

# Integrated Decision and Control: Towards Interpretable and Efficient Driving Intelligence

Yang Guan<sup>1</sup>, Yangang Ren<sup>1</sup>, Shengbo Eben Li<sup>1</sup>, Haitong Ma<sup>1</sup>, Jingliang Duan<sup>1</sup>, Bo Cheng<sup>1</sup>

**Abstract**—Decision and control are two of the core functionalities of high-level automated vehicles. Current mainstream methods, such as functionality decomposition or end-to-end reinforcement learning (RL), either suffer high time complexity or poor interpretability and limited safety performance in real-world complex autonomous driving tasks. In this paper, we present an interpretable and efficient decision and control framework for automated vehicles, which decomposes the driving task into multi-path planning and optimal tracking that are structured hierarchically. First, the multi-path planning is to generate several paths only considering static constraints. Then, the optimal tracking is designed to track the optimal path while considering the dynamic obstacles. To that end, in theory, we formulate a constrained optimal control problem (OCP) for each candidate path, optimize them separately and choose the one with the best tracking performance to follow. More importantly, we propose a model-based reinforcement learning (RL) algorithm, which is served as an approximate constrained OCP solver, to unload the heavy computation by the paradigm of offline training and online application. Specifically, the OCPs for all paths are considered together to construct a multi-task RL problem and then solved offline by our algorithm into value and policy networks, for real-time online path selecting and tracking respectively. We verify our framework in both simulation and the real world. Results show that our method has better online computing efficiency and driving performance including traffic efficiency and safety compared with baseline methods. In addition, it yields great interpretability and adaptability among different driving tasks. The real road test also suggests that it is applicable in complicated traffic scenarios without even tuning.

**Index Terms**—Automated vehicle, Decision and control, Reinforcement learning, Model-based.

## I. INTRODUCTION

Intelligence of automobile technology and driving assistance system has great potential to improve safety, reduce fuel consumption and enhance traffic efficiency, which will completely change the way people travel. High quality perception, decision and control are required by high-level automated vehicles, in which the decision and control are in charge of computing the expected instructions of steering and acceleration relying on the information given by the perception module, to make the automated vehicle drive automatically and satisfy the requirements of safety, compliance, efficiency, comfort and economy. It is generally believed that there are two technical routes for the decision and control of automated vehicles: decomposed scheme and end-to-end scheme.

Decomposed scheme splits the decision and control process into several submodules, such as scene understanding, prediction, behavior selection, trajectory planning and upper control. Each submodule can be designed separately. Scene understanding is to judge characters of road and traffic, e.g. the traffic density and driving aggressiveness of each vehicle, to support parameter tuning for other submodules. Prediction is to predict the future trajectory of traffic participants, so as to construct the feasible region in a certain future time. It is further decomposed into behavior recognition and trajectory prediction [1], [2]. Since the algorithm of behavior recognition and trajectory prediction usually works on each surrounding vehicle, it means that the more the number of vehicles, the more computation is needed. Behavior selection methods are mostly relying on rules, such as finite state machine (FSM) [3]. Based on the current behavior selection results, a collision free space-time curve satisfying vehicle dynamics is calculated according to the predicted trajectories and road constraints by the trajectory planning submodule. There mainly exists three categories of the planning algorithms, i.e., optimization-based, search-based and sample-based. A mainstream algorithm of the optimization-based methods is the model predictive control (MPC), which formulates the planning problem into an optimization problem, where specific indexes of trajectory are optimized and constraints are considered. However, it suffers from long computational time for nonlinear and non-convex problem. The search-based methods represented by A\* are more efficient [4], but they usually lead to low-resolution paths and can barely take dynamic obstacles into consideration. The sample-based methods also have poor computing efficiency because they need to sample points and interpolate them evenly in the whole state space. Xin et al. proposed a combination of the optimization-based and search-based methods, where a trajectory is searched in the space-time by A\* and then smoothed by MPC, yielding the best performance in terms of planning time and comfort. Finally, the upper controller is used to follow the planned trajectory and calculate the expected controls. The decomposed scheme requires large amount of human design, which significantly increases the manpower burden but is still hard to cover all possible driving scenarios. Besides, the real time decision and control cannot be guaranteed because it is time-consuming to complete all the works serially in a limited time for an industrial computer.

End-to-end scheme computes the expected instructions directly from inputs given by perception module using a policy usually carried out by a deep neural network (NN). RL methods do not rely on labelled driving data but learn by trial-and-error in real-world or a high fidelity simulator. Duan et

<sup>1</sup>School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China. All correspondence should be sent to S. Eben Li. <lisb04@gmail.com>.

al. realized decision making under a virtual two-lane highway using hierarchical RL, which designs complicated reward functions for its high-level manoeuvre selection and three low-level manoeuvres respectively [5]. Guan et al. achieved centralized control in a four-leg single-lane intersection with sparse rewards, but took days to find a good driving strategy [6]. Current RL methods are mostly task-specific, which means they have poor adaptability among different driving scenarios and tasks. That is because in each task, a set of complicated reward functions is required to offer guidance for complex driving tasks with long term goal, which is usually non-trivial to design and needs a lot human efforts to tune. Besides, the outcome of the policy is hard to interpret, which makes it barely used in real autonomous driving tasks. Moreover, they cannot deal with safety constraints explicitly and suffer from low convergence speed.

In this paper, we propose an integrated decision and control framework for automated vehicles, which has great interpretability and online computing efficiency, and is applicable to different driving tasks without even tuning. Concretely, we decompose a driving task into multi-path planning and optimal tracking hierarchically. The high-level multi-path planning is used to generate multiple paths only considering static constraints such as road topology, traffic lights, thus can be completed extremely fast. The low-level optimal tracking is used to select the optimal path and track it considering dynamic obstacles. For each path, a finite-horizon constrained optimal control problem is constructed and optimized. The optimal path is selected as the one with the lowest optimal cost function. The key of the optimal tracking is that we unload the heavy computations of path selecting and tracking by an offline training and online application paradigm. Specifically, in offline, we first developed a model-based RL algorithm, which is served as the solver of large-scale constrained optimal control problems (OCP) to obtain the optimal cost function and the optimal control policy in form of NNs, then we formulate the constrained OCPs of all paths into one multi-task RL problem and solve it by the proposed algorithm, leading to interpretable value and policy neural solutions in the sense that they are the approximation of the optimal cost and the optimal action, respectively. Finally, the solved policy and value function are used online for path selecting and tracking, saving time for online optimization. The contributions are summarized as follows.

- 1) We proposed an interpretable hierarchical integrated decision and control framework for automated vehicles, which equips with high online computing efficiency and great adaptability to different tasks.

- 2) We developed a model-based RL algorithm for the purpose of solving large-scale constrained optimal control problem approximately, which to the best of our knowledge, is the first neural solver for constrained OCPs.

- 3) The proposed method is evaluated thoroughly in both simulation and in real-world road. The results shows the potential of the method to be applied in real-world autonomous driving tasks.

## II. HIERARCHICAL INTEGRATED DECISION AND CONTROL FRAMEWORK

In this section, we formulate the framework of hierarchical integrated decision and control. As shown in Fig. 1, the framework consists of two layers: multi-path planning and optimal tracking. Different from the existing schemes, the upper layer aims to generate multiple candidate paths only considering static information such as road structure, speed limit, traffic signs and lights. Note that these paths will not include time information, but each is attached with an expected velocity, which is simply determined by rules and human experience.

The lower layer further considers the static paths generated by the upper layer and the dynamic information such as surrounding vehicles, pedestrians and bicycles. Specifically, for each pair of path and expected velocity, an constrained OCP is designed and optimized, where the objective function is to minimize the tracking error within a finite horizon and the constraints characterize safety requirements. In each time step, the optimal path is chosen as the one with the lowest optimal cost function and tracked as reference. The core of our method is to substitute all the expensive online optimizations with feed-forward of two NNs trained offline by RL. To do that, we first reformulate the constrained OCP for RL, including considering all possible path candidates in the problem, replacing the actions to be optimized by a policy network, and adding a value network to evaluate the cost function. Then we develop a model-based RL algorithm for solving general OCPs and use it in the problem we formulated to obtain the optimal policy and value networks offline. After that, we implement the trained networks online, with the value network identifying the optimal path and the policy network determining the optimal control command.

This framework has several advantages compared to conventional methods. First, it has high online computing efficiency. The upper layer can embed key parameters of multiple paths priorly into the electronic map and read directly for application, which is promise to improve the efficiency of path planning. Furthermore, the lower layer utilizes two trained networks to compute optimal path and control command, which is also time saving due to the extremely fast propagation of neural networks. Second, it can be easily transferred among different driving scenarios without a lot of human design. As the upper layer only uses the static information, the multiple paths can be easily generated by the road topology of various scenes such as intersections, roundabouts and camps. Besides, the lower-layer always formulates a similar tracking problem with safety constraints no matter what the task is, which all can be solved by the developed model-based solver, saving time to design separate reward functions for different tasks. Third, the learned policy and value function are interpretable in the way that they approximate the optimal value and the optimal action of the constrained OCP.

