2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Scalability of Platoon-based Coordination for Mixed Autonomy Intersections

Zhongxia Yan, Cathy Wu

***Abstract*— As transportation systems see gradual deployment of connected and automated vehicles (CAVs), there is increasing opportunity for intelligent coordination of CAVs towards system-wide objectives. While numerous previous works have modeled single junctions (*e.g.* intersections and merges) and investigated control theory-based strategies for vehicle-based coordination, this work investigates the scalability of vehicular control approaches to large networks of intersections, where interactions among multiple intersections may amplify traffic disturbances. Moreover, this work focuses on mixed autonomy networks where the highly nonlinear behavior of human-driven vehicles (HDVs) complicate overall system dynamics, and where the formation of CAV-led platoons may be advantageous. Two approaches are considered for the studied settings: model predictive control (MPC) and model-free reinforcement learning (RL), both adapted from previous methods designed for single intersection and/or full autonomy settings. Results in a network of two intersections demonstrate that MPC faces significant challenges in low-level nonlinear trajectory optimization as well as high-level crossing scheduling, while the RL policies implicitly optimizes for both low-level control and high-level coordination. Scalability analysis in large networks with hundred of intersections reveal that policies derived from additional finetuning only suffer mild degradation in performance despite the numerous out-of-distribution traffic conditions that may emerge under large scale.**

I. Introduction

In accordance with technological advances in sensing, communication, control theory, and artificial intelligence, recent adoption of autonomous driving technology raises the possibility of intelligent coordination of connected and automated vehicles (CAVs) towards system objectives, such as reducing congestion and fuel consumption. One prominent vision for CAVs invoked by numerous previous studies is the possibility of optimized vehicle coordination at *junctions*—critical regions in space where multiple streams of vehicles must interact mutually exclusively [1], [2]. Nevertheless, in the near future, partial adoption of autonomous driving technology may result in *mixed autonomy* traffic, where a significant fraction of vehicles are still human-driven, complicating the overall system dynamics with nonlinear car-following behavior. Despite only partial penetration of CAVs, previous works have demonstrated that model-free reinforcement learning (RL)-based control of CAVs may significantly reduce congestion in single-intersection mixed autonomy systems [3]. However, even for *full autonomy* systems, previous studies in kinematic vehicle coordination often restrict their analysis to single intersections [1], [2] or small grids of intersections [4]–[6], disregarding traffic disturbances which may arise due to compounding and feedback in large traffic systems.

In the context of these works, this article designs and analyzes extensions of model predictive control (MPC)-based and model-free RL-based approaches for mixed autonomy coordination in traffic networks with multiple intersections. To focus on scaling, we restrict our study to two-way and four-way intersections geometries without turns, as respectively studied by [2] and [1]; coordinating such intersections requiring decision-making on both the acceleration-level and the crossing order-level. For the MPC approach, inspired by analogous approaches in full autonomy [2], the trajectories of each CAV and its immediate human followers (grouped into a *platoon*) are jointly obtained through nonlinear trajectory optimization in first-come-first-serve (FCFS) order of intersection crossing. For the model-free RL approach, we leverage the factorized multi-agent RL approach designed by [3] and apply a heuristic sequence of transfer finetuning to medium-sized intersection networks and zero-shot transfer to large intersection networks. We find that the trained RL policy performs better or equal to the MPC approach under all conditions in a small intersection network. Though the planning time of MPC becomes impractical for larger intersection networks, we analyze the trained RL policy under networks of hundreds of intersections and thousands of vehicles. Interestingly, we find little or no degradation in throughput performance for full autonomy coordination, though performance mildly worsens with decreasing penetration of CAVs.

Our code, models, and additional videos of results can be found on GitHub.

II. Related Work

A. Coordination of CAVs at Signal-free Intersections

In a seminal work, [7] introduces the concept of autonomous intersection management by coordinating CAVs following a first-come-first-serve-based reservation protocol, which discretizes the intersection into tiles. Many subsequent works model a single intersection in continuous space and tackle a bilevel optimization problem [8], with high-level discrete decisions specifying the order of CAV crossings at the intersection and the low-level continuous control of vehicle accelerations for fulfilling the crossing order [1], [2], [9], [10]. For example, [2] models a two-way intersection as a polling system with two queues and compares the system behavior under two heuristic crossing orders (exhaustive

This work was supported by MIT SuperCloud, the MIT Amazon Science Hub, MIT-IBM Watson AI Lab, and the National Science Foundation (NSF) under grant number 2149548.

Zhongxia Yan and Cathy Wu are with the Laboratory for Information & Decision Systems (LIDS), Massachusetts Institute of Technology, Cambridge, MA 02139, USA. Email:{zxyan,cathywu}@mit.edu

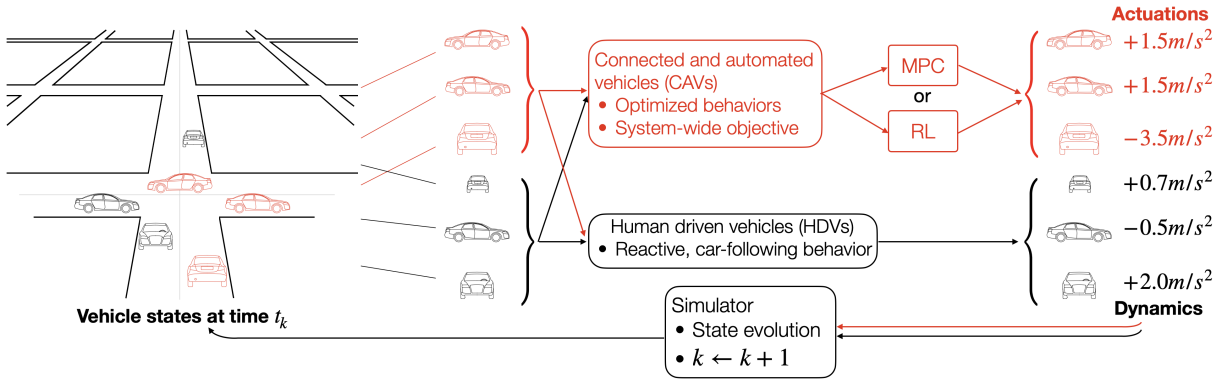


Fig. 1: **An overview of mixed autonomy coordination.** At each time t_k, HDVs follow highly nonlinear human-driving behaviors while MPC or RL dictate the actuations of CAVs given the traffic states of surrounding vehicles. A microscopic simulator obtains evolved vehicle states for the next timestep.

