

7 Appendix

7.1 Algorithm

Algorithm 1 TraCo for agent i

- 1: **Input:** Randomly initialize TraCo, actor and critic network f , π and V with weights φ , θ_π and θ_v
 - 2: **for** episode=1, T **do**
 - 3: Get agents' observations $\{o_1, \dots, o_n\}$
 - 4: Get $s_i = \{o_k, \dots, o_p\}$ according to the distance
 - 5: Compute $z_i = f(o_i, s_i)$, $a_i = \pi(o_i, z_i)$
 - 6: Compute counterfactual advantage function A_i^D according to equations (5, 10, 11)
 - 7: Compute A_i according to equation (6)
 - 8: Update with PPO rules
 - 9: **end for**
-

7.2 Experiment Platform and Scenarios

We use MetaDrive [36] as a simulator, which is capable for generating infinite scenarios with various road maps and traffic settings to enable generalizable RL. In our setup, we use current state, navigation info, and surrounding data encoded in a vector of 72 lidar-like measurements as agent observations, while the policy output is the acceleration and steering of the vehicle. As the mutual influence between vehicles decreases with distance, we define the in-domain state for the traffic coordinator as the information splicing of different vehicles within a 40-meter radius of the ego-vehicle.

As shown in Figure 5, we benchmark our method in four common autonomous driving tasks, which are described in detail as follows:

Bottleneck: The Bottleneck is to set up a narrow bottleneck lane between the eight lanes, forcing vehicles to give way and queue up to pass. The environment is initialized with 20 cars.

Tollgate: The Tollgate environment models the real-world behavior of vehicles passing through a tollgate, where agents are required to wait for a permission signal for 3 seconds before continuing. Failure to comply with this rule results in a failed episode. The environment is initialized with 40 cars.

Parking lot: The parking lot scenario in our simulation consists of 8 parking spaces. Spawn points for vehicles are scattered both within and outside the parking lot, leading to simultaneous entry and exit of vehicles and thereby increasing the level of difficulty. The environment is initialized with 10 cars.

Intersection: At an unprotected intersection scenario, vehicles are required to negotiate and judge the potential intentions of other parties in order to complete the task. The environment is initialized with 30 cars.

In this paper, we use three indicators to evaluate the performance of multi-agent algorithms. *success rate* is the ratio of vehicles successfully reaching the destination, *safety* is the vehicle non-collision rate, *efficiency* ≥ 0 indicates the difference between successes and failures in a unit of time $(N_{success} - N_{failure})/T$. Vehicles may travel at low speeds for the safety of driving, but this is not conducive to the effective passage of vehicles.

7.3 Ablation Studies

In our previous experiments, we employed the traffic coordinator network solely as a feature extraction network, without considering the counterfactual advantage function. Therefore, it is crucial to verify the validity of this function. As illustrated in Figure 7, TraCo w/o CAF performs worse than TraCo w/ CAF in all four autonomous driving tasks. This is because the traffic coordinator network, when equipped with a counterfactual advantage function, not only extracts in-domain features but also evaluates the agent's behavior based on these features. This evaluation allows for the

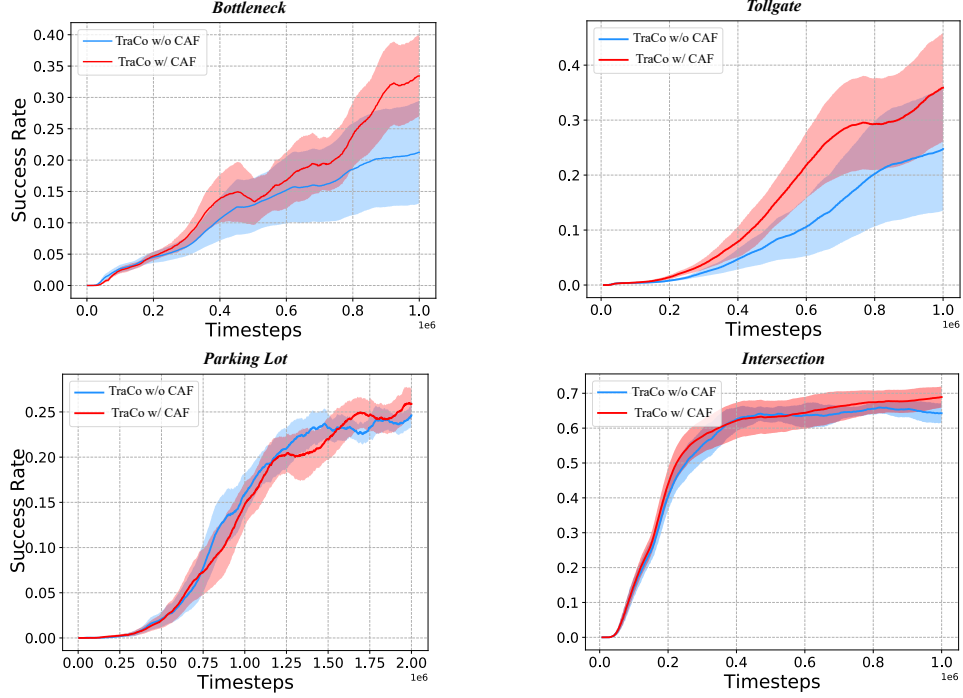


Figure 7: Performance comparison of TraCo with and without counterfactual advantage functions.

Table 1: The traffic coordinator network re-issues the command z according to the current situation at different time intervals, and the command z remains unchanged during this time interval.

	Bottleneck			Tollgate			Parking lot			Intersection		
	Success Rate	Efficiency	Safety	Success Rate	Efficiency	Safety	Success Rate	Efficiency	Safety	Success Rate	Efficiency	Safety
TraCo/1	0.36 ± 0.13	0.26	0.36	0.36 ± 0.19	0.22	0.38	0.27 ± 0.04	0.21	0.27	0.73 ± 0.05	0.51	0.73
TraCo/2	0.37 ± 0.09	0.27	0.37	0.32 ± 0.15	0.18	0.34	0.15 ± 0.07	0.11	0.16	0.72 ± 0.01	0.51	0.72
TraCo/4	0.38 ± 0.07	0.28	0.38	0.17 ± 0.16	0.09	0.2	0.14 ± 0.06	0.11	0.15	0.72 ± 0.03	0.51	0.72
TraCo/6	0.42 ± 0.09	0.3	0.42	0.25 ± 0.19	0.14	0.28	0.17 ± 0.04	0.13	0.17	0.73 ± 0.04	0.51	0.73
TraCo/8	0.34 ± 0.12	0.25	0.34	0.29 ± 0.18	0.16	0.31	0.21 ± 0.04	0.16	0.21	0.74 ± 0.02	0.53	0.74

372 measurement of the agent’s contribution to itself and the surrounding team, effectively addressing
373 the interests balance problem.

374 Taking inspiration from the behavior of real-life traffic coordinators, who issue commands based
375 on vehicle behavior and intersection information at time intervals rather than continuously directing
376 vehicles, we designed different time intervals for the Traffic Coordinator Network (TroCo) to extract
377 features. As shown in Table 1, our experiments reveal that in complex traffic environments such
378 as Tollgate and Parking lot, where obstacles are numerous, roads are congested, and the behavior
379 of domain agents is difficult to predict, frequent direction is necessary to ensure optimal vehicle
380 decision-making. However, in Bottleneck and Intersection tasks, where the purpose of the vehicle
381 is clear, and the behavior is more predictable, frequent direction may interfere with the agent’s
382 decision-making. In such cases, an appropriate time interval can enhance the consistency of the
383 agent’s behavior.