## III. MULTI-PATH PLANNING

This module aims to generate multiple candidate paths for optimal tracking of the lower layer meanwhile maintaining

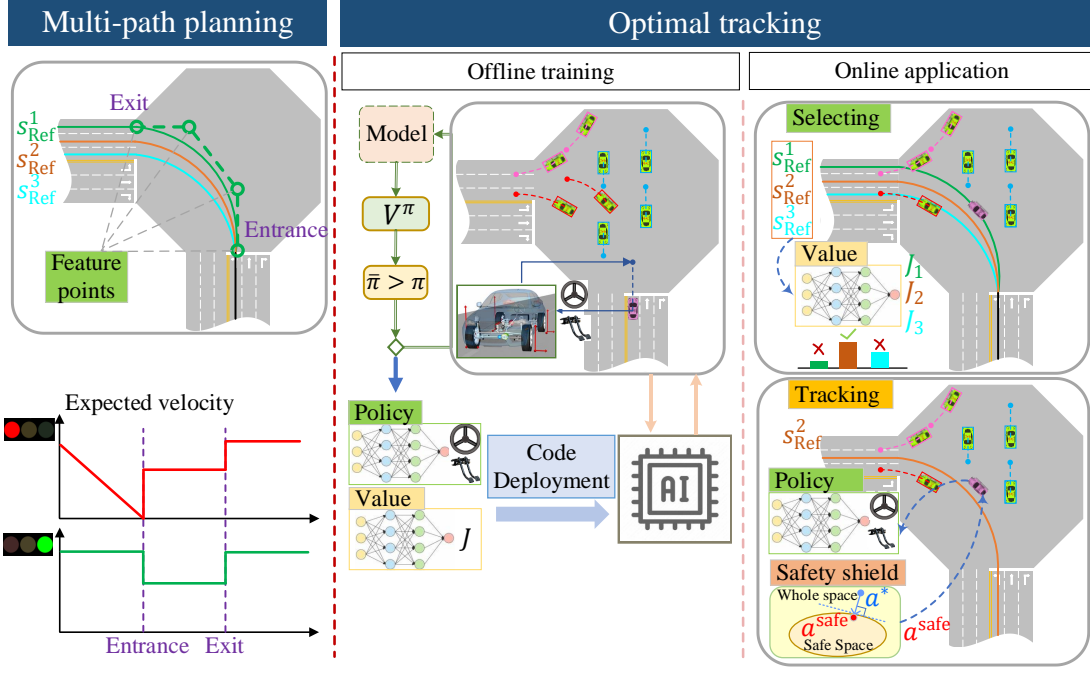


Fig. 1: Hierarchical integrated decision and control framework

high efficiency and feasibility. For that purpose, Cubic Bezier curve is adopted to obtain a continuous and smooth path. Given the road map and driving destination, we will choose four feature points to control the shape of Bezier curve and generate multiple paths with consideration of lane number of destination, as summarised in Algorithm 1. Overall, there exists two methods to generate candidate paths, shown as Fig. 2. One is the static multi-path planning, in which these paths are produced priorly according to road structure and will not change with the riding of ego vehicle. The other one is called dynamic multi-path planning where the start points of reference paths are always line with the current position of ego vehicle. That is, these paths will be generated between ego vehicle and its destination in a real-time manner. Apparently, static multi-path planning is simple and convenient to be directly embedded in the map, and also with a higher efficiency because it only needs to read the pre-stored paths every step. By contrast, dynamic multi-path planning has higher flexibility and may conduct more complex driving behaviors, but its online computing efficiency will be a bit lower than the former. However, both methods will be much more efficient than current existing methods based on searching or optimizing algorithm as the multi-path planning module aims to generate potential candidate paths rather than seek for the optimal path.

As for the expected velocity, we heuristically assign different speed levels with respect to road regions, traffic signals as well as traffic rules such as speed limits or stop signs, which can be quickly designed according to human knowledge, shown as Fig. 2c. Take an intersection with red light as an example, the ego vehicle is expected to slow down before entering the intersection to obey the signal, but drive in certain speeds when it is already in or leaves the intersection so as to pass it carefully. Similar to the candidate paths, the

#### Algorithm 1 Multi-path planning

**Initialize:** target lane number  $N$ , discrete point number  $M$

**for** each target lane **do**

Choose four feature points on map topology:  $(X_0, Y_0)$ ,  $(X_1, Y_1)$ ,  $(X_2, Y_2)$ ,  $(X_3, Y_3)$

**for**  $t = 1 : M$  **do**

$$p_x^{\text{ref}}(t) = X_0(1-t)^3 + 3X_1t(1-t)^2 + 3X_2t^2(1-t) + X_3t^3$$

$$p_y^{\text{ref}}(t) = Y_0(1-t)^3 + 3Y_1t(1-t)^2 + 3Y_2t^2(1-t) + Y_3t^3$$

$$\phi^{\text{ref}}(t) = \arctan\left(\frac{Y(t) - Y(t-1)}{X(t) - X(t-1)}\right)$$

**end for**

Output one path  $\{(p_x^{\text{ref}}, p_y^{\text{ref}}, \phi^{\text{ref}})\}$

**end for**

expected velocity provides a goal for the lower layer to track but not necessarily to follow strictly, so that the ego vehicle seeks to minimize the tracking error while satisfying safety constraints. Actually, it can be simply a fixed value, the lower layer will still always learn a driving policy to balance the safety requirements and tracking errors.

## IV. OPTIMAL TRACKING

### A. Problem formulation

In each time step  $t$ , provided multiple candidate paths generated by the upper layer, the lower layer is designed to first select an optimal path  $\tau^* \in \Pi$  according to a certain criterion, where  $\Pi = \{\tau_i\}_{i=1:N}$  denotes a collection of  $N$  reference paths and  $\tau_i$  is the  $i$ -th candidate path. And then it obtains the control quantities  $u_t$  by optimizing an finite horizon constrained optimal control problem, in which the

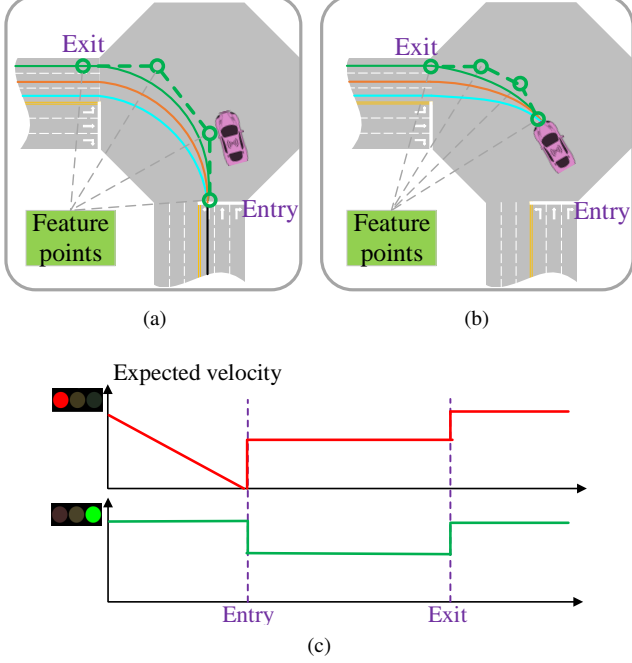


Fig. 2: The upper layer for path planning. (a) Static multi-path planning. (b) Dynamic multi-path planning. (c) Speed planning.

objective is to minimize the tracking error as well as the control energy, and the constraints are to keep a safe distance from dynamic obstacles, as shown in (1),

$$\begin{aligned}
 \min_{u_{i|t}, i=0:T-1} \quad & J = \sum_{i=0}^{T-1} (x_{i|t}^{\text{ref}} - x_{i|t})^\top Q (x_{i|t}^{\text{ref}} - x_{i|t}) + u_{i|t}^\top R u_{i|t} \\
 \text{s.t.} \quad & x_{i+1|t} = F_{\text{ego}}(x_{i|t}, u_{i|t}), \\
 & x_{i+1|t}^j = F_{\text{pred}}(x_{i|t}^j), \\
 & (x_{i|t} - x_{i|t}^j)^\top M (x_{i|t} - x_{i|t}^j) \geq D_{\text{veh}}^{\text{safe}}, \\
 & (x_{i|t} - x_{i|t}^{\text{road}})^\top M (x_{i|t} - x_{i|t}^{\text{road}}) \geq D_{\text{road}}^{\text{safe}}, \\
 & x_{i|t} \leq L_{\text{stop}}, \text{ if light} = \text{red} \\
 & x_{0|t} = x_t, x_{0|t}^j = x_t^j, u_{0|t} = u_t \\
 & i = 0 : T-1, j \in I
 \end{aligned} \tag{1}$$

where  $T$  is the prediction horizon,  $x_{i|t}$  and  $u_{i|t}$  are the ego vehicle state and control in the virtual predictive time step  $i$  starting from the current real time step  $t$ ,  $x_{i|t}^{\text{ref}}$  and  $x_{i|t}^{\text{road}}$  are the closest point from  $x_{i|t}$  on the selected reference  $\tau^*$  and on the road edge, respectively.  $x_{i|t}^j$  is the state of the  $j$ -th vehicle in the interested vehicle set  $I$ .  $Q, R, M$  are positive-definite weighting matrices.  $F_{\text{ego}}$  represents the bicycle vehicle dynamics with linear tire model.  $F_{\text{pred}}$ , on the other hand, is the surrounding vehicle prediction model. Besides,  $D_{\text{veh}}^{\text{safe}}$  and  $D_{\text{road}}^{\text{safe}}$  denote the safe distance from other vehicles and the road edge.  $L_{\text{stop}}$  is the position of stop line. Note that in (1) the virtual states are all produced by the dynamics model and the prediction model except that the  $x_{0|t}$  is assigned with the current real state  $x_t$ . The described variables and functions

above are further defined as:

$$\begin{aligned}
 x_{i|t}^{\text{ref}} &= \begin{bmatrix} p_x^{\text{ref}} \\ p_y^{\text{ref}} \\ v_{\text{lon}}^{\text{ref}} \\ 0 \\ \phi^{\text{ref}} \\ 0 \end{bmatrix}_{i|t} & x_{i|t}^{\text{road}} &= \begin{bmatrix} p_x^{\text{road}} \\ p_y^{\text{road}} \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{i|t} & x_{i|t} &= \begin{bmatrix} p_x \\ p_y \\ v_{\text{lon}} \\ v_{\text{lat}} \\ \phi \\ \omega \end{bmatrix}_{i|t} \\
 x_{i|t}^j &= \begin{bmatrix} p_x^j \\ p_y^j \\ v_{\text{lon}}^j \\ 0 \\ \phi^j \\ 0 \end{bmatrix}_{i|t} & u_{i|t} &= \begin{bmatrix} \delta \\ a \end{bmatrix}_{i|t} & M &= \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \\
 F_{\text{ego}} &= \begin{bmatrix} p_x + \Delta t(v_{\text{lon}} \cos \phi - v_{\text{lat}} \sin \phi) \\ p_y + \Delta t(v_{\text{lon}} \sin \phi + v_{\text{lat}} \cos \phi) \\ v_{\text{lon}} + \Delta t(a + v_{\text{lat}} \omega) \\ \frac{mv_{\text{lon}}v_{\text{lat}} + \Delta t[(L_f k_f - L_r k_r)\omega - k_f \delta v_{\text{lon}} - mv_{\text{lon}}^2 \omega]}{mv_{\text{lon}} - \Delta t(k_f + k_r)} \\ \phi + \Delta t\omega \\ \frac{-I_z \omega v_{\text{lon}} - \Delta t[(L_f k_f - L_r k_r)v_y - L_f k_f \delta v_{\text{lon}}]}{\Delta t(L_f^2 k_f + L_r^2 k_r) - I_z v_{\text{lon}}} \end{bmatrix} \tag{2} \\
 F_{\text{pred}} &= \begin{bmatrix} p_x^j + \Delta t(v_{\text{lon}}^j \cos \phi^j - v_{\text{lat}}^j \sin \phi^j) \\ p_y^j + \Delta t(v_{\text{lon}}^j \sin \phi^j + v_{\text{lat}}^j \cos \phi^j) \\ v_{\text{lon}}^j \\ 0 \\ \phi^j + \Delta t\omega_{\text{pred}}^j \\ 0 \end{bmatrix}
 \end{aligned}$$

where  $p_x, p_y$  are the position coordinates, for ego and other vehicles, it is the position of their respective center of gravity (CG),  $v_{\text{lon}}, v_{\text{lat}}$  are the longitudinal and lateral velocities,  $\phi$  is the heading angle,  $\omega$  is the yaw rate,  $\delta$  and  $a$  are the front wheel angle and the acceleration commands, respectively. The ego dynamics model is discretized by the first-order Euler method, which has been proved to be numerically stable at any low speed [7]. Other vehicle parameters are listed in Table I. The vehicle prediction model is a simple deduction from the current state with constant speed and turning rate  $\omega_{\text{pred}}^j$ , which depends on the driving scenarios and will be specified in the experiments.

