CTDE vs. Independent PPO: Benchmarking Multi-Agent Coordination in Flow-SUMO Vehicle Platoons

Stanford CS229 Project

Taylor Tam, Vansh Gadhia, Leon Liu
Department of Computer Science
Stanford University
taylor52@stanford.edu, vanshg@stanford.edu, lliu32@stanford.edu

1 Introduction

As autonomous vehicles (AV) become mainstream, the control problem evolves from optimizing individual vehicle trajectories to orchestrating coordinated behaviors across many coupled agents. Vehicle platooning, where multiple AVs travel in tight formations, offers a practical solution, generating 7-20% fuel savings, greater throughput from reduced headways, and improved safety through synchronized V2V braking. These benefits have driven major deployments (PATH AHS, SARTRE, and Peloton) and position platooning core to future automated highway systems.

Recent work applies multi-agent reinforcement learning (MARL) to cruise control, platoon formation, and cooperative merging, raising central design questions around interaction modeling, continuous control, and stability in mixed-autonomy settings. A major distinction is between independent learning (IL), where each agent trains separately, and centralized training with decentralized execution (CTDE), which conditions learning on joint state information. Despite growing interest in CTDE for cooperative driving, previous work has largely overlooked disturbance-propagation metrics that define platoon stability, limiting our understanding of how centralized training influences real-world longitudinal dynamics.

We present a controlled comparison of IL and CTDE in a multi-vehicle platooning environment, measuring the impact of centralized critics and coordinated training signals on (1) convergence behavior, (2) inter-vehicle gap regulation, and (3) emergent platoon efficiency. Our results clarify how centralized training influences stability and coordination in tightly coupled multi-agent driving systems.

2 Related Work

Research on learning-based traffic control has expanded significantly with the introduction of the Flow framework by Wu et al. (2017), which integrates reinforcement learning with SUMO to provide realistic vehicle dynamics and reproducible multi-agent environments. Flow has been used to demonstrate that RL agents can reduce congestion and dampen stop-and-go traffic patterns (Vinitsky et al., 2018), offering a scalable alternative to traditional analytical controllers. Most of these studies focus on single-agent or lightly coupled settings.

A parallel line of research examines multi-agent reinforcement learning under cooperative and mixed environments. CTDE frameworks such as MADDPG (Heskes et al., 2020) introduced the idea of centralized critics to stabilize learning in multi-agent settings. CTDE methods generally show improved coordination compared to independent learning, though many evaluations rely on simplified simulation tasks and do not address stability or safety.

Work specifically targeting coordinated vehicle behavior provides additional context. Li et al. (2021) review MARL methods for platoon formation and cooperative longitudinal control, noting that learning-based controllers can adapt to nonlinear dynamics. Other studies look at multi-vehicle

coordination in the context of mixed autonomous–human traffic (Yu et al., 2020), demonstrating that MARL agents can still learn under partial observability. However, these works focus more on lane change or overall traffic flow than on the disturbance control and spacing stability that matter most in platooning.

3 Dataset and Features

3.1 Network Configuration

We use the Flow framework (Wu et al., 2017) interfaced with SUMO (López et al., 2018) to simulate a continuous-control mixed-autonomy highway environment. Flow provides microscopic vehicle dynamics, realistic leader–follower and lane-change behavior, and an RLlib-compatible API for multi-agent training.

The environment is a partially observable multi-agent MDP on a straight four-lane highway (single directed edge). The road length is

$$L \in \{1000 \text{ m (training)},\ 3000 \text{ m (visualization)}\},$$

with speed limits

$$v_{\max} \in \{30 \text{ m/s (training)},\ 15 \text{ m/s (visualization)}\}.$$

SUMO advances dynamics at $\Delta t = 0.1$ s (10 Hz), and each episode lasts $T = 1500$ steps (150 s).

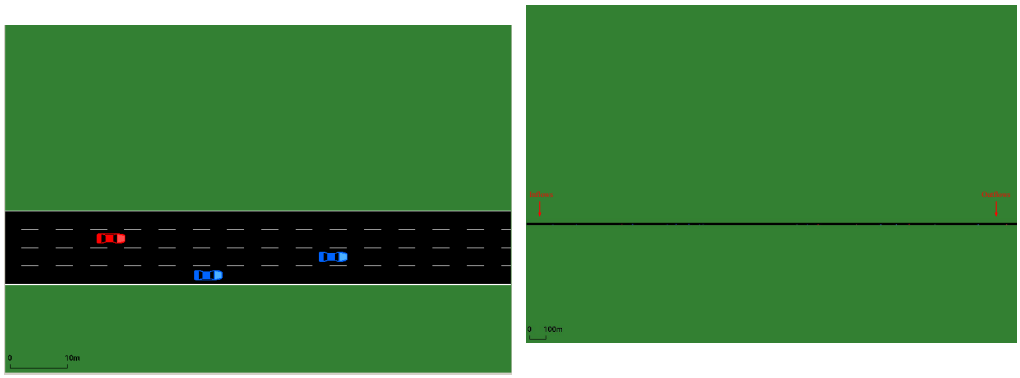

Figure 1: Visualization of the Flow–SUMO highway environment. Left: four-lane straight highway used for training and evaluation. Right: inflow/outflow configuration with mixed human and RL vehicle insertion rates.

3.2 Vehicle Models

Human-driven and RL vehicles enter according to independent Poisson processes with rates

$$\lambda_{\text{human}} = 900 \text{ veh/h}, \qquad \lambda_{\text{RL}} = 270 \text{ veh/h},$$

yielding a 70/30 human–RL mix.