and k-limited polling policies), obtaining vehicle trajectories with trajectory optimization conditioned on the crossing order. Inheriting this design, our designed MPC approach adapts these concepts to *mixed autonomy* within an extended *network* of multiple such two-way intersections.

Relatively sparse prior work has focused on CAV coordination in multiple intersections, perhaps due to the difficulty of scaling planning to large multi-agent systems. [4] designs an A* based algorithm for a four-intersection network which dynamically routes CAVs based on estimated travel time. [5] designs a bilevel optimization algorithm to coordinate CAVs in a network of three two-lane intersections without turns while considering travel time and fuel. [6] demonstrates an extensive composite technique for route and trajectory optimization in a grid network of six intersections with turns. [11] devises a large-scale spatial routing algorithm in sync with a global "rhythm", but assumes that all vehicles travel at constant speed. Our work focuses on analyzing the scalability of coordination algorithms for crossing order and motion-level decisions while controlling for route-level decisions.

B. Mixed Autonomy Coordination of Traffic

Mixed autonomy, in contrast with *full autonomy*, denotes the coexistence of coordinated CAVs and uncoordinated HDVs in a traffic system. Restricted by the nonlinearity of HDV dynamics, model-based methods for mixed autonomy control typically focus a single platoon consisting of a leading CAV and several following HDV [12], [13]. [14] studies a first-come-first-serve scheme for scheduling CAV passages through the intersection while regulating human vehicles with fixed-time traffic signal. On the other hand, model-free RL decouples the modeling of vehicle driving dynamics from the optimization of CAV decisions and has been leverage by the Flow framework [15] to automatically discover vehicular coordination strategies in various mixed autonomy settings [3] ranging from highway merges to four-way intersections. While these works still focus on coordination of single intersections, this paper aims to investigate mixed autonomy coordination in large-scale intersection networks with orders of magnitude more CAVs.

C. Coordination of Signalized Intersections

In intersections systems with only HDVs, fixed-time control of traffic signals may be programmed to address historical demand patterns throughout the day [16], while adaptive control of signals may address current traffic conditions [17]. Recently, RL-based signal control methods have demonstrated superior performance to traditional methods in simulated large-scale traffic networks with thousands of intersections [18]. In this work, we design *ideal* traffic signals as a comparison for CAV-based coordination.

III. Problem Statement

The geometry of our problem formulation is a multi-intersection extension of those in [1], [2]. Consider a M by N network of MN unsignalized intersections on a Cartesian plane. The center of intersection (m, n) where $m \in \mathbb{Z}_{1:M}$ and $n \in \mathbb{Z}_{1:N}$ is at position (mL, nL), where L is the distance between adjacent intersections. The (m, n)th intersection region corresponds to $[mL - w, mL + w] \times [nL - w, nL + w] \subset \mathbb{R}^2$, where we assume that the intersection width is twice the lane width w for notational clarity. Each right-moving lanes is defined as the rectangle $[mL, mL + L] \times [nL - w, nL]$ for $m \in \mathbb{Z}_{0:M}$ and $n \in \mathbb{Z}_{1:N}$; each up-moving lanes is defined as the rectangle $[mL, mL+w] \times [nL, nL+L]$ for $m \in \mathbb{Z}_{1:M}$ and $n \in \mathbb{Z}_{0:N}$. For four-way intersections, the left- and down-moving lanes are defined analogously. Each horizontal route is an directed chain of $N + 1$ lanes, and each vertical route is an directed chain of $M + 1$ lanes, and no turns are permitted.

We define the discrete time $k \equiv t_k = k\Delta t$ for $k \in \mathbb{Z}_{0:H}$ where Δt is the time step and H is the problem horizon. We define vehicle ν's state in the reference frame of its route ρ: $x_k^\nu = \begin{bmatrix} p_k^\nu \\ v_k^\nu \end{bmatrix}$ where speed $v_k^\nu \in [0, v_{\max}]$ while front bumper position $p_k^\nu \in [0, \ell^\rho]$ increases along ρ and is upper-bounded by the length of ρ. The vehicle's state evolves according to second order dynamics

$$x_{k+1}^\nu = \begin{bmatrix} p_{k+1}^\nu \\ v_{k+1}^\nu \end{bmatrix} = \begin{bmatrix} \min(\ell^\rho,\ p_k^\nu + v_{k+1}^\nu \Delta t) \\ \max(0, \min(v_{\max},\ v_k^\nu + \dot{v}_k^\nu \Delta t)) \end{bmatrix}. \quad (1)$$

For each CAV, the acceleration is directly controlled; for each HDV, the acceleration is modeled by the Intelligent Driver Model [19], a reactive car-following model which is a function of the states of vehicle ν itself and its leader ν':

$$\dot{v}_k^\nu = \begin{cases} u_k^\nu \in [u_{\min}, u_{\max}] & \text{if } \nu \in \mathcal{A} \\ f(x_k^\nu, x_k^{\nu'}) & \text{otherwise} \end{cases}, \tag{2}$$

where, omitting ν and k for clarity,

$$f(x, x') = c_a \left(1 - \left(\frac{v}{v_{\text{des}}} \right)^{c_\delta} - \left(\frac{d_{\min} + v c_\tau + \frac{v(v-v')}{2\sqrt{c_a c_b}}}{p' - p - \ell^\nu} \right)^2 \right). \tag{3}$$

The penetration rate $\phi \in [0, 1]$ is the fraction of CAVs among all vehicle inflows. However, while previous mixed autonomy intersection work considers CAVs to inflow at *regular* intervals [3], *i.e.* every $\frac{1}{\phi}$th vehicle on each route is CAV, we also consider *irregular* CAV penetration, where each vehicle has independent probability ϕ of being CAV.

