and the cost term receives more weight in the policy update. The baseline and attention-only variants set λ and the cost critic to zero, allowing the contribution of the constraint to be measured independently of the encoder.

3.8 Innovations and Complexity

The proposed architecture has three methodological innovations. First, it uses an ego-query attention mechanism instead of a fixed neighbor order, which is aligned with the asymmetric information required for gap acceptance. Second, it applies temporal memory after interaction aggregation, so the recurrent state represents changes in relevant relationships rather than a raw flattened history. Third, it treats safety as a constrained return with its own critic, separating the learning signal for task progress from the learning signal for risk control. With N neighbor tokens, embedding width d , and history length T , the dominant attention term is approximately $O(Nd^2+N^2d)$, while the GRU contributes $O(Td_h^2)$. In this environment $N=8$ and $T=4$ are fixed, so the extra computation remains bounded.

4 EXPERIMENTAL DESIGN AND RESULTS

4.1 Environment and Training Protocol

The experiment uses Eclipse SUMO 1.27.1, TraCI 1.27.1, Python 3.14.0, PyTorch 2.11.0+cu128, and NumPy 2.5.3. Every model is trained for 80 episodes and evaluated for 20 episodes per seed. Three training seeds (101, 102, and 103) are used. The environment, reward, action set, optimizer, horizon, and safety mode are held constant across all methods (Table 1).

Table 1 Environment and Training Configuration

Item	Setting
Traffic simulator	Eclipse SUMO 1.27.1 with TraCI 1.27.1
Road	One-way, three lanes, 2 km
Control interval	0.5 s; maximum 80 steps per episode
Observation	$4 \times 9 \times 6$ token tensor
Action	Keep lane / higher-index lane / lower-index lane
PPO	$\gamma=0.99$; GAE $\lambda=0.95$; learning rate 3×10^{-4} ; $\epsilon=0.2$
Risk constraint	$\delta=0.05$; multiplier step $\eta=0.05$
Replication	3 seeds; 80 training + 20 evaluation episodes per seed

4.2 Baselines and Ablations

The four settings differ only in the representation and risk-constraint modules. Vanilla PPO flattens the observation tensor and uses an MLP. Attention PPO adds neighbor attention and a four-frame GRU but does not use a cost critic. Risk PPO keeps the MLP encoder and adds the cost critic and Lagrangian multiplier. Full PPO combines both modules (Table 2).

Table 2 Controlled Comparison and Ablation Settings

Model	Spatial encoder	Temporal encoder	Cost critic / constraint	Purpose
Vanilla PPO	Flattened MLP	None	No	Algorithm baseline
Attention PPO	Neighbor multi-head attention	Four-frame GRU	No	Representation ablation
Risk PPO	Flattened MLP	None	Yes	Risk-constraint ablation
Full PPO	Neighbor attention	Four-frame GRU	Yes	Joint model

4.3 Evaluation Metrics and Statistical Unit

The evaluation files report six episode-level metrics: episode reward, success, episode risk cost, mean speed, episode steps, and collision. Success means that the ego vehicle reached the target lane before collision or timeout, and episode risk cost is the sum of the surrogate c_t values defined in equation (3). Mean speed is the average ego speed over the steps of an episode, and steps record the episode length, which is bounded by the 80-step horizon. Training seeds, rather than individual evaluation episodes, are treated as independent replication units: the episode-level summaries in Figure 2 and Table 3 describe variability within a seed, whereas Figure 3, Figure 4 and the differences reported with them use the three seed-level means. With three seeds, seed-level standard deviations are reported as a description of run-to-run spread, and no significance test is claimed.

4.4 Main Comparison Across the Four Policies

Each policy was evaluated on 20 episodes per seed, giving 60 episode-level observations per model. Table 3 reports the episode-level means and sample standard deviations for the six metrics, and Figure 2 shows the corresponding distributions.