The optimal path is chosen as the one with the best tracking performance while satisfying the safety requirements. This means that for each path candidate  $\tau_i \in \Pi$  we ought to construct such a problem (1) and to optimize it to obtain its optimal value  $J_i^*$ . The optimal path is, therefore, the one with minimal  $J$ , i.e.,

$$\tau^* = \arg \min_{\tau} \{J_i^* | \tau_i \in \Pi\} \tag{3}$$

Such a criterion for path selection is consistent with the objective of the optimal control problem (1) and is relatively reasonable in the sense that it optimizes the problem with respect to paths within a finite set, compared to the original optimization with respect to control quantities.

In such framework of selecting and tracking, the lower layer is able to well determine a control quantity that yields good driving efficiency and, meanwhile, the safety guarantee. But, unfortunately, the time complexity is extremely high for an

on-board computing platform of a vehicle, as in each time step we need to solve  $N$  constrained optimal control problem, where  $N$  is the number of candidate paths. However, we argue that this framework naturally corresponds to the actor-critic architecture of RL, where the critic, with the function of judging state goodness, can be served as the path selector while the actor is in charge of action output and used for tracking. With the help of RL, under a paradigm of offline training and online application, the computation burden in the lower layer can be almost nearly eliminated.

## B. Offline training

1) *Solver - Model-based RL*: In this section, we put forward a model-based RL solver for the discrete-time nonlinear constrained OCPs, and then use it to solve the lower layer optimal tracking problem offline for the purpose of online application. Consider the following OCP,

$$\begin{aligned} \min_{a_{i|t}, i=0:T-1} \quad & J = \sum_{t=0}^{T-1} l(s_{i|t}, a_{i|t}) \\ \text{s.t.} \quad & s_{i+1|t} = f(s_{i|t}, a_{i|t}) \\ & g(s_{i|t}) \geq 0 \\ & s_{0|t} = s_t \\ & i = 0 : T - 1 \end{aligned} \quad (4)$$

where  $l$  is the utility function,  $f$  is the system model. For simplicity, we only consider one form of state constraint, i.e.,  $g$ , without losing generality.  $s, a$  denotes a more general form of state and action distinct from that in (1). In order to employ RL to solve it, we notice that there are several significant differences between the OCP and RL problems, originated from the online or offline optimizations. The OCP, which is designed for online optimization, seeks to find an single optimal control quantity of a single state at time step  $t$  while satisfying the constraints with it. RL problems, on the other hand, aim to solve a control policy called actor that maps from state space  $\mathcal{S}$  to control space  $\mathcal{A}$  as well as a value function called critic that evaluates the preference to a state, in an offline style. Thus in them the variable to be optimized is no longer the control quantity but the parameters of the actor and critic. In addition, the objective function and constraints of RL is not about a single state any more, but about a state distribution in the state space. The converted RL problem is shown as:

$$\begin{aligned} \min_{\theta} \quad & J_{\text{actor}} = \mathbb{E}_{s_{0|t}} \left\{ \sum_{i=0}^{T-1} l(s_{i|t}, \pi_{\theta}(s_{i|t})) \right\} \\ \text{s.t.} \quad & s_{i+1|t} = f(s_{i|t}, \pi_{\theta}(s_{i|t})) \\ & g(s_{i|t}) \geq 0 \\ & s_{0|t} = s_t \sim d, \\ & i = 0 : T - 1 \end{aligned} \quad (5)$$

$$\begin{aligned} \min_w \quad & J_{\text{critic}} = \mathbb{E}_{s_{0|t}} \left\{ \left( \sum_{i=0}^{T-1} l(s_{i|t}, \pi_{\theta}(s_{i|t})) - V_w(s_{0|t}) \right)^2 \right\} \\ \text{s.t.} \quad & s_{i+1|t} = f(s_{i|t}, \pi_{\theta}(s_{i|t})) \\ & s_{0|t} = s_t \sim d, \\ & i = 0 : T - 1 \end{aligned} \quad (6)$$

where  $\pi_{\theta} : \mathcal{S} \rightarrow \mathcal{A}$  and  $V_w : \mathcal{S} \rightarrow \mathbb{R}$  are actor and critic, parameterized by  $\theta$  and  $w$  that are generally in form of NNs, respectively. The problem of actor (5) has been proved to be equivalent to the OCP (4) with arbitrary initial state  $s_t$ , provided a powerful approximation function [8]. It means that, given a arbitrary state  $s_t$ , the optimal action  $a^*$  from (4) would be equal to the one mapped by the optimal policy  $\pi_{\theta^*}$  of (5) from  $s_t$ . Consequently, the optimal value  $J^*$  from (4), of course, would be equal to the one mapped by the optimal value function  $V_{w^*}$  of (6) from  $s_t$ , i.e.,

$$\begin{aligned} a^* &= \pi_{\theta^*}(s_t), \quad \forall s_t \in \mathcal{S} \\ J^* &= V_{w^*}(s_t), \quad \forall s_t \in \mathcal{S} \end{aligned} \quad (7)$$

To solve the converted RL problem, we adopt a general iterative framework called policy iteration, wherein two procedures, namely policy evaluation (PEV) and policy improvement (PIM), are alternatively performed to update the critic and actor. Since the critic update is an unconstrained problem that can be optimized by ordinary gradient descent methods, we mainly focus on the actor update which is quite tricky because of its large-scale parameter space, nonlinear property and infinite number of state constraints. To tackle this, we propose a generalized penalty function method adapted from the one introduced in the optimization. It first transforms the constrained problem (5) into an unconstrained one by the exterior penalty function, shown as:

$$\begin{aligned} \min_{\theta} \quad & J_p = \mathbb{E}_{s_{0|t}} \left\{ \sum_{i=0}^{T-1} l(s_{i|t}, \pi_{\theta}(s_{i|t})) \right\} + \rho \mathbb{E}_{s_{0|t}} \left\{ \sum_{i=0}^{T-1} \varphi_i(\theta) \right\} \\ \text{s.t.} \quad & s_{i+1|t} = f(s_{i|t}, \pi_{\theta}(s_{i|t})) \\ & \varphi_i(\theta) = [\max\{0, -g(s_{i|t})\}]^2 \\ & s_{0|t} = s_t \sim d, \\ & i = 0 : T - 1 \end{aligned} \quad (8)$$

where  $\varphi$  is the penalty function,  $\rho$  is the penalty factor. Note that in penalty function method, the second expectation in the objective function should originally be the sum over the state space, but that would make the problem unsolvable. Since the penalty factor before it can be arbitrary scalars, we simply do the replacement without breaking theoretical requirements. After that, we alternatively optimize the policy parameters by gradient descent methods and increase the penalty factor by multiplying a scalar greater than 1 every certain iterations. It can be seen from its way to solve (5) that it has great adaptability to large-scale parameter space and scalability to numerous state constraints. The work flow of our model-based RL solver is summarised in the Fig. 3.

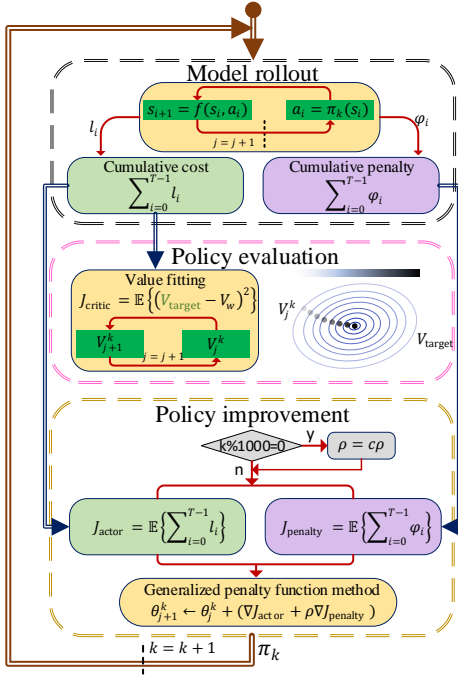


Fig. 3: The work flow of the model-based solver.

2) *Multi-task formulation*: We aim to solve the path selecting and path tracking problems (1) and (3) described in Section IV-A, by our model-based RL solver. The tracking problem (1) is to solve the optimal control given a certain path, while the selection problem (3), on the other hand, is to solve the optimal path given the optimal values. Since there exists a natural correspondence between the tracker and the actor, as well as the selector and the critic, inspired by (7), a naive idea is to design the state  $s$  to include the information of the ego vehicle, surrounding vehicles and the road, and then to train a pair of actor and critic for each candidate path. But obviously, it would become cumbersome and inefficient to train due to the duplicated works. Instead, inspired by the RL problem solving a collection of state all in one go, we also incorporate the information of the path as a part of our state and train only one pair of actor and critic if, as is the case, we know the set of candidate path in advance. It subsequently follows from (7) that given a state of an arbitrary path, the optimal policy and value function can output the optimal control and value under the path. The resulting algorithm for the offline training is shown in 2. Detailed state design will be presented in V-B2.