A rear-end collision occurs along a route ρ if $p_k^\nu \le p_k^{\nu'}$ for any two vehicles ν and ν' along ρ. In the reference frame of ρ, let the position of an intersection χ along ρ be p^χ. The vehicle ν is *within* intersection χ during time k if and only if $p_{k-1}^\nu \le p^\chi + w + \ell$ and $p_k^\nu \ge p^\chi - w$. A crossing collision occurs if any two orthogonally-heading vehicles are within the intersection at any time.

Vehicles start trips at position 0 of each route ρ at an inflow rate of F^ρ vehicles per hour upon availability of space, with initial speed v_{init}, and leave the traffic network when their positions exceed ℓ^ρ. The *throughput* is the number of vehicles that leave the network within some unit of time.

The overall optimization problem is to maximize the throughput across a time horizon H with respect to the control inputs u for all CAVs, subject to kinematics constraints (Equation 1), acceleration behavior (Equation 2 and 3), rear-end collision avoidance, and crossing collision avoidance.

TABLE I: **Summary of symbols and notations.**

Variables	Descriptions
t; $k \equiv t_k$	Time; discrete time
ν; $\alpha \in A$; ξ	Vehicle; CAV (agent); platoon
ρ; χ	Route; intersection
F^-; $F^\vert$; F^ρ	Inflow rate on horizontal routes; vertical routes; ρ
ϕ	Penetration rate (fraction of CAVs)
x; p; v; $\dot{v}$	Vehicle state; position; speed; acceleration
Constants	
$\Delta t = 0.5$s	Timestep size
$H = 2000$	Problem horizon
$h_p = 20$; $h_c = 1$	MPC predictive horizon; MPC control horizons
$L = 100$m	Inter-intersection distance
$w = 5.6$m	Half-intersection width; lane width
$\ell^\nu = 5$m	Vehicle length
$v_{\text{init}} = 0$m/s	Initial speed
$v_{\max} = 13$m/s	Maximum speed
$u_{\max} = 1.5$m/s^2	Maximum acceleration control
$u_{\min} = -3.5$m/s^2	Maximum deceleration control
$d_{\min} = 2.5$m	(IDM) minimum space gap
$c_a = 2.6$m/s^2	IDM maximum acceleration
$c_b = 4.5$m/s^2	IDM comfortable deceleration
$c_\tau = 1$s	IDM desired time headway
$v_{\text{des}} = 13$m/s	IDM desired speed
$c_\delta = 4$	IDM exponent
Vector Notations	
$x_{k_1:k_2} = x$	$x_{k_1} = x_{k_1+1} = \cdots = x_{k_2} = x$
$\mathbb{Z}_{k_1:k_2}$	$\{k \mid k \in \mathbb{Z},\ k_1 \le k \le k_2\}$

IV. Model Predictive Control (MPC)

We design a MPC approach considering the structure of the present problem. The problem of interest is intractable to solve directly due to integral components in the objective and constraints, nonlinearity of car-following behavior in the constraints, the large total number of vehicles involved, and the temporally extended nature (a long horizon H) of the throughput objective. Therefore, similar to [2] but for mixed autonomy, we design a decomposition-based approach, Algorithm 1, which computes high-level discrete decisions heuristically while optimizing low-level acceleration decisions with nonlinear programming-based trajectory optimization.

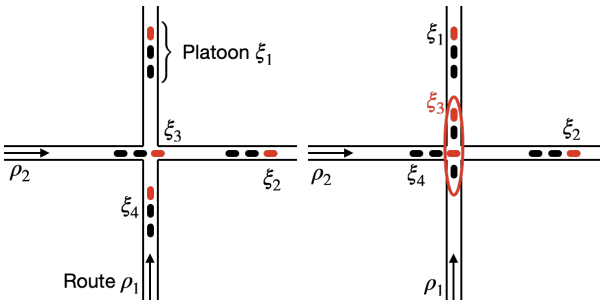
A. Speed-based Objective

As the throughput objective is a function of number of leaving vehicles, and thus inherently discrete and temporally extended, we consider a proxy objective over a more manageable prediction horizon $h_p \ll H$ to be the sum over all vehicle speeds. This objective provides immediate feedback for vehicle decisions and helps distinguish decision quality for small prediction horizon h_p, while largely aligning with the original throughput objective. We denote the MPC control horizon by $h_c \le h_p$.

B. Multi-level Control of Platoons

Only CAVs are coordinated in mixed autonomy traffic, and HDVs can only be controlled indirectly by controlling the CAV leading them. If the leading CAV is too distant to effect HDV followers, the ability to control HDVs is lost. Therefore, we associate each CAV with any HDVs that immediate follows it into a *platoon* ξ whose motion should be jointly considered during optimization; every vehicle is part of exactly one platoon.