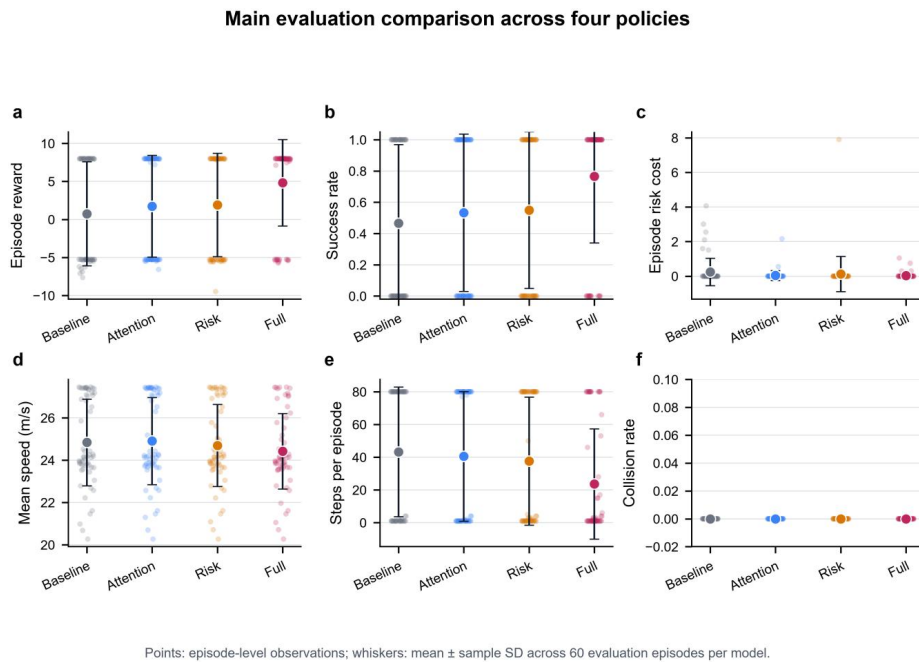


Figure 2 Main Evaluation Comparison Across the Four Policies. (a) Episode reward. (b) Success rate. (c) Episode risk cost. (d) Mean speed. (e) Steps per episode. (f) Collision rate. Points are individual evaluation episodes; whiskers are the mean $\pm$ sample SD across the 60 episodes of each model

Vanilla PPO completed 46.7% of its episodes with a mean episode reward of 0.75, and its reward distribution was wide (SD 6.84) because unsuccessful episodes accumulate a negative terminal term. The two single-module variants improved completion modestly: Attention PPO reached the target lane in 53.3% of episodes with a mean episode risk cost of 0.045, and Risk PPO reached 55.0% with a risk cost of 0.132. The combined model separated itself from both: it completed 76.7% of episodes, recorded the highest mean reward (4.84), and shortened the mean episode from 43.3 steps to 23.7 steps, which is consistent with faster target-lane acquisition. Its mean speed was 0.42 m/s lower than that

of Vanilla PPO, so the completion gain was obtained with a small speed cost rather than with a faster and more risk-seeking policy.

Collision rate was zero for every policy, seed, and episode, 240 episodes in total (Figure 2f). This metric therefore does not order the policies, and the comparison rests on completion and surrogate risk. The consequences of that uniform collision result are examined in Section 4.8.

Table 3 Main Evaluation Comparison

Model	Episode reward	Success rate	Episode risk cost	Mean speed (m/s)	Steps
Vanilla PPO	0.75 ± 6.84	0.467 ± 0.503	0.247 ± 0.800	24.84 ± 2.04	43.3 ± 39.6
Attention PPO	1.75 ± 6.69	0.533 ± 0.503	0.045 ± 0.287	24.90 ± 2.06	40.6 ± 39.6
Risk PPO	1.91 ± 6.78	0.550 ± 0.502	0.132 ± 1.020	24.69 ± 1.94	37.7 ± 39.1
Full PPO	4.84 ± 5.67	0.767 ± 0.427	0.040 ± 0.172	24.42 ± 1.79	23.7 ± 33.7

Note: Episode-level mean and sample standard deviation across 60 evaluation episodes per model (three seeds $\times$ 20 episodes).

4.5 Seed-Level Uncertainty

The episode-level spread in Figure 2 mixes variation between training runs with variation between episodes. Figure 3 aggregates each training run to one mean per metric, so the whiskers describe run-to-run reproducibility.

Success rate was the most reproducible metric for the two policies without a risk constraint: Vanilla PPO and Attention PPO varied by only 0.029 across seeds (0.467 ± 0.029 and 0.533 ± 0.029). The combined model reached the highest mean success rate (0.767) but with the widest spread (SD 0.252), because one seed completed every evaluation episode while the other two reached 0.50 and 0.80. Risk PPO was intermediate in both mean and spread (0.550 ± 0.173). Episode risk cost and mean speed followed the same ordering: the combined model had the lowest seed-level risk cost (0.040 ± 0.035), and its seed-level mean speed (24.42 ± 0.50 m/s) was the only metric on which it was worse than Vanilla PPO (24.84 ± 0.37 m/s). With three seeds these intervals are descriptive, and they show that the run-to-run spread of the full model at the seed level is wider than a single-seed comparison would suggest.