### C. Online application

Ideally, we expect to get an optimal policy which is able to output the optimal action within the safety action space in any given point in the state space. Unfortunately, it is impossible to acquire such a policy in both theory and practice. First, the converted RL problem (5) enforces constraint on every single state point, leading to an infinite number of constraints in the continuous state space. But nevertheless when we solve the equivalent unconstrained problem (8), we approximate the expectation by an average of samples, that means we only

### Algorithm 2 Optimal tracking - Offline training

**Initialize:** critic network  $V_w$  and actor network  $\pi_\theta$  with random parameters  $w, \theta$ , buffer  $\mathcal{B} \leftarrow \emptyset$ , learning rates  $\beta_w, \beta_\theta$ , penalty factor  $\rho = 1$ , penalty amplifier  $c$ , update interval  $m$

**for** each iteration  $k$  **do**

// Sampling

Randomly select a path  $\tau \in \Pi$ , initialize ego state  $x_t$  and vehicle states  $x_t^j, j \in I$

**for** each environment step **do**

$s_t \leftarrow \{\tau, x_t, x_t^j, j \in I\}$

$\mathcal{B} \cup \{s_t\}$

$a_t = \pi_\theta(s_t)$

Apply  $a_t$  to observe  $x_{t+1}$  and  $x_{t+1}^j, j \in I$

**end for**

// Optimizing (model-based RL solver)

**Model rollout:** Sample a batch of states from  $\mathcal{B}$  and for each state, predict  $\sum_{i=0}^{T-1} l_i$  and  $\sum_{i=0}^{T-1} \varphi_i$  by  $F_{\text{ego}}, F_{\text{pred}}$ , and  $\pi_\theta$

**PEV:**  $w \leftarrow w - \beta_w \nabla_w J_{\text{critic}}$

**PIM:** if  $k \bmod m$ :  $\rho \leftarrow c\rho$ ;  $\theta \leftarrow \theta + \beta_\theta \nabla_\theta J_p$

**end for**

consider finite constraints of the set of sample that can vary in different iterations. So there is no safety guarantee of the policy in a certain state but only an approximately safe performance. Second, the condition of (7) that the approximation function has infinite fitting power cannot be established in practical, resulting in a suboptimal solution without safety guarantee. To ensure the safety performance, we adopt a multi-step safety shield after the output of the policy.

1) *Multi-step safety shield*: The safety shield aims to find the nearest actions in the safe action space in state  $s_t$ , which is formulated as a quadratic programming (QP) problem:

$$a_t^{\text{safe}} = \begin{cases} a_t^*, & \text{if } a_t^* \in \mathcal{A}_{\text{safe}}(s_t) \\ \arg \min_{a \in \mathcal{A}_{\text{safe}}(s_t)} \|a - a_t^*\|^2, & \text{else} \end{cases} \quad (9)$$

where  $a_t^*$  is the policy output. Rather than designing  $\mathcal{A}_{\text{safe}}(s_t)$  to guarantee the safety of only the next state, which may cause the problem to be infeasible, we design it to guarantee that the next  $n_{ss}$  prediction states are safe, i.e., collision-free with the surrounding vehicles and road edges. Formally,

$$\mathcal{A}_{\text{safe}}(s_t) = \{a_t | g(s_{n_{ss}|t}) \geq 0\} \quad (10)$$

2) *Algorithm for online application*: Given the trained policy and value functions are designed to handle arbitrary candidate paths, therefore, in online application, we simply construct a set of states for different paths, then pass them to the trained value function to select the one with the largest value, which is next passed to the trained policy to get the optimal control, as summarised in Algorithm 3.



---

**Algorithm 3** Optimal tracking - Online application
 

---

**Initialize:** Path set  $\Pi$  from upper layer, trained critic network  $V_{w^*}$  and actor network  $\pi_{\theta^*}$ ,  $\lambda$ , ego state  $x_t$  and vehicle states  $x_t^j, j \in I$

**for each environment step do**

// Selecting

**for each**  $\tau_i \in \Pi$  **do**

$s_{t,i} \leftarrow \{\tau_i, x_t, x_t^j, j \in I\}$

$V_i^* = V_{w^*}(s_t)$

**end for**

$\tau^* = \arg \min_{\tau} \{V_i^* | \tau_i \in \Pi\}$

// Tracking

$s_t \leftarrow \{\tau^*, x_t, x_t^j, j \in I\}$

$a_t^* = \pi_{\theta^*}(s_t)$

Calculate  $a_{\text{safe}}^*$  by (9)

Apply  $a_{\text{safe}}^*$  to observe  $x_{t+1}$  and  $x_{t+1}^j, j \in I$

**end for**

---

## V. SIMULATION VERIFICATION

### A. Scenario and task descriptions

We first carried out our experiments on a regular signalized four-way intersection built in the simulation, where the roads in different directions are all the six-lane dual carriageway, as shown in Fig. 4. The junction is a square with a side length of 50m. Each entrance of the intersection has three lanes, each with a width of 3.75m, for turning left, going straight and turning right, respectively. With the help of SUMO [9], we introduce a dense traffic flow by generating 800 vehicles per hour on each lane of the entrance. These vehicles are controlled by the car-following and lane-changing models in the SUMO, producing a variety of traffic behaviors. Moreover, traffic lights in the east-west and north-south streets are included to control the traffic flow of turning left and going straight. We verify our algorithm in three tasks: turn left, go straight and turn right. In each task, the ego vehicle is initialized randomly in the corresponding lane of the south road and is expected to drive safely and efficiently to pass the intersection.

### B. Implementation of our algorithm

1) *Path planning:* In this paper, we adopt the static path planning method to generate multiple candidate paths. Specially, in our scenario, each task will be assigned three paths according to the lane number of and exits. As illustrated in Fig. 2, the paths are generated by the cubic Bezier curve featured by four key points. Two of them are determined by the positions of the entrance and exit, and the other two are auxiliary points located within the junction. In addition to the paths, the expected velocity are chosen as a fixed value for simplicity. We will demonstrate that, even so, a satisfied driving policy could still be well optimized by the lower layer.

2) *State, action and utility function:* In this section, we will introduce detailed settings of the lower layer in this scenario. As mentioned in section IV-B2, the state should be designed

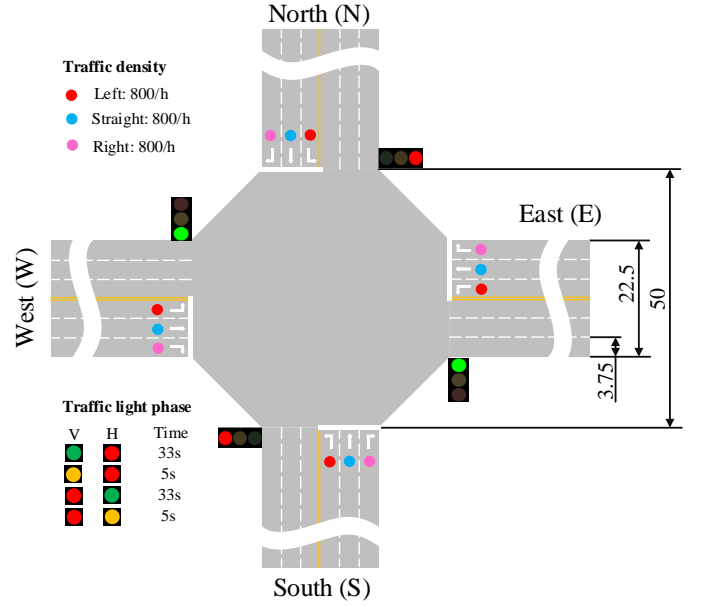


Fig. 4: The scenario used for experiment verification.

to include information of the ego vehicle, the surrounding vehicles and the reference path, i.e.,

$$s_t = [s_t^{\text{ego}}, s_t^{\text{other}}, s_t^{\text{ref}}] \quad (11)$$

where  $s_t^{\text{ego}} = x_t$  is the ego dynamics defined in section IV-A,  $s_t^{\text{other}}$  is the concatenation of the interested surrounding vehicles. Take the turn left task as an example, it defined as:

$$s_t^{\text{other}} = [\delta p_x^j, \delta p_y^j, \phi^j, v_{\text{lon}}^j]_{t,j \in I_{\text{left}}}$$

$$I_{\text{left}} = [\text{SW1}, \text{SW2}, \text{SN1}, \text{SN2}, \text{NS1}, \text{NS2}, \text{NW1}, \text{NW2}] \quad (12)$$

where  $\delta p_x^j, \delta p_y^j$  are relative positions to the ego vehicle.  $I_{\text{left}}$  is an ordered list of vehicles that have potential collision risk with the ego. They are encoded by their respect route start and end, as well as on which the orders. Correspondingly, one can define the go straight and turn right task in a similar way. The information of the reference  $s_t^{\text{ref}}$ , however, is designed in an implicit way by the tracking errors with respect to the position, the heading angle, and the velocity, formulated as:

$$s_t^{\text{ref}} = [\delta_p, \delta_\phi, \delta_v]_t \quad (13)$$

where  $\delta_p$  is the position error,  $|\delta_p| = \sqrt{(p_x - p_x^{\text{ref}})^2 + (p_y - p_y^{\text{ref}})^2}$ ,  $\text{sign}(\delta_p)$  is positive if the ego is on the left side of the reference path, or else is negative.  $\delta_\phi = \phi - \phi^{\text{ref}}$  is the error of heading angle, and  $\delta_v = v_{\text{lon}} - v_{\text{lon}}^{\text{ref}}$  is the velocity error. The overall state design is illustrated in Fig. 5. The action  $a_t = u_t$ , and the weighting matrices in the utility function are designed as  $Q = \text{diag}(0.04, 0.04, 0.01, 0.01, 0.1, 0.02)$ ,  $R = \text{diag}(0.1, 0.005)$ . The predictive horizon  $T$  is set to be 25, which is 2.5s in practical.

3) *Constraint construction:* Slightly different from the one in (1), we further refine the constraint in a way that represents the ego vehicle and each of the surrounding vehicles by two circles as illustrated by Fig. (6), where  $r_{\text{veh}}$  and  $r_{\text{ego}}$  are

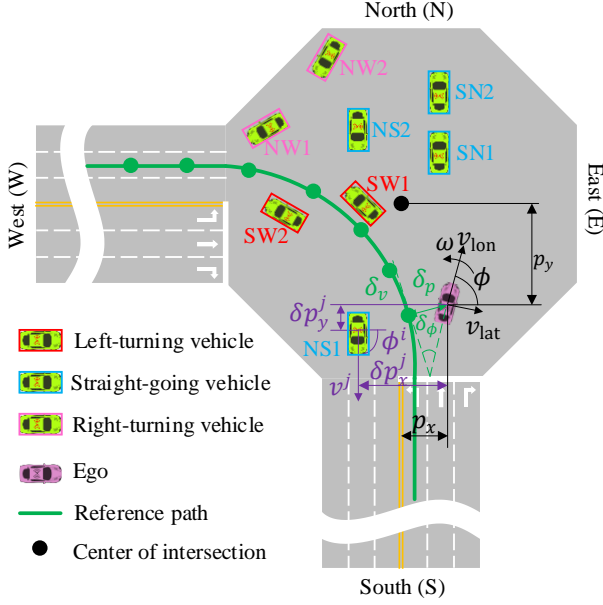


Fig. 5: State design in the scenario.

radii of circles of a vehicle and the ego. Then, in each time step, we impose four constraints for each vehicle between each center of the ego circle and that of the other vehicle rather than between only their CGs. Similarly, in each time step, the number of constraint between the ego and the road edge also increases by one. Parameters for constraints:  $M = \text{diag}(1, 1, 0, 0, 0, 0)$ ,  $D_{\text{veh}}^{\text{safe}} = r_{\text{veh}} + r_{\text{ego}}$ ,  $D_{\text{road}}^{\text{safe}} = r_{\text{ego}}$ . For the traffic light constraint, we convert it to constraints between ego and vehicles by placing two virtual vehicles on the stop line, as shown in Fig. (6).