We would like to enforce the *uninterrupted crossing* of platoons at intersections: any platoon must finish crossing the intersection completely before another platoon begins to cross. Figure 2 demonstrates an uninterrupted crossing and an interrupted crossing. This constraint ensures that CAVs always retain control over HDVs in their platoons. In practice, interrupting a platoon's crossing often results in significant slowdowns to the interrupted vehicles, worse performance overall, and difficulty in enforcing safety anyways, so we believe that any suboptimality introduced by this constraint should be small. Any solution satisfying the uninterrupted crossing constraint is then associated with a set of intersection crossing orders, where the sequence of platoons crossing each intersection χ can be summarized as a *crossing order* Q_χ. The optimal solution satisfying the uninterrupted crossing constraint is therefore the optimal set of crossing orders (high-level discrete decisions) for all intersections coupled with the optimal trajectories of



(a) Uninterrupted (b) Interrupted

Fig. 2: **Interrupted vs uninterrupted crossings.**

platoon vehicles (low-level continuous control) conditioned on the crossing orders. Therefore, we devise Algorithm 1 to condition the low-level trajectory optimization of platoon vehicles on the high-level crossing orders. This separation of high-level crossing decisions and low-level continuous control is typical of model-based approaches for *full autonomy* intersection management [8].

C. High-level Platoon Crossing Orders

We decide the high-level crossing orders with a reservation-based algorithm employing a first-come-first-serve (FCFS) heuristic, which has also been used by other intersection coordination algorithms [7], [14]. Each intersection χ keeps track of a FCFS crossing order Q_χ of platoons. A platoon is appended to Q_χ when it first enters the lane immediately approaching χ. According to prior studies on crossing order in full autonomy settings, first-come-first-serve is often a reasonable heuristic crossing order [9]. The FIFO crossing order for platoons is also similar to the k-limited polling policy studied by [2], for "k" equal to $\frac{1}{\phi}$.

D. Topological Factoring of Low-level Control

Even given the crossing orders at all intersections, joint trajectory optimization of all vehicles is still impractical due to the nonlinear dynamics of the car-following models of the HDVs. Thus, we approximately optimize the joint trajectories of all vehicles while by optimizing vehicle trajectories one platoon at a time. Intuitively, optimizing the trajectory of a platoon ξ approaching intersection χ requires boundary conditions provided by the trajectories of at most two other platoons: 1) the platoon ξ_{lead} directly leading ξ along the same route, and 2) the platoon ξ_{cross} crossing χ immediately before ξ. We can therefore construct a directed dependency graph of platoons based on route-precedence of platoons and the crossing orders Q at all intersections, and we cannot optimize the trajectory of a platoon until we obtain the trajectories of its dependencies. A two-way acyclic traffic network naturally yields an acyclic dependency graph, so we may carry out platoon-level trajectory optimization in the topological ordering of this dependency graph. While the optimization order may be addressed even for cyclic dependency graphs, this is orthogonal to the focus of our work, scalability of coordination.

E. Platoon-level Trajectory Optimization

Let ξ represent a list of vehicles with ξ_1 denoting the leading CAV and $\xi_{|\xi|}$ denoting the last HDV. Setting current time to 0 for notational clarity, we denote the platoon-level trajectory optimization problem for ξ by $\mathcal{P}(x_0^\xi, \overline{p}_{1:h_p})$, where x_0^ξ is the initial state of ξ and $\overline{p}_{1:h_p}$ is the upper-bounds on the positions of ξ_1 from time 1 to h_p to avoid collision with ξ_{lead} and ξ_{cross}. The nonlinear optimization problem $\mathcal{P}$ is thus:

$$\underset{p^\xi_{1:h_p}, v^\xi_{1:h_p}}{\operatorname{argmax}} \sum_{\nu\in\xi}\sum_{k=1}^{h_p} v_k^\nu \tag{4}$$

subject to, $\forall j \in \mathbb{Z}_{2:|\xi|}$,

$$\begin{aligned}
&p^{\xi_1}_{1:h_p} \le \overline{p}_{1:h_p} \qquad p^{\xi_j}_{1:h_p} - p^{\xi_{j-1}}_{1:h_p} \ge \ell^\nu \qquad 0 \le v^\xi_{1:h_p} \le v_{\max}\\
&\frac{p^\xi_{1:h_p} - p^\xi_{0:h_p-1}}{\Delta t} = v^\xi_{1:h_p} \qquad u_{\min} \le \frac{v^{\xi_1}_{1:h_p} - v^{\xi_1}_{0:h_p-1}}{\Delta t} \le u_{\max}\\
&\frac{v^{\xi_j}_{1:h_p} - v^{\xi_j}_{0:h_p-1}}{\Delta t} = f\left(x^{\xi_j}_{0:h_p-1}, x^{\xi_{j-1}}_{0:h_p-1}\right),
\end{aligned} \tag{5}$$

which respectively enforce CAV rear-end and crossing safety, HDV rear-end safety, speed range, position update, CAV acceleration range, and HDV car-following behavior.

Algorithm 1 MPC for Two-way Junction Network

procedure SIMULATE
  $Q \leftarrow$ empty crossing orders for every χ
  for $k = 1$ to $k = H$ **do**
    Add default plans for all new inflow CAVs
    Append new platoons approaching χ to Q_χ
    if $k \equiv 0 \mod h_p$ **then**
      $x_k \leftarrow$ all vehicle states
      New CAV plans $\leftarrow$ PLAN(x_k, Q)
    Evolve states by CAV plans and HDV dynamics

procedure PLAN(x_0, Q)
  ▷ *Topological order: Section IV-D*
  for each $\xi \in$ topological platoon order **do**
    $\chi \leftarrow$ intersection approached by ξ
    Let all position p be defined along ξ's route
    $\overline{p}_{1:h_p} \leftarrow \infty$
    if leading platoon ξ_{lead} exists **then**
      $\nu \leftarrow$ last vehicle in ξ_{lead}
      $\overline{p}_{1:h_p} \leftarrow p^\nu_{1:h_p} - \ell - d_{\min}$
    if previous crossing platoon ξ_χ exists **then**
      $k_\chi \leftarrow$ last timestep that ξ_χ occupies χ
      $p^\chi \leftarrow$ mid-position of χ along ξ's route
      $\overline{p}_{1:k_\chi} \leftarrow \min\{\overline{p}_{1:k_\chi}, p^\chi - w\}$
    ▷ *Platoon-level problem: Section IV-E*
    $x^\xi_{1:h_p}$, $u^\xi_{0:h_p-1} \leftarrow$ solve $\mathcal{P}(x_0^\xi, \overline{p}_{1:h_p})$
  return $u_{0:h_c-1}$