Seed-level uncertainty for the key evaluation metrics

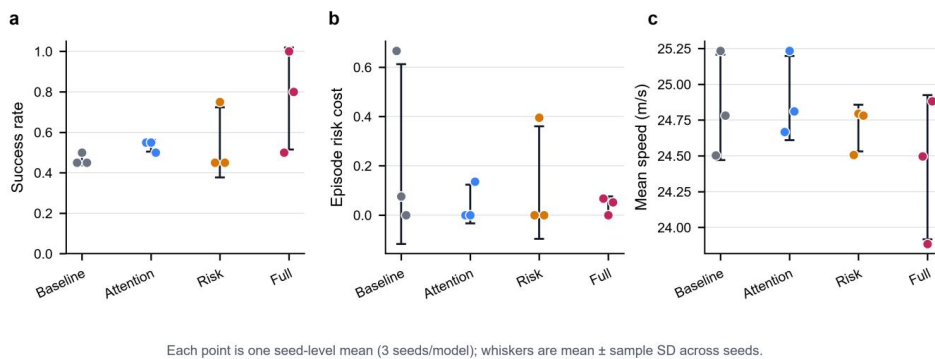


Figure 3 Seed-Level Uncertainty for the Key Evaluation Metrics. (a) Success rate. (b) Episode risk cost. (c) Mean speed. Each point is the mean of one training seed (three seeds per model); whiskers are the mean $\pm$ sample SD across seeds

4.6 Paired Ablation Effects

Figure 4 reports the ablation effects as differences paired by seed against Vanilla PPO, which removes the between-seed component that dominates the episode-level reward distributions of Figure 2.

Attention alone raised the success rate by 0.067 ± 0.029 and lowered the episode risk cost by 0.202 ± 0.416 , with almost no change in mean speed ($+0.065 \pm 0.088$ m/s). The risk constraint alone raised success by 0.083 ± 0.144 and lowered risk cost by 0.116 ± 0.503 , but with a speed cost of 0.145 ± 0.254 m/s and with one seed unchanged on all three metrics. The combination was the only configuration whose success gain was consistent in direction across all three seeds ($+0.300 \pm 0.229$), and it produced the largest reduction in episode risk cost on two of the three seeds. It also produced the largest speed reduction (0.419 ± 0.807 m/s), which indicates that part of the completion gain is a trade of speed for successful maneuvers rather than a uniform improvement in every objective.

The per-seed differences also locate where the modules do not add. For episode risk cost, the attention-only run and the combined model changed the second seed by the same amount (-0.076), and changed the third seed by $+0.136$ and $+0.053$ respectively; the cost advantage of the full model is therefore concentrated in the seed where Vanilla PPO itself recorded its highest cost (0.666). Attention and the risk budget thus act on partly overlapping failure cases: the representation change reduces how often high-cost episodes occur, and the budget limits how large they become.

Paired ablation effects relative to Baseline PPO

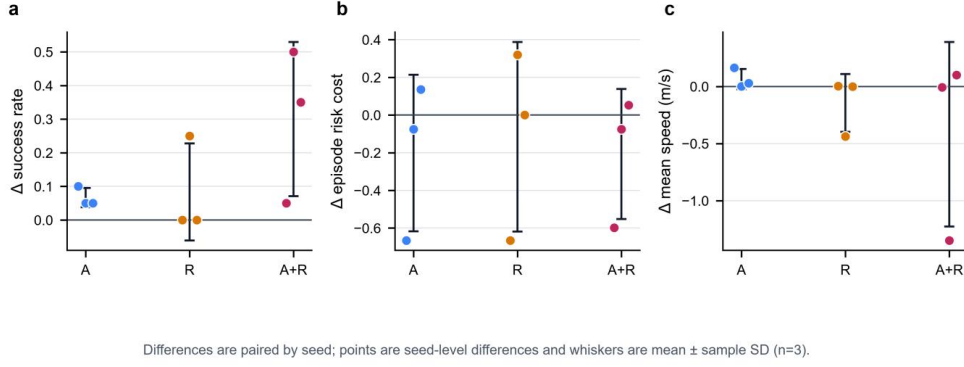


Figure 4 Paired Ablation Effects Relative to Baseline PPO. (a) Change in success rate. (b) Change in episode risk cost. (c) Change in mean speed. Differences are paired by seed; points are seed-level differences and whiskers are the mean $\pm$ sample SD across the three seeds

4.7 Training Dynamics

Figure 5 shows the training curves. All four configurations started from a similar reward level (7.88 to 7.97 over the first ten training episodes), so the differences below arise during training rather than from initialization.