4) *Vehicle dynamics and prediction model*:  $F_{\text{ego}}$  has been shown in (2), where all the vehicles parameters are displayed in Table. I. Moreover, according to the type and position of the vehicle  $j, j \in I$ , the turning rate  $\omega_{\text{pred}}^j$  in the prediction model is determined, as shown in Table. II. The vehicles to turn left and turn right are assumed to turn in radii of 26.875m and 15.625m, respectively.

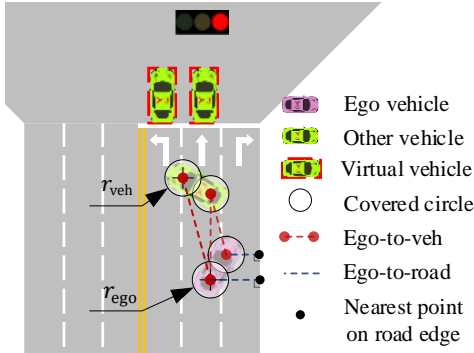


Fig. 6: Design of the state constraints.

5) *Training settings*: We implement our offline training algorithm 2 in an asynchronous learning architecture proposed by Guan et al. [10]. For value function and policy, we use a fully connected neural network (NN) with 2 hidden

TABLE I: Parameters for  $F_{\text{ego}}$

| Parameter  | Meaning                         | Value                     |
|------------|---------------------------------|---------------------------|
| $k_f$      | Front wheel cornering stiffness | -88000 [N/rad]            |
| $k_r$      | Rear wheel cornering stiffness  | -94000 [N/rad]            |
| $L_f$      | Distance from CG to front axle  | 1.14 [m]                  |
| $L_r$      | Distance from CG to rear axle   | 1.40 [m]                  |
| $m$        | Mass                            | 1500 [kg]                 |
| $I_z$      | Polar moment of inertia at CG   | 2420 [kg·m <sup>2</sup> ] |
| $\Delta t$ | System frequency                | 0.1 [s]                   |

TABLE II: Parameters for  $F_{\text{pred}}$

| $\omega_{\text{pred}}^j$ |                | Position relative to the intersection |                            |
|--------------------------|----------------|---------------------------------------|----------------------------|
|                          |                | Out of                                | Within                     |
| Vehicle type             | Left-turning   | 0                                     | $v_{\text{lon}}^j/26.875$  |
|                          | Straight-going | 0                                     | 0                          |
|                          | Right-turning  | 0                                     | $-v_{\text{lon}}^j/15.625$ |

layers, consisting of 256 units per layer, with Exponential Linear Units (ELU) each layer [11]. The Adam method [12] with a polynomial decay learning rate is used to update all the parameters. Specific hyperparameter settings are listed in Table. III. We train 5 different runs of each algorithm with different random seeds on a single computer with a 2.4 GHz 50 core Inter Xeon CPU., with evaluations every 100 iterations.

TABLE III: Detailed hyperparameters.

| Hyperparameters              | Value                                     |
|------------------------------|---|
| Optimizer                    | Adam ( $\beta_1 = 0.9, \beta_2 = 0.999$ ) |
| Approximation function       | MLP                                       |
| Number of hidden layers      | 2   |
| Number of hidden units       | 256                                       |
| Nonlinearity of hidden layer | ELU                                       |
| Replay buffer size           | 5e5                                       |
| Batch size                   | 1024                                      |
| Policy learning rate         | Linear decay 3e-4 $\rightarrow$ 1e-5      |
| Value learning rate          | Linear decay 8e-4 $\rightarrow$ 1e-5      |
| Penalty amplifier $c$        | 1.0                                       |
| Total iteration              | 200000                                    |
| Update interval $m$          | 10000                                     |
| Safety shield $n_{ss}$       | 5   |
| Number of Actors             | 4   |
| Number of Buffers            | 4   |
| Number of Learners           | 30  |

### C. Simulation results

Followed by the multi-path planning algorithm 1, the planned paths are shown in Fig. 7a. We also demonstrate the comprehensive tracking and safety performances of different tasks during the training process, indicated by  $J_{\text{actor}}$  and  $J_{\text{penalty}}$  respectively, and the value loss  $J_{\text{critic}}$  to exhibit the performance of the value function. The training curves are shown in Fig. 7. Along the training process, the policy loss, the penalty and the value loss decrease consistently for all the tasks, indicating an improving tracking and safety performance. Specially, the penalty and value loss decrease to 0 approximately, proving the effectiveness of the proposed RL-based solver for constrained OCPs. In addition, the convergence speed, variance across different seeds, and the final performance vary with tasks. That is because the set of surrounding vehicles that have potential collision risk with the



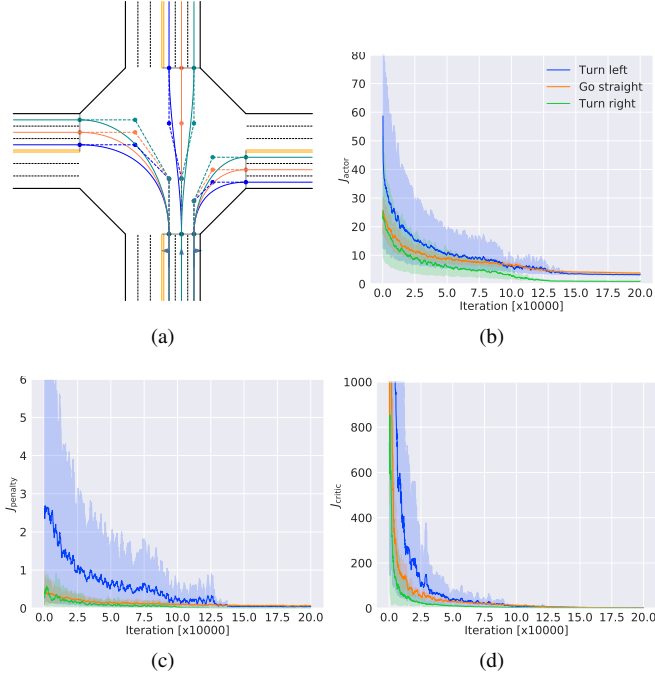


Fig. 7: Results of multi-path planning and optimal tracking. (a) Planned paths for each task. (b) Tracking performance during training process. (c) Safety performance during training process. (d) Value loss during training process. For (b)-(c), The solid lines correspond to the mean and the shaded regions correspond to 95% confidence interval over 5 runs.

ego is different across tasks, leading to the difference in the task difficulty.

In addition, we run the trained policy of the left-turning task in the environment to visualize one typical case where a dense traffic and different phase of traffic light are included, and display its corresponding states, as well as the decision and control process. Note that different from the training, we perform all the simulation tests on a 2.90GHz Intel Core i9-8950HK CPU. As shown in Fig. 8, when the traffic light is red, the ego pulls up to avoid collision and to obey the rule. Then when the light turns green, the ego starts itself off to enter the junction, where it meets several straight-going vehicles from the opposite direction. Therefore, the ego chooses the outer path and slows down to try to bypass the first one. After that, it speeds up to follow the middle path, the one with highest value in that case, with which it can go first and meanwhile avoid the right-turning vehicles so that the velocity tracking error can be largely reduced. The computing time in each step is under 10ms, making our method considerably fast to be applied in the real world.

#### D. Experiment 1: Comparison with MPC

To verify the control precision and computing efficiency of our model-based solver on constrained optimal problem, we conduct comparison with the classic Model Predictive Control (MPC), which utilizes the receding horizon optimization online and can deal with constraints explicitly. Formally, the

problem (1) is defined on one certain path, and thus MPC method will solve the same number of problems as that of candidate paths and calculate cost function of each path respectively. Then optimal path will be selected by the minimum cost function and its corresponding actions will serve as the input signal of ego vehicle. Here we adopt the Ipopt solver [13] to obtain the exact solution of the constrained optimal control problem, which has been an open and popular package to solve nonlinear optimization problem. Fig. 9 demonstrates the comparison of our algorithm on control effect and computation time. Results show that the optimal path of two methods are extremely same and the output actions, steer wheel and acceleration, have similar trends, which indicates our proposed algorithm can approximate the solution of MPC in a small margin error. However, there exists the obvious difference in computation time that our method can output the actions within 10ms while MPC will take 1000ms to perform that. Although MPC can find the optimal solution by its online optimization, its computation time also increases sharply with the number of constraints, probably violating the real-time requirements of autonomous driving.

#### E. Experiment 2: Comparison of driving performance

1) *Baseline algorithms*: Two baseline algorithms are chosen including a rule-based separated hierarchical approach and an model-free reinforcement learning approach. The rule-based approach consists of decision stage and control stage. The decision stage uses an A\* algorithm to generate the spatio-temporal trajectories. The inputs of decision stage is the directed acyclic graph (DAG) consisting the information of the map and surroundings. The control stage adopts a PID controller to track the trajectory. The model-free reinforcement learning approach use a sparse-reward design like the general reinforcement learning tasks. The agent receive a reward of 100 if reaching the desired destination, and gets a penalty of -100 if causing a collision, going out of the road or braking traffic rules. An neural network policy whose input is the state vector of ego and surrounding vehicles, and output is the desired acceleration and steering angle is trained and implemented. Compared to the state space of the proposed model-based approach, the states of tracking error is neglected.