TABLE II: **Summary of MDP and RL symbols.**

Symbols	Descriptions
$s_k \equiv x_k \in \mathcal{S}$; $a_k \equiv u_k \in \mathcal{A}$	System state; joint action
$r(s_k, a_k, s_{k+1})$	System reward
$T(s_k, a_k, s_{k+1})$	Transition function
$\gamma = 0.99$; $H_\gamma = \frac{1}{1-\gamma}$	Discount factor; effective horizon
π_θ	Policy parameterized by θ
z	Observation function
π_θ^α; $o_k^\alpha = z(s_k, \alpha) \in \mathcal{O}^\alpha$; $a_k^\alpha \in \mathcal{A}^\alpha$	Agent policy; observation; action
$\lambda_c = 5$	Collision penalty coefficient

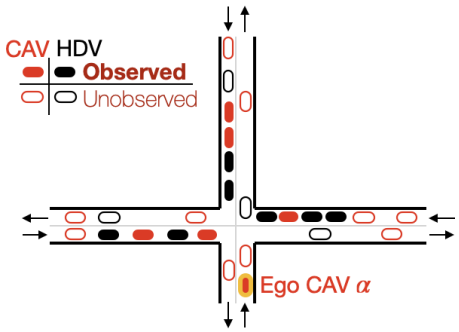


Fig. 3: **Observation space.**

V. Model-free Reinforcement Learning

Previous have found that RL discovers platooning-based strategies automatically [3] for mixed autonomy intersections, so we aim to analyze the effectiveness of such strategies under large scale and varying conditions. We formulate the coordination problem as a finite horizon Markov Decision Process (MDP) and extend the model-free RL method proposed by [3] from single intersections to up to hundreds of intersections by incorporating transfer finetuning to progressively larger traffic networks and zero-shot transfer to very large traffic networks. We summarize the MDP- and RL-related symbols in Table II, but notably we use $s_k \equiv x_k$ and $a_k \equiv u_k$ to be consistent with RL notation. The overall approach follows the centralized training with decentralized execution (CTDE) multi-agent paradigm, which requires global throughput rewards during training but allows each CAV to only acquire local information during execution.

A. Multi-agent Policy Decomposition

Decision problems with a large number of action dimensions may be challenging for learning algorithms. However, if the decision dimensions are somewhat decoupled, multi-agent policy decomposition may be suitable [3]. MDPs with multiple partially decoupled control dimensions can naturally be formulated as a decentralized partially observable MDP (Dec-POMDP). Let A_k denote the set of CAVs (*agents*) at time k. The policy factorizes into the product of per-agent policies: $\pi_\theta(a_k|s_k) = \prod_{\alpha \in A_k} \pi_\theta^\alpha(a^\alpha|o^\alpha)$. A state $s_k \in \mathcal{S}$ of the original MDP may be restricted by the observation function z to obtain per-agent observations $o_k^\alpha = z(s_k, \alpha) \in \mathcal{O}^\alpha$ for each agent α. Each agent's observation contains only a subset of the information in s_k, thus $\bigcup_{\alpha \in A} o_k^\alpha \subseteq s_k$. The joint action space of the original MDP factorize into the product of agent action spaces $a_k^\alpha \in \mathcal{A}^\alpha$, satisfying $\bigtimes_{\alpha \in A} \mathcal{A}^\alpha = \mathcal{A}$.

The observation function should extract relevant *ego* aspects of the state with respect a given CAV approaching some intersection χ. As illustrated in Figure **??**, we allow each CAV to observe itself and the first 4 vehicles along each other approach to intersection χ. Each observed vehicle is featurized by its distance to χ, its speed, and whether or not it is a CAV. The per-agent action space is a discrete bang-off-bang acceleration space $\mathcal{A}^\alpha = \{u_{\min}, 0, u_{\max}\}$ which empirically improves coordination between multiple CAVs compared to a continuous acceleration space [3]. The policy is a fully connected network mapping the input observation to action probabilties.

B. Centralized Objective and Reward

Since the ultimate objective of interest is the system throughput (veh/hour), the main reward function $r(s_k, a_k)$ is the number of vehicles leaving the traffic network at step k. Thus, the cumulative reward is equal to the throughput. Since the simulator cannot fully guarantee safety of vehicles, we additionally penalize for the number of collided vehicles at step k, with coefficient λ_c. Since the overall global state s_k may be large and difficult to encode with a learned value function, we apply reward normalization for variance reduction [20] rather than training a value-based critic.

C. Multi-task and Transfer Learning

In general, it would be cumbersome to train a separate RL policy for every unique system. Therefore, we incorporate multi-task and transfer learning to share learning experience across multiple different systems.

With multi-task learning, we simultaneously initialize multiple systems corresponding to different inflow rates. We use the same policy to collect trajectories across all systems. The collected trajectories are batched to perform each gradient update of the policy. When some coordination strategies may be shared across different inflow rates, multi-task learning aims to obtain a single policy with good performance while avoiding the computational cost of training a separate policy for each system.

With limited computational resources, large traffic settings may involve too many vehicles to directly optimize with RL. Therefore, we consider two types of transfer learning techniques: *zero-shot transfer* and *transfer with finetuning*; the former directly evaluates a trained policy from a source setting to a target setting without additional training, while the latter continues to finetune the policy on the target setting after transfer. Transfer learning not only may reduce the computational burden of training on many different settings, but also enables us to probe the performance of learned RL policies on traffic settings that are too large to train on. In this paper, we analyze the effectiveness of zero-shot transfer and also characterize any practical advantages of transfer with finetuning over training from scratch in terms of the three

transfer objectives: learning speed improvement, asymptotic improvement, and jumpstart improvement [21].