The two risk-enabled variants converged to similar and very stable plateaus: over the last five training episodes the five-episode moving-mean reward was 7.98 ± 0.01 for Risk PPO and 7.97 ± 0.01 for the full model, and their training episode cost fell to zero in every seed (Figure 5b). Vanilla PPO and Attention PPO ended lower and far less consistently (6.37 ± 2.21 and 5.68 ± 2.71), and both retained non-zero training cost until the end of training, with Attention PPO reaching the largest single-seed excursion (a five-episode moving-mean cost of 1.97 at episode 42). The curves therefore separate the policies by stability as well as by level: the cost critic and multiplier coincide with the flat reward plateau, whereas the attention encoder shows its effect at evaluation time rather than in the training trajectory.

Whether the constraint is binding at convergence cannot be read from these curves alone, because the multiplier trajectory and the separate cost returns are not plotted here. They are reported as a reporting improvement in Section 5.2.

Training dynamics across three random seeds

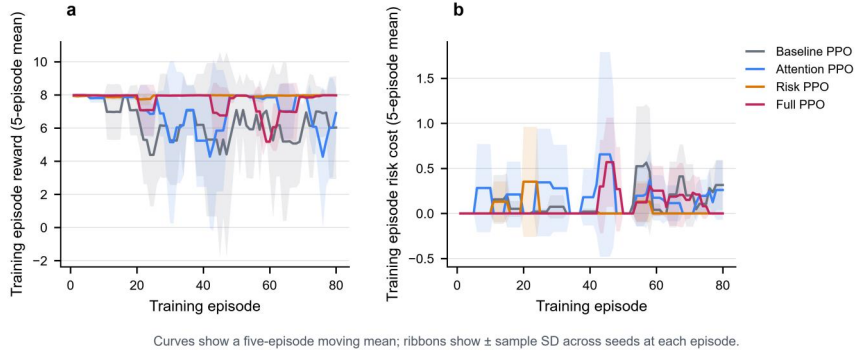


Figure 5 Training Dynamics Across Three Random Seeds. (a) Training episode reward. (b) Training episode risk cost. Curves are five-episode moving means; ribbons are $\pm$ sample SD across seeds at each episode

4.8 Validity of the Safety Evidence

All 240 evaluation episodes ended without a recorded collision, for every policy and every seed (Figure 2f). This uniform result has two consequences for how the comparison should be read.

First, the environment applies the SUMO lane-change safety mode with laneChangeMode=512, which keeps collision avoidance and safety-gap enforcement active for externally requested maneuvers [33]. An unsafe request is therefore modified or rejected by the simulator instead of being executed, so the collision count indicates whether this backstop failed, not whether the learned policy would have been safe without it. Zero collisions across all four policies is consistent with the backstop working as intended, and it is not evidence that the four policies are equally safe. Second, the quantities that separate the policies in this study are the surrogate episode cost, the completion rate, and the episode length. The surrogate is a bounded function of closing TTC and longitudinal gap with fixed thresholds, so its absolute values are threshold choices of this study rather than probabilities of collision.

Intervention events were not logged separately in these experiments, so the frequency with which the simulator modified or rejected a requested lane change is not quantified. This is the main limitation of the present safety evidence, and it is carried into Section 5.2.

5 CONCLUSION

5.1 Summary

This paper formulated mixed-traffic lane changing as a constrained discrete-action decision problem and developed a risk-constrained spatio-temporal attention PPO method. The method uses an ego-query attention layer to select relevant neighboring vehicles, a four-frame GRU to capture the evolution of traffic relations, and an independent cost critic with a Lagrangian risk budget. The SUMO environment, risk surrogate, training protocol, and four-way ablation design were specified so that the effect of each module can be evaluated under matched conditions. Under matched seeds and budgets, the combined model reached the target lane in 76.7% of its 60 evaluation episodes with a mean episode risk cost of 0.040, compared with 46.7% and 0.247 for Vanilla PPO, while no policy produced a recorded collision; the same comparison also showed that the safety-backstop behaviour of the simulator, rather than the collision count, limits what the present evaluation can claim about safety.