2) *Evaluation indicator*: We choose several indicators to evaluate the performance of our algorithm and baselines, including driving safety, comfort, efficiency, decision compliance, failure rate, and computing efficiency. Driving safety is defined by the number of collisions happened during passing the intersection 100 times. Collision is determined by a six-circle safety distance shown in Fig. x. Safety distance of each vehicle is  $r_{\text{car}} = l_{\text{car}}/6$ , where  $l_{\text{car}}/6$  denotes the vehicle car's length. Distance  $d$  between any two center of different vehicles lower than the sum of their safety distance, i.e.,  $d \leq r_{\text{ego}} + r_{\text{sur}}$ , is counted as one collision. Driving comfort is calculated by the root square of lateral and longitudinal acceleration [14]:

$$I_{\text{comfort}} = 1.4 \sqrt{(a_x^2) + (a_y^2)} \quad (14)$$

The driving efficiency is evaluated by the time used to pass the intersection. We evaluate the decision compliance by the

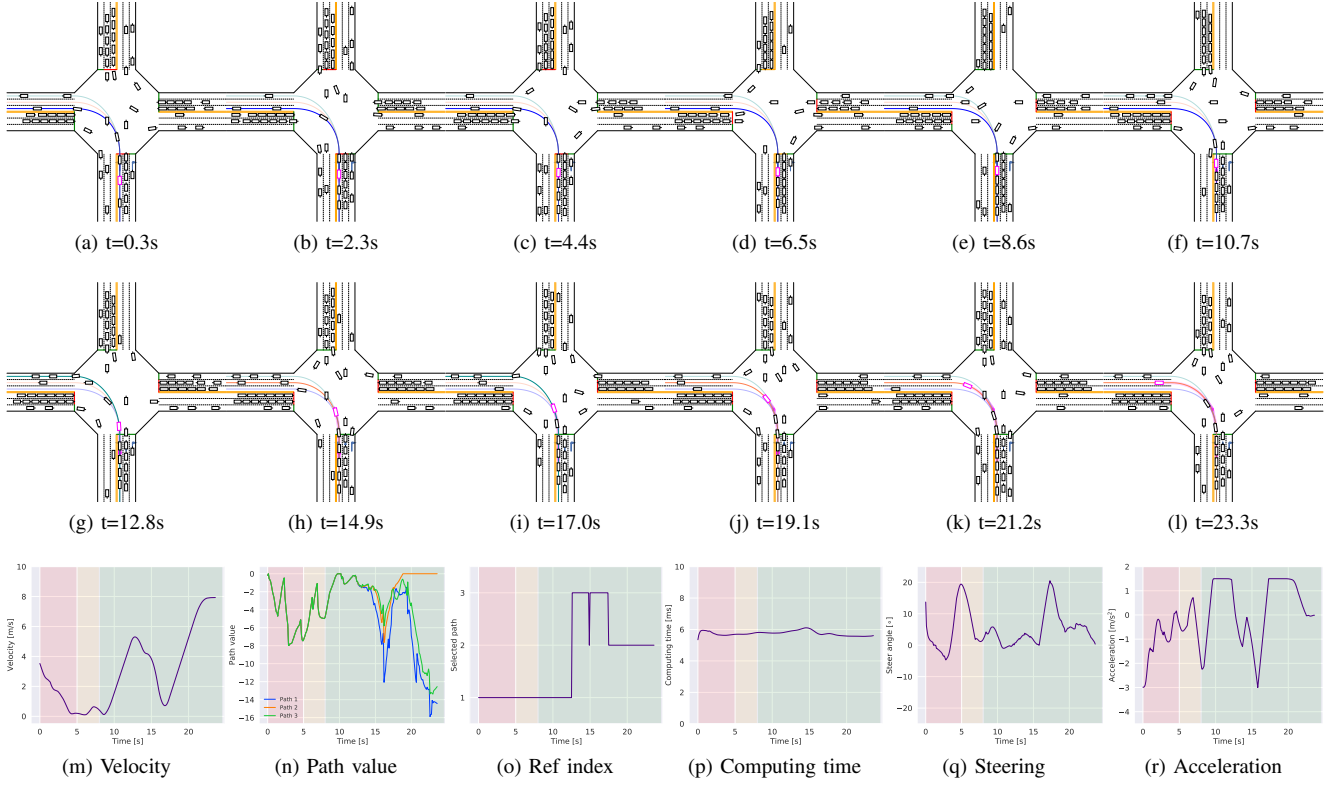


Fig. 8: Visualization of one typical episode driven by the trained policy.

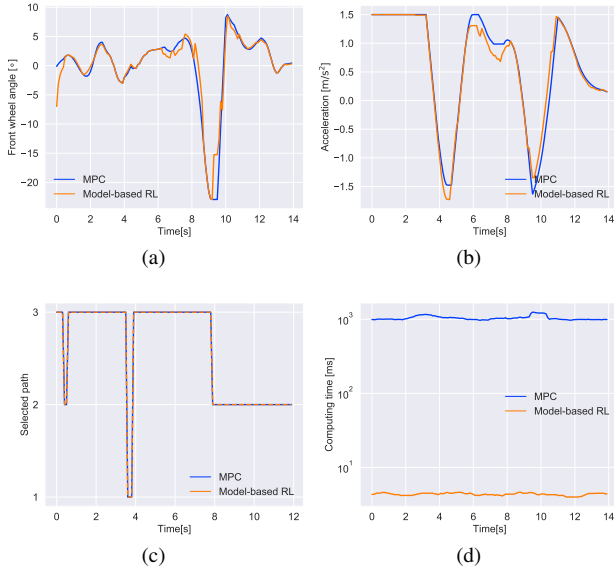


Fig. 9: Comparison with MPC. (a) Front wheel angle. (b) Acceleration. (c) Optimal path. (d) Computing time

number of actions violating the traffic regulations including breaking red light, overspeed, driving on the solid line, etc. The computing efficiency is evaluated by single-step decision and control time. The decision failure rate is counted by the failure times. A single time of passing the intersection is counted as decision failure if the decision algorithm fail to generate the

decision signal for larger than 3 seconds.

3) *Comparison*: A total of 100 times to pass the intersection is recorded to compare the performance of different approaches. The results of driving comfort, efficiency, and computing efficiency are shown in Fig. 10, and the results of driving safety, decision compliance and failure rate are listed in Table IV.

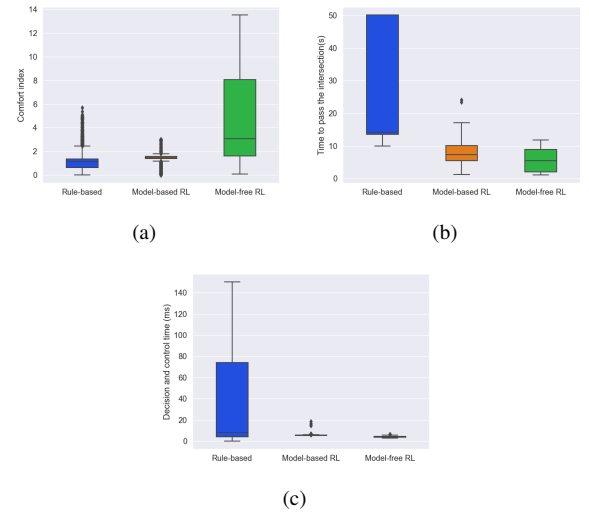


Fig. 10: Comparison of driving performance. (a) Riding comfort. (b) Driving Efficiency. (c) Computing efficiency.

The rule-based decision is more likely to stop and wait

for other vehicles, so the passing time is usually higher and sometimes reaches the upper limit of 50 seconds. The riding comfort of rule-based decision approach is better than model-based one for the same reason, but the difference is small. The model-free controller generally choose reckless actions. This results in the less passing time but worse driving safety and comfort performance. As for the computing efficiency, model-based and model-free approaches have similar computing time since they only need to step forward the neural networks. On the contrary, rule-based algorithm needs online optimization. The computing time increases as the traffic flow gets more complicated.

TABLE IV: Comparison of driving performance.

|                     | Rule-based | Model-based RL | Model-free RL |
|---------------------|------------|----------------|---------------|
| Collisions          | 8          | 3              | 31            |
| Failure Rate        | 13         | -              | -             |
| Decision Compliance | 2          | 3              | 17            |

#### F. Experiment 3: Application of distributed control

We also apply our trained policies in multiple vehicles for distributed control. Our method yields a surprisingly good performance by showing a group of complex and intelligent driving trajectories (Fig. 11), which demonstrates the potential of the proposed method to be extended to the distributed control of large-scale connected vehicles. More videos are available online (<https://youtu.be/J8aRgcjJukQ>).

### VI. TEST ON REAL-WORLD ROADS

#### A. Scenario and equipment

In the real-world test, we choose an intersection of two-way streets located at ( $31^{\circ}08'13''N$ ,  $120^{\circ}35'06''E$ ), shown as Fig. 12a. The east-west street has a eight-lane dual carriageway from both directions, but the north-south street has a only four-lane dual carriageway. The detailed size and functionality of each lane are illustrated in Fig. 12b. Note that we did not utilize the traffic flow and traffic signals of the real intersection, which it actually does not have. Instead, these traffic elements are designed in SUMO, and then provided for the ego vehicle in real time. That means, in fact, we have virtual surrounding vehicles and traffic lights. The experiment vehicle is Changan CS55 equipped with RTK GPS produced by Shanghai CHC Navigation, which realizes precise localization of the ego vehicle. In each time step, the ego states gathered from the CAN bus and the RTK are mapped into SUMO traffic to obtain the current traffic states including the states of surrounding vehicles and traffic signals. Then they both are sent to the industrial computer, where our online optimal tracking algorithm, together with the trained policies and value functions, is embedded. The computer is KMDA-3211 with a 2.6 GHz Intel Core I5-6200U CPU. Processed by our online algorithm, the safe actions including the steering angle and the expected acceleration are then delivered from the CAN bus to the real vehicle for its real time control. The experiment settings are illustrated in Fig. 13. Similar to section V, the ego vehicle enters the intersection from south, and is required to complete the same tasks (turn left, go straight, and turn right) under the signal control and a dense traffic flow.

#### B. Experiment 1: Functionality verification

This experiment is aim to verify the functionality of the hierarchical integrated decision and control framework under different tasks and scenarios. In total, nine runs were carried out, three for each task. In each run, the ego vehicle are initialized before the south entrance with random states, meanwhile, the surrounding vehicles and signals are also initialized randomly. Following the diagram 13, the run keeps going on until the ego passes the intersection successfully, i.e., without colliding with obstacles or breaking traffic rules. The diversity among different runs is guaranteed by using different random seeds. All the videos are available online (<https://youtu.be/adqjor5KXXQ>).