VI. Experimental Setup

We perform all simulations within the SUMO microscopic simulator [22], which closely models the description in Section III but also adds collisions-preventing mechanisms. In practice, if two or more slow-moving vehicles (*e.g.* HDVs) approach the intersection from different lanes, the behavior resembles that of a stop sign, where vehicles slow down to pass alternatively. We obtain a randomized initialization by warm-start steps during which all vehicles behave like HDVs. Subsequently, additional warm-start steps are executed to allow traffic dynamics to reach steady state under the evaluated coordination method. Thereafter, the performance is measured on a final $H = 2000$ simulation steps.

We perform all nonlinear trajectory optimization with Drake's [23] Sparse Nonlinear OPTimizer (SNOPT) [24]. We train and finetune RL policies by performing update steps with the Trust Region Policy Gradient (TRPO) [25] algorithm until policy performance converges. The per-agent policy network has three fully-connected layers with hidden size of 64. Policy parameters θ are shared across all CAVs to leverage experience sharing. Finetuning initializes θ from a previously trained checkpoint. We perform each gradient step on a batch of 480 collected trajectories. Training/finetuning time is proportional to traffic network size, and finetuning on a 3x3 traffic network takes roughly one day on a CPU machine with 48 Intel Xeon Platinum CPU cores. We do not observe significant differences in performance between training runs, despite the stochasticity of training. To discourage the policy from overfitting to a particular vehicle density condition, we always multi-task over combinations of horizontal per-route inflow rate F^{-} and vertical per-route inflow rate $F^{|}$ such that $1400 \leq F^{-} + F^{|} \leq 1850$. If the total inflow rate reaching an intersection is too high, then congestion is unmitigable; if the total inflow rate reaching an intersection is too low, then the problem is trivial.

For trained policies, we evaluate the checkpoint with the best mean performance during training. When evaluating, we take the mean performance across 10 random seeds.

To ground our evaluation of the MPC and RL approaches, we construct a strong and *ideally*-tuned traffic signal baseline with 0% CAV penetration. For each particular combination of horizontal and vertical inflow rates, we empirically sweep over a range of traffic signal phase lengths along both directions in a 1x1 traffic network and pick the combination of phase lengths which empirically maximizes the throughput, and deploy these phase lengths for all conditions with the same inflow rate conditions. We believe the ideal traffic signals serves as a reference for near-optimal throughput, though it is not a direct comparison as the mode of actuation differs (signal vs CAVs) and privileged knowledge of the system inflow rates along each approach is required.

VII. Experimental Results

A. Performance on Small- and Medium-scaled Tasks

In Table III, we compare performance of MPC and RL in the two-way 2x1 traffic network, two-way 3x3 traffic network, and four-way 1x1 with regular CAV penetration considered by [3]. RL exceed the Ideal TL performance in all but the two-way 2x1 setting with 10% penetration rate, while MPC is comparable to Ideal TL under 50% penetration. The comparisons suggest that both RL perform closer to the optimal throughput than MPC under a larger range of settings. We find that lower penetration rate (*i.e.* 33%) increases the difficulty for MPC since the number of HDVs per platoon is increased, compounding the nonlinearity of car-following model constraints. Interestingly, MPC also performs worse at high penetration rates (*i.e.* 100%), likely due to the deficiencies of our FCFS-based crossing order heuristic, which suboptimally leads to a platoon of one CAV crossing the intersection at a time under 100% penetration. Finally, we note that the inference time of RL is negligible (2-3 orders of magnitude less than the MPC planning time) and much less than the simulator's state evolution time; thus RL scales much more easily to the two-way 3x3 traffic network. Due to inferior performance, scaling limitations, and inapplicability to four-way traffic networks, we do not further analyze the performance of MPC in other settings.

B. Time-space Diagrams of Vehicle Trajectories

To compare the behaviors of Ideal TL, MPC, and RL, we illustrate representative vehicle trajectories for each policy within the two-way 2x1 setting with inflow rates $(F^{-}, F^{|}) = (1000, 850)$ and 50% regular CAV penetration in Figure 4. The time-space diagram illustrates the trajectories of vehicles approaching the *upstream* intersection. MPC and RL demonstrate somewhat similar behavior in alternating *uninterrupted* platoons of two vehicles (though RL is slightly smoother), and the Ideal TL alternatingly allows 8 vehicles to pass at a time. Though our goal is not to compare mixed autonomy control with traffic signal control, Figure 4 demonstrates that even under *mixed* autonomy, each individual vehicle's delay at a given intersection may be much lower than that of traffic signal control, which resembles conclusions reported by full autonomy works [2].

C. Effect of Transfer Finetuning

As a first step towards studying the scalability of RL policies, we examine the benefits of finetuning a transferred policy from a different source setting when compared to training in the target setting directly. In Figure 5, we compare the performance progress of training a policy from scratch for on a two-way 3x3 setting versus finetuning policies pretrained on two-way 1x1, 2x1, and 2x2 source settings under 33% regular CV penetration. For each training or finetuning setting, we take the mean and standard deviation of the inflow-averaged throughput over three independent trials. We see both asymptotic and jumpstart improvements [21] by the finetuned policies over the policy trained from scratch. The jumpstart improvement (diminishingly) increases as the

TABLE III: **Throughput performance of MPC and RL relative to Ideal TL for various ϕ on three intersection grids.**