5.2 Future Work

Future experiments will vary traffic density, human-driver aggressiveness, target-lane distance, and the penetration rate of automated vehicles. Additional road geometries, including on-ramps and bottlenecks, will test whether the attention weights and risk budget generalize beyond the straight three-lane segment. A further direction is to replace the hand-designed surrogate with a distributional or uncertainty-aware cost model and to evaluate policies with scenario-based stress testing. These extensions can move the study from a controlled module comparison toward a broader validation of safety-aware decision making in mixed traffic.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017. DOI: 10.48550/arXiv.1707.06347.
- [2] Shi T, Wang P, Cheng X, et al. Driving decision and control for automated lane change behavior based on deep reinforcement learning. In: 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019: 2895-2900. DOI: 10.1109/ITSC.2019.8917392.
- [3] Li G, Yang Y, Li S, et al. Decision making of autonomous vehicles in lane change scenarios: Deep reinforcement learning approaches with risk awareness. *Transportation Research Part C: Emerging Technologies*, 2022, 134: 103452. DOI: 10.1016/j.trc.2021.103452.
- [4] Yao X, Du Z, Sun Z, et al. Cooperative lane-changing in mixed traffic: a deep reinforcement learning approach. *Transportmetrica A: Transport Science*, 2024, 22(1). DOI: 10.1080/23249935.2024.2343048.
- [5] Lu C, Lu H, Chen D, et al. Human-like decision making for lane change based on the cognitive map and hierarchical reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 2023, 156: 104328. DOI: 10.1016/j.trc.2023.104328.
- [6] Zhang S, Zhuang W, Li B, et al. Integration of planning and deep reinforcement learning in speed and lane change decision-making for highway autonomous driving. *IEEE Transactions on Transportation Electrification*, 2025, 11(1): 521-535. DOI: 10.1109/TTE.2024.3393293.
- [7] Wang T, Ma M, Liang S, et al. Robust lane change decision for autonomous vehicles in mixed traffic: A safety-aware multi-agent adversarial reinforcement learning approach. *Transportation Research Part C: Emerging Technologies*, 2025, 172: 105005. DOI: 10.1016/j.trc.2025.105005.
- [8] Deng Z, Hu W, Sun C, et al. Eliminating uncertainty of driver's social preferences for lane change decision-making in realistic simulation environment. *IEEE Transactions on Intelligent Transportation Systems*, 2025, 26(2): 1583-1597. DOI: 10.1109/TITS.2024.3512784.
- [9] Yang K, Tang X, Qiu S, et al. Towards robust decision-making for autonomous driving on highway. *IEEE Transactions on Vehicular Technology*, 2023, 72(9): 11251-11263. DOI: 10.1109/TVT.2023.3268500.
- [10] Wang S, Wang Z, Jiang R, et al. A multi-agent reinforcement learning-based longitudinal and lateral control of CAVs to improve traffic efficiency in a mandatory lane change scenario. *Transportation Research Part C: Emerging Technologies*, 2024, 158: 104445. DOI: 10.1016/j.trc.2023.104445.
- [11] Achiam J, Held D, Tamar A, et al. Constrained policy optimization. In: *Advances in Neural Information Processing Systems*, 2017, 30.
- [12] Dai J, Ji J, Yang L, et al. Augmented proximal policy optimization for safe reinforcement learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023, 37(6): 7288-7295. DOI: 10.1609/aaai.v37i6.25888.