We visualize one of the left run by snapshotting its featured time steps shown in Fig. 14. Besides, we draw its corresponding ego states, including the speed, the yaw rate, the heading angle, the tracking errors, the steering angle and the acceleration; the decision and controls, including the selected path, the expected steering angle and acceleration; and the online computing time. All the curves are shown in Fig. 15. In this run, at the beginning, the ego pull up before the stop line due to the negative expected acceleration, waiting for the green light (Fig. 14a). When that comes, the ego accelerates into the intersection to reduce the velocity tracking error (Fig. 14b). In the center of the intersection, it encounters a straight-going vehicle with high speed from the opposite direction. In order to avoid collision, the ego slows itself down and switches to the path 4, with which it is able to bypass the vehicle from the back. (Fig. 14c). However, another straight-going vehicle comes over after the previous one passes through, but with a relative low speed. This time, the ego no longer waits for it, but chooses to accelerate to pass first. Interestingly, as the vehicle approaches, the optimal path is automatically selected away from it, i.e., changing from the path 4 to the path 3 and finally the path 2, to minimize the tracking errors (Fig. 14c and Fig. 14d). Following the path 2, the ego finally passes the intersection successfully. The computing time of each step is basically within 15ms, showing the superior of our method in terms of the online computing efficiency.

#### C. Experiment 2: Robustness to noise

This experiment is to compare the driving performance under different levels of noises added manually for verifying the robustness of the trained policies. Referring to [15], we take similar measure to divide the noises into 7 levels, i.e. 0-6, where all the noises are in form of Gaussian white noise with different variances varying with the level, as shown in Table. V. We apply the noise in several dimensions of the RL state, including the tracking error in terms of the ego position and heading angle, as well as the positions, velocities, and heading angles of the surrounding vehicles.

We choose the left-turning task to perform seven experiments, one for each noise level in the Table V, to show its influence on the effect of the proposed framework. The random seed is fixed across all experiments. For each noise level, i.e., each experiment, we make statistic analysis on the parameters related to vehicle stability, namely the yaw rate and

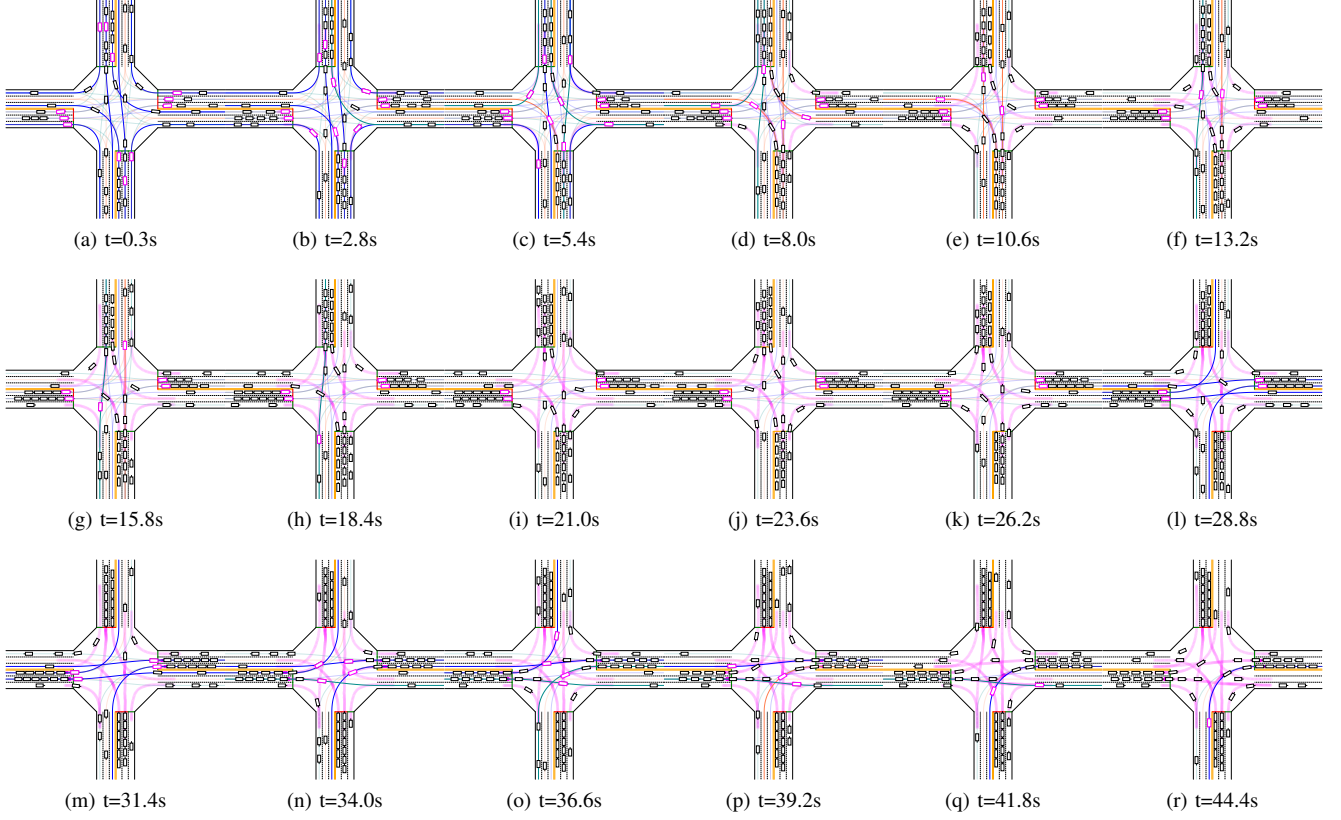


Fig. 11: Demonstration of the distributed control carried out by the trained policies.

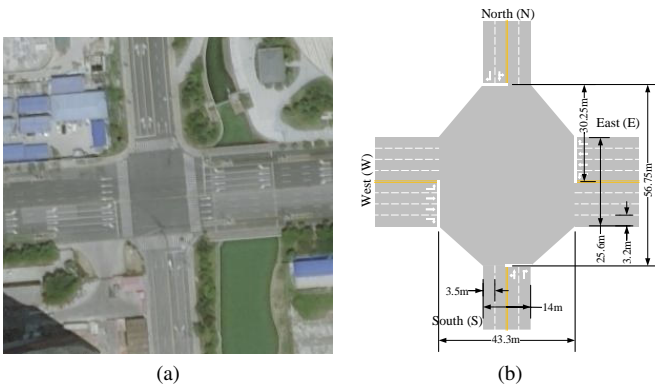


Fig. 12: The intersection for the real test.

lateral speed, and the control quantities, i.e., steering angle and acceleration, as shown in Fig. 16. From it we can see that our method is rather robust to low level noises (0-3) because the data distributions in those noises, including the median value, the standard variance, the quantile values and the bounds, have no significant change. However, these distribution parameters, especially the variance and the bounds, are inevitably enlarged if we add more noise. The fluctuations of the lateral velocity and the yaw rate are mainly caused by the sensitivity of the steering wheel, because large noises tend to yield large variance of the steering angle, which further leads to the swing of the vehicle body. But nevertheless the bounds of

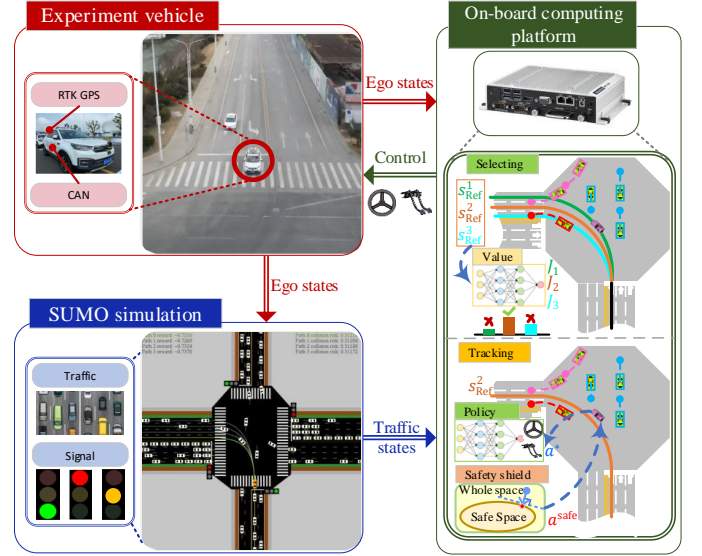


Fig. 13: Diagram of the real-world road test.

the featured values of stability always remain in a reasonable range, proving the robustness of the proposed method.

#### D. Experiment 3: Robustness to human disturbance

The experiment is carried out to verify the ability of the framework to cope with human disturbance. We also use a left-turning case, during which we perform two times of



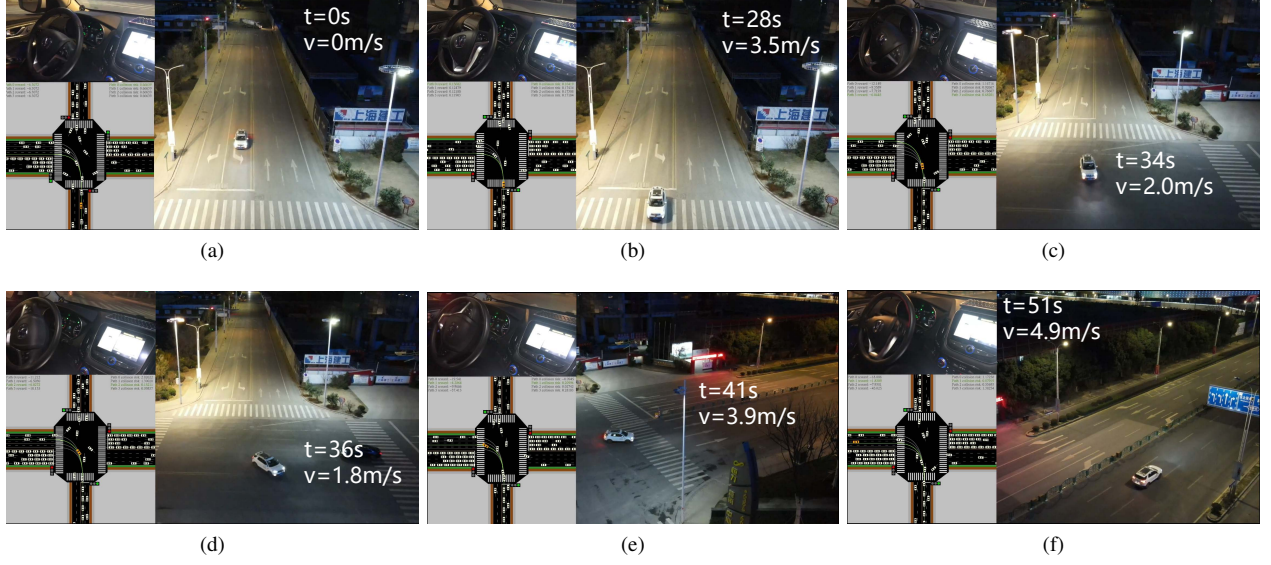


Fig. 14: Featured time steps of the left run.