F^-	$F^\|$	Two-way 2x1								Two-way 3x3			Four-way 1x1		
(veh/hr/route)		Id. TL (veh/hr)	$\phi = 1$		$\phi = 1/2$		$\phi = 1/3$		$1/10$	Id. TL (veh/hr)	$\phi = 1/2$	$\phi = 1/3$	Id. TL (veh/hr)	$\phi = 1/2$	$\phi = 1/3$
			$\frac{\text{MPC}}{\text{TL}}$	$\frac{\text{RL}}{\text{TL}}$	$\frac{\text{MPC}}{\text{TL}}$	$\frac{\text{RL}}{\text{TL}}$	$\frac{\text{MPC}}{\text{TL}}$	$\frac{\text{RL}}{\text{TL}}$	$\frac{\text{RL}}{\text{TL}}$		$\frac{\text{RL}}{\text{TL}}$	$\frac{\text{RL}}{\text{TL}}$		$\frac{\text{RL}}{\text{TL}}$	$\frac{\text{RL}}{\text{TL}}$
1000	850	2662	0.77	**0.99**	0.95	**1.02**	0.97	**1.00**	0.94	5006	1.04	1.01	3318	1.02	1.03
850	1000	2526	0.80	**1.02**	1.01	**1.04**	0.92	**1.03**	1.00	5006	1.04	1.01	3318	1.02	1.03
850	850	2525	0.79	**1.01**	**1.01**	**1.01**	1.00	**1.01**	1.00	5013	1.02	1.02	3313	1.02	1.03
850	550	2255	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.99	4205	1.00	1.00	2800	1.00	1.00
700	700	2106	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	1.00	4199	1.00	1.00	2800	1.00	1.00
400	1000	1809	**0.99**	**0.99**	0.99	**1.00**	0.97	**1.00**	0.99	4202	1.00	1.00	2800	1.00	1.00

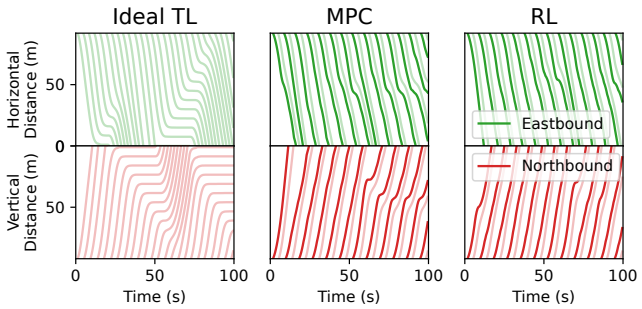


Fig. 4: **Time-space diagrams of vehicle trajectories under the TL Oracle, RL, MPC policies for two-way tasks with 50% regular penetration and size $g^+ = (2, 1)$.**

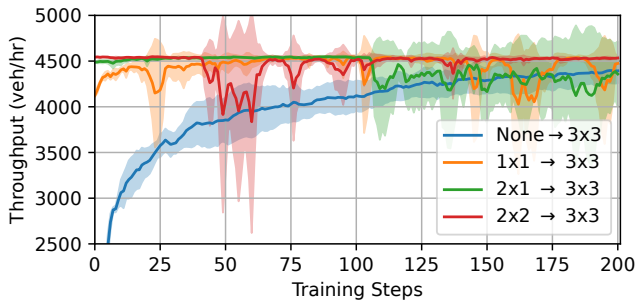


Fig. 5: **Transfer finetuning of RL policies on the two-way 3x3 traffic network**

source setting approaches the size of the target setting, whereas the asymptotic improvement is the same regardless of the source setting. These results suggest that benefits exist even when transferring policies trained on much smaller settings. Besides these benefits, we experimentally find that finetuning a transferred policy provides an additional advantage over training from scratch on larger settings: a randomly initialized policy may lead to extreme and unrepresentative congestion conditions.

D. Transfer Performance on Large Tasks

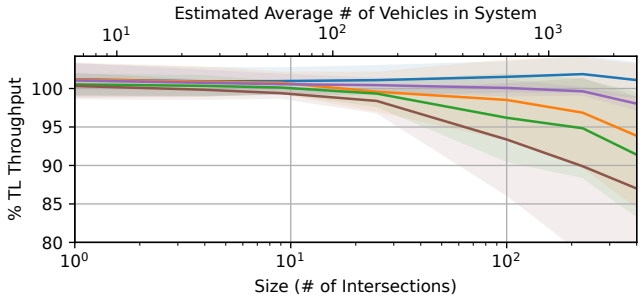
We investigate the zero-shot transfer performance of the RL policy to very large traffic networks, which are difficult to train or finetune directly due to GPU memory constraints. Thus, we would like to explore the zero-shot transfer performance of policies trained on smaller source settings to large target settings, up to two-way 20x20 settings and four-way 10x10 settings. Each policy here is trained or finetuned on the specified penetration rate and irregularity with multi-task learning across all inflow rates. The size of the source setting is progressively increased until finetuning no longer improves performance on the source tasks, which typically occurs at around 2x2 or 3x3 due to our computational constraints limiting training batch sizes used. Each policy is then zero-shot transferred to large scale settings with the *same* penetration rate and regularity that the policy saw at training time.

One potential challenge of zero-shot transfer is that suboptimality of decisions may compound and amplify in larger settings. In Figure 6, we experimentally find that the performances of our policy degrade relative to the Ideal TL under various penetration rate and penetration regularities. The exception is the two-way 100% penetration tasks corresponding to *full* autonomy systems, which continues to outperform the Ideal TL even in systems with 400 intersections. In general, for both regular and irregular CAV penetration, the transfer performances to large target settings under lower penetration rates tend to decline more. Moreover, regular CAV penetration tends to outperform irregular CAV penetration when holding the penetration rate constant. These experiments suggest that suboptimal behaviors due to irregularity-induced underactuation may cascade and amplify more in larger tasks with more complex (*e.g.* four-way) intersections. On average, we observe roughly 1400 vehicles in the two-way 20x20 network and roughly 1900 vehicles in the four-way 10x10 traffic network at any given time.

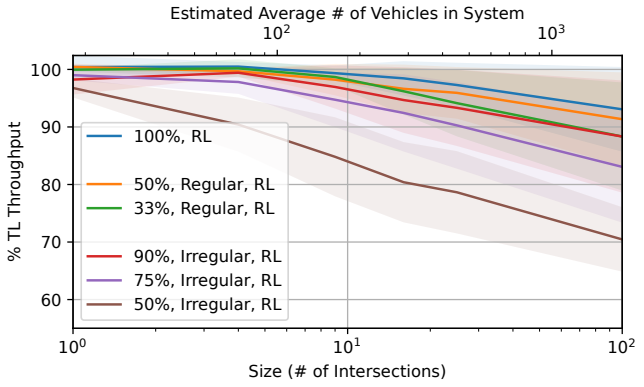
VIII. Conclusion

This article contributes an investigation into model-based and model-free coordination strategies for mixed autonomy intersections, including very large settings with up to hundreds of intersections. The model-based MPC approach leverages and adapts strategies employed by previous works [2] while the model-free RL approach infuses transfer learning to scale previous approaches [3] beyond a single junction. We conclude that model-free RL is more effective and scalable than model-based MPC for the large, underactuated, and highly nonlinear settings of mixed autonomy coordination.

Insights from our work may assist in policy- and decision-making regarding the adoption of autonomous vehicles in multi-faceted intelligent transportation systems. The richness of intelligent traffic systems leaves much future work to be desired. For instance, future work may analyze systems with



(a) Two-way scaling



(b) Four-way scaling

Fig. 6: **Zero-shot transfer and scaling of RL policies to large two-way and four-way traffic networks under various penetration rates and irregularities, relative to Ideal TL.**

richer geometry (*e.g.* turns), additional control elements such as traffic signals in conjunction with CAVs, heterogeneous vehicles, mixtures of possibly conflicting objectives (*e.g.* throughput vs fuel), and non-stationary traffic regimes (*e.g.* with temporally varying inflow rates).

REFERENCES

[1] A. A. Malikopoulos, C. G. Cassandras, and Y. J. Zhang, "A decentralized energy-optimal control framework for connected automated vehicles at signal-free intersections," *Automatica*, 2018.

[2] D. Miculescu and S. Karaman, "Polling-systems-based autonomous vehicle coordination in traffic intersections with no traffic signals," *IEEE Transactions on Automatic Control*, 2019.

[3] Z. Yan, A. R. Kreidieh, E. Vinitsky, A. M. Bayen, and C. Wu, "Unified automatic control of vehicular systems with reinforcement learning," *IEEE Transactions on Automation Science and Engineering*, 2022.

[4] M. Hausknecht, T.-C. Au, and P. Stone, "Autonomous intersection management: Multi-intersection optimization," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2011.

[5] B. Chalaki and A. A. Malikopoulos, "Optimal control of connected and automated vehicles at multiple adjacent intersections," *IEEE Transactions on Control Systems Technology*, 2021.

[6] Y. Wang, P. Cai, and G. Lu, "Cooperative autonomous traffic organization method for cavs in multi-intersection road networks," *Transportation research part C: emerging technologies*, 2020.

[7] K. Dresner and P. Stone, "A multiagent approach to autonomous intersection management," *J of artificial intelligence research*, 2008.

[8] P. Bender, Ö. Ş. Taş, J. Ziegler, and C. Stiller, "The combinatorial aspect of motion planning: Maneuver variants in structured environments," in *2015 IEEE Intelligent Vehicles Symposium (IV)*, 2015.

[9] H. Xu, C. G. Cassandras, L. Li, and Y. Zhang, "Comparison of cooperative driving strategies for cavs at signal-free intersections," *IEEE Transactions on Intelligent Transportation Systems*, 2021.

[10] B. Chalaki and A. A. Malikopoulos, "A priority-aware replanning and resequencing framework for coordination of connected and automated vehicles," *IEEE Control Systems Letters*, 2021.

[11] X. Lin, M. Li, Z.-J. M. Shen, Y. Yin, and F. He, "Rhythmic control of automated traffic—part ii: Grid network rhythm and online routing," *Transportation Science*, 2021.

[12] A. I. Mahbub and A. A. Malikopoulos, "Platoon formation in a mixed traffic environment: A model-agnostic optimal control approach," in *2022 American Control Conference (ACC)*, IEEE, 2022.

[13] J. Wang, Y. Zheng, Q. Xu, and K. Li, "Data-driven predictive control for connected and autonomous vehicles in mixed traffic," in *2022 American Control Conference (ACC)*, IEEE, 2022.

[14] G. Sharon and P. Stone, "A protocol for mixed autonomous and human-operated vehicles at intersections," in *International Conference on Autonomous Agents and Multiagent Systems*, Springer, 2017.

[15] C. Wu, K. Parvate, N. Kheterpal, L. Dickstein, A. Mehta, E. Vinitsky, and A. M. Bayen, "Framework for control and deep reinforcement learning in traffic," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2017.

[16] J. D. Little, M. D. Kelson, and N. H. Gartner, "Maxband: A versatile program for setting signals on arteries and triangular networks," 1981.

[17] P. Hunt, D. Robertson, R. Bretherton, R. Winton, Transport, and R. R. Laboratory, *SCOOT: A Traffic Responsive Method of Coordinating Signals*. TRRL Laboratory report, 1981.

[18] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, "Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control," in *Proceedings of the AAAI conference on artificial intelligence*, 2020.

[19] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, 2000.

[20] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry, "Implementation matters in deep rl: A case study on ppo and trpo," in *Int conference on learning representations*, 2019.

[21] A. Lazaric, "Transfer in reinforcement learning: a framework and a survey," in *Reinforcement Learning*, Springer, 2012.

[22] 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 *21st Int Conference on Intelligent Transportation Systems (ITSC)*, IEEE, 2018.

[23] R. Tedrake and the Drake Development Team, "Drake: Model-based design and verification for robotics," 2019.

[24] P. E. Gill, W. Murray, and M. A. Saunders, "Snopt: An sqp algorithm for large-scale constrained optimization," *SIAM review*, 2005.

[25] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International conference on machine learning*, PMLR, 2015.