- [13] Gao Z, Hao H, Gao F, et al. Constrained reinforcement-learning-enabled policies with augmented Lagrangian for cooperative intersection management. *IEEE Internet of Things Journal*, 2025, 12(5): 5396-5411. DOI: 10.1109/JIOT.2024.3487854.
- [14] Zhang X, Peng Y, Luo B, et al. Model-based safe reinforcement learning with time-varying constraints: Applications to intelligent vehicles. *IEEE Transactions on Industrial Electronics*, 2024, 71(10): 12744-12753. DOI: 10.1109/TIE.2023.3317853.
- [15] Lin K, Li Y, Chen S, et al. Motion planner with fixed-horizon constrained reinforcement learning for complex autonomous driving scenarios. *IEEE Transactions on Intelligent Vehicles*, 2024, 9(1): 1577-1588. DOI: 10.1109/TIV.2023.3273857.
- [16] Gao F, Wang X, Fan Y, et al. Constraints driven safe reinforcement learning for autonomous driving decision-making. *IEEE Access*, 2024, 12: 128007-128023. DOI: 10.1109/ACCESS.2024.3454249.
- [17] Zhao W, He T, Chen R, et al. State-wise safe reinforcement learning: A survey. In: *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023: 6814-6822. DOI: 10.24963/ijcai.2023/763.
- [18] Wachi A, Shen X, Sui Y. A survey of constraint formulations in safe reinforcement learning. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024: 8262-8271. DOI: 10.24963/ijcai.2024/913.
- [19] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems*, 2017, 30: 5998-6008.
- [20] Guo H, Keyvan-Ekbatani M, Xie K. Modeling coupled driving behavior during lane change: A multi-agent Transformer reinforcement learning approach. *Transportation Research Part C: Emerging Technologies*, 2024, 165: 104703. DOI: 10.1016/j.trc.2024.104703.
- [21] Wang J J, Zhang Q C, Zhao D B. Highway lane change decision-making via attention-based deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica*, 2022, 9(3): 567-569. DOI: 10.1109/JAS.2021.1004395.
- [22] Xu D, Liu P, Li H, et al. Multi-view graph convolution network reinforcement learning for CAVs cooperative control in highway mixed traffic. *IEEE Transactions on Intelligent Vehicles*, 2024, 9(1): 2588-2599. DOI: 10.1109/TIV.2023.3297310.
- [23] Liu H, Huang Z, Mo X, et al. Augmenting reinforcement learning with transformer-based scene representation learning for decision-making of autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024, 9(3): 4405-4421. DOI: 10.1109/TIV.2024.3372625.
- [24] Mozaffari S, Alipour Sormoli M R, Koufos K, et al. Multimodal manoeuvre and trajectory prediction for automated driving on highways using transformer networks. *IEEE Robotics and Automation Letters*, 2023, 8(10): 6123-6130. DOI: 10.1109/LRA.2023.3301720.
- [25] Ashfaq F, Ghoniem R M, Jhanjhi N Z, et al. Using dual attention BiLSTM to predict vehicle lane changing maneuvers on highway dataset. *Systems*, 2023, 11(4): 196. DOI: 10.3390/systems11040196.
- [26] Cho K, van Merriënboer B, Gülçehre Ç, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. DOI: 10.48550/arXiv.1406.1078.
- [27] Guo H, Xie K, Keyvan-Ekbatani M. Modeling driver's evasive behavior during safety-critical lane changes: Two-dimensional time-to-collision and deep reinforcement learning. *Accident Analysis & Prevention*, 2023, 186: 107063. DOI: 10.1016/j.aap.2023.107063.
- [28] Choi J H, Kim S. Predictive risk-aware reinforcement learning for autonomous vehicles using safety potential. *Electronics*, 2025, 14(22): 4446. DOI: 10.3390/electronics14224446.
- [29] Jiao Y, Calvert S C, van Cranenburgh S, et al. Learning collision risk proactively from naturalistic driving data at scale. *Nature Machine Intelligence*, 2026, 8(3): 337-350. DOI: 10.1038/s42256-026-01189-w.
- [30] Miqdady T, de Oña R, Casas J, et al. Studying traffic safety during the transition period between manual driving and autonomous driving: A simulation-based approach. *IEEE Transactions on Intelligent Transportation Systems*, 2023, 24(6): 6690-6710. DOI: 10.1109/TITS.2023.3241970.
- [31] Schuurmans M, Katriniok A, Meissen C, et al. Safe, learning-based MPC for highway driving under lane-change uncertainty: A distributionally robust approach. *Artificial Intelligence*, 2023, 320: 103920. DOI: 10.1016/j.artint.2023.103920.
- [32] Eclipse SUMO. TraCI documentation. 2026. <https://sumo.dlr.de/docs/TraCI/>.
- [33] Eclipse SUMO. Change vehicle state: TraCI documentation. 2026. https://sumo.dlr.de/docs/TraCI/Change_Vehicle_State.html.
- [34] Yang Z, Zheng Z, Kim J, et al. Eco-driving strategies using reinforcement learning for mixed traffic in the vicinity of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 2024, 165: 104683. DOI: 10.1016/j.trc.2024.104683.
- [35] Mamond A W, Kundroo M, Yoo S E, et al. FLDQN: Cooperative multi-agent federated reinforcement learning for solving travel time minimization problems in dynamic environments using SUMO simulation. *Sensors*, 2025, 25(3): 911. DOI: 10.3390/s25030911.
- [36] Jo Y, Oh C. A lane change strategy to enhance traffic safety in the coexistence of autonomous vehicles and manual vehicles. *Journal of Advanced Transportation*, 2024, 2024(1): 6126204. DOI: 10.1155/2024/6126204.