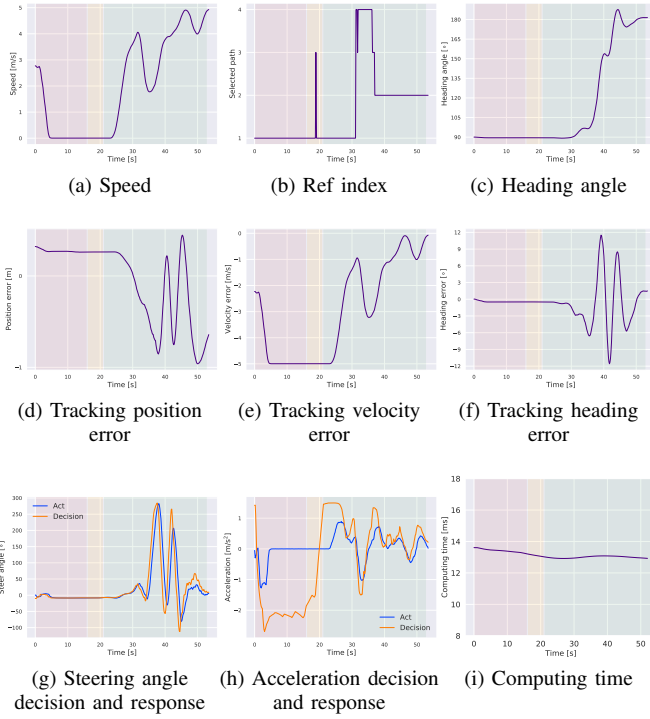


Fig. 15: Key parameters in the left run.

TABLE V: Noise level and the corresponding standard deviation.

| Noise level        | 0 | 1     | 2     | 3     | 4     | 5     | 6     |
|--------------------|---|-------|-------|-------|-------|-------|-------|
| $\delta_p$ [m]     | 0 | 0.017 | 0.033 | 0.051 | 0.068 | 0.085 | 0.102 |
| $\delta_\phi$ [°]  | 0 | 0.017 | 0.033 | 0.051 | 0.068 | 0.085 | 0.102 |
| $p_x^j, p_y^j$ [m] | 0 | 0.05  | 0.10  | 0.15  | 0.20  | 0.25  | 0.30  |
| $v_{lon}^j$ [m/s]  | 0 | 0.05  | 0.10  | 0.15  | 0.20  | 0.25  | 0.30  |
| $\phi^j$ [°]       | 0 | 1.4   | 2.8   | 4.2   | 5.6   | 7.0   | 8.4   |

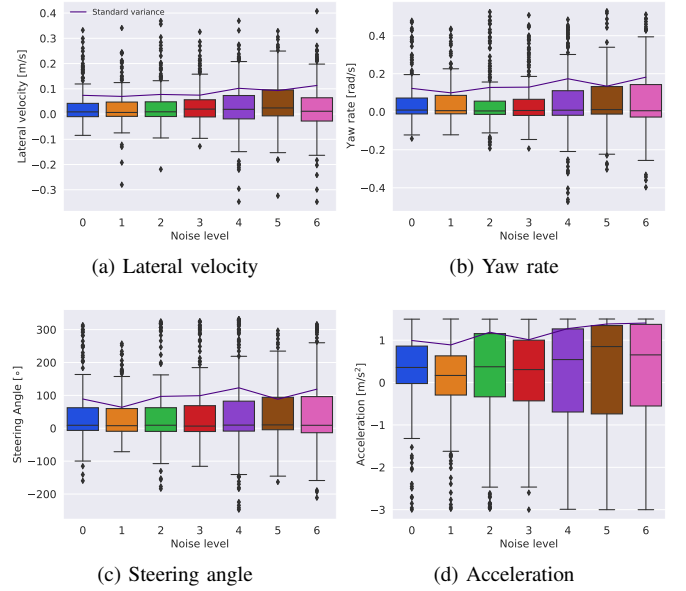


Fig. 16: States and actions in different noise level.

human intervention on the steer wheel, and then switch to the autonomous driving mode. We draw the states of the ego vehicle in Fig. 17, where the colored region is when the human disturbance is acted on. The first one is acted on 10s, when the ego just enters the crossroad and we turn the steering wheel left to  $100^\circ$  from  $0^\circ$  to make the ego head to the left. After that the driving system turn the steering wheel right immediately to correct the excessive ego heading. The second one happens at 16s, when the ego is turning to the left to pass the crossroad. We turn the steering wheel right from  $90^\circ$  to  $0^\circ$  to interrupt the process. After the take-over, the driving system is able to turn the steering wheel left to  $240^\circ$  right away to continue to complete the turn left operation. Results show

that the proposed method is capable of dealing with the abrupt human disturbance on the steering wheel by quick responses to the interrupted state after taking over.

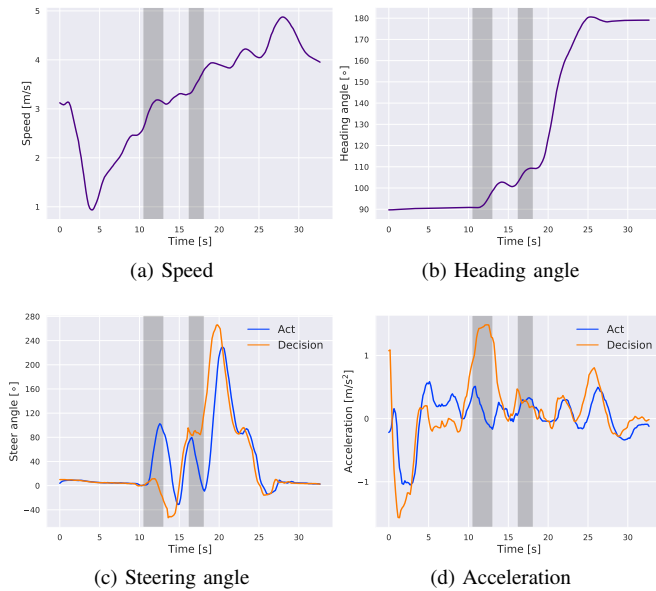


Fig. 17: States and actions under human disturbance.

## VII. CONCLUSION

In this paper, we propose integrated decision and control framework for automated vehicles, for the purpose of building an interpretable learning system with high online computing efficiency and applicability among different driving tasks and scenarios. The framework decomposes the driving task into the multi-path planning and the optimal tracking hierarchically. The former is in charge of generating multiple paths only considering static constraints, which are then sent to the latter to be selected and tracked. The latter first formulates the selecting and tracking problem as constrained OCPs mathematically to take dynamic obstacles into consideration, and then solves it offline by a model-based RL algorithm we propose to seek an approximate solution of an OCP in form of neural networks. Notably, these solved approximation functions, namely value and policy, have a natural correspondence to the selecting and tracking problems, which originates the interpretability. Finally, the value and policy functions are used online instead, releasing the heavy computation due to online optimizations. We verify our framework in both simulation and real world. Results show that our method has better online computing efficiency compared with the traditional rule-based method. In addition, it yields better driving performance in terms of traffic efficiency and safety, and shows great interpretability and adaptability among different driving tasks. The real road test suggests that it is applicable in complicated traffic scenario without even tuning.

## REFERENCES

[1] G. Li, S. E. Li, B. Cheng, and P. Green, "Estimation of driving style in naturalistic highway traffic using maneuver transition probabilities,"

*Transportation Research Part C: Emerging Technologies*, vol. 74, pp. 113–125, 2017.

[2] L. Hou, L. Xin, S. E. Li, B. Cheng, and W. Wang, "Interactive trajectory prediction of surrounding road users for autonomous driving using structural-lstm network," *IEEE Transactions on Intelligent Transportation Systems*, 2019.

[3] M. Montemerlo, J. Becker, S. Bhat, H. Dahlkamp, D. Dolgov, S. Ettinger, D. Haehnel, T. Hilden, G. Hoffmann, B. Huhne *et al.*, "Junior: The stanford entry in the urban challenge," *Journal of field Robotics*, vol. 25, no. 9, pp. 569–597, 2008.

[4] D. Dolgov, S. Thrun, M. Montemerlo, and J. Diebel, "Path planning for autonomous vehicles in unknown semi-structured environments," *The International Journal of Robotics Research*, vol. 29, no. 5, pp. 485–501, 2010.

[5] J. Duan, S. E. Li, Y. Guan, Q. Sun, and B. Cheng, "Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data," *IET Intelligent Transport Systems*, vol. 14, no. 5, pp. 297–305, 2020.

[6] Y. Guan, Y. Ren, S. E. Li, Q. Sun, L. Luo, and K. Li, "Centralized cooperation for connected and automated vehicles at intersections by proximal policy optimization," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 12 597–12 608, 2020.

[7] Q. Ge, S. E. Li, Q. Sun, and S. Zheng, "Numerically stable dynamic bi-cycle model for discrete-time control," *arXiv preprint arXiv:2011.09612*, 2020.

[8] S. E. Li, *Reinforcement learning and control. Lecture notes of Tsinghua University*, 2019.

[9] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. [Online]. Available: <https://elib.dlr.de/124092/>

[10] Y. Guan, J. Duan, S. E. Li, J. Li, J. Chen, and B. Cheng, "Mixed policy gradient," *arXiv preprint arXiv:2102.11513*, 2021.

[11] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[12] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[13] J. A. E. Andersson, J. Gillis, G. Horn, J. B. Rawlings, and M. Diehl, "CasADi – A software framework for nonlinear optimization and optimal control," *Mathematical Programming Computation*, vol. 11, no. 1, pp. 1–36, 2019.

[14] ISO Central Secretary, "Environmental management — Life cycle assessment — Principles and framework," International Organization for Standardization, Geneva, CH, Standard, 2006. [Online]. Available: <https://www.iso.org/standard/37456.html>

[15] J. Duan, "Study on distributional reinforcement learning for decision-making in autonomous driving," Ph.D. dissertation, Tsinghua University, Beijing, China, 2021.