Human vehicles follow the Intelligent Driver Model (IDM):

$$\dot{v}_i = a_{\max}\left(1 - (v_i/v_0)^4 - (s^*/h_i)^2\right), \qquad s^* = s_0 + v_i\tau + \frac{v_i \Delta v_i}{2\sqrt{a_{\max} b_{\max}}}.$$

We use standard parameters:

$$a_{\max} = 1.8,\ b_{\max} = 4.5,\ \tau = 1.0,\ s_0 = 4.0.$$

RL vehicles follow simple kinematic dynamics:

$$u_i(t) \in [-3, 3], \qquad v_{i,t+\Delta t} = v_{i,t} + u_i \Delta t, \quad x_{i,t+\Delta t} = x_{i,t} + v_i \Delta t + \tfrac{1}{2} u_i \Delta t^2.$$

3.3 Observation and Action Spaces

Each RL agent receives a local observation $o_i(t) \in \mathbb{R}^d$, with dimension

$$d = \begin{cases} 5 & \text{(longitudinal only)}, \\ 12 & \text{(lane changing)}. \end{cases}$$

Features include ego speed, headway, relative speed, lane index, normalized speed, adjacent-lane indicators, lateral velocity, and neighbor distances. Actions consist of acceleration only, or acceleration + discrete lane change:

$$\mathcal{A} = \begin{cases} [-3, 3]^N, & \text{longitudinal}, \\ ([-3, 3] \times [-1, 1])^N, & \text{with lane change}. \end{cases}$$

3.4 Reward Function Design

Each agent receives a composite reward

$$r_i = r_{\text{speed}} + r_{\text{slow}} + r_{\text{stuck}} + r_{\text{sync}} + r_{\text{safety}} + r_{\text{smooth}} + r_{\text{lane}},$$

encouraging high throughput, safe headways, smooth accelerations, synchronized platoon motion, and effective lane changes.

Speed rewards (maintaining v_{target}), penalties for near-zero or unnecessarily slow motion, and free-road bonuses follow standard Flow practice. Stuck penalties increase when the leader is slow or when the agent is part of a larger congestion chain. Synchronization rewards reduce deviations from the mean RL speed.

Safety penalties apply for headways < 6 m; smoothness penalties apply for $|a_i| > 2\,\text{m/s}^2$. Lane-change terms reward successful escapes from slow leaders and penalize oscillatory or late changes. The reward ranges roughly from -8 to $+4.8$.

4 Methods

4.1 Independent PPO (IL)

In the independent learning (IL) baseline, each RL vehicle is treated as its own agent and learns a separate policy $\pi_i(a_i \mid o_i)$ using PPO. Agents do not share parameters, observations, or gradients. Each agent receives only its local observation $o_i \in \mathbb{R}^9$ and optimizes its own return. This setup represents the common decentralized-learning paradigm in traffic RL, where coordination must emerge solely through environment interactions.

4.2 Centralized Training with Decentralized Execution (CTDE PPO)

In the CTDE variant, all RL vehicles share a single decentralized policy $\pi_\theta(a_i \mid o_i)$ but learn using a centralized value function $V_\phi(s)$ conditioned on global state information (e.g., joint observations, leader states, and inter-vehicle spacing). During execution, the critic is discarded and each agent acts using only its local observation, preserving decentralization. Centralized critics stabilize gradient estimates and provide richer credit assignment in tightly coupled multi-agent systems such as platoons.

4.3 Training Setup

Training takes place on a 1000 m, four-lane highway segment with a speed limit of 30 m/s. Each episode lasts 3000 simulation steps with timestep 0.1 s (300 s). Traffic contains 10 vehicles: 2 RL agents and 8 human-driven IDM vehicles.

Each RL agent receives a 9-dimensional observation vector containing ego speed, headway, relative speed, lane index, normalized speed, estimated acceleration, normalized time-to-collision, normalized leader speed, and a slow-leader indicator.

CTDE and IL use identical reward functions, action spaces, and observation structures, isolating the contribution of the centralized critic. Additionally, both CTDE and IL use identical hyperparameters for fair comparison.

Hyperparameter	Value
Learning rate	3×10^{-4}
Discount factor γ	0.99
GAE parameter λ	0.95
PPO clip ratio	0.1
SGD iterations per batch	10
Minibatch size	128
Training batch size	4000 agent-steps
Network architecture	2×256 FC layers, `tanh` activation
Training duration	150 iterations

Table 1: PPO hyperparameters used for both IL and CTDE.

4.4 Evaluation Metrics

We evaluate IL and CTDE across four dimensions (stability, coordination, efficiency, and safety) using the following metrics:

- **Spacing variance:** variability of inter-vehicle headways (stability).
- **Oscillation amplitude:** magnitude of stop-and-go waves via velocity derivatives.
- **Damping ratio:** attenuation of disturbances across the platoon.
- **Synchronization index:** alignment of RL agents' velocities (coordination).
- **Avg. velocity & speed variance:** overall efficiency and smoothness.
- **Throughput:** vehicles per second exiting the network.
- **Safety metrics:** minimum TTC, near-collision count, collision count.
- **Policy divergence & action correlation:** behavioral diversity and interaction structure.
- **String stability ratio:** disturbance amplification from leader to follower (> 1 indicates instability).

Together, these metrics capture the stability, safety, coordination, and efficiency properties of multi-agent platooning, enabling a direct comparison of IL and CTDE.

5 Experiments and Results

5.1 Training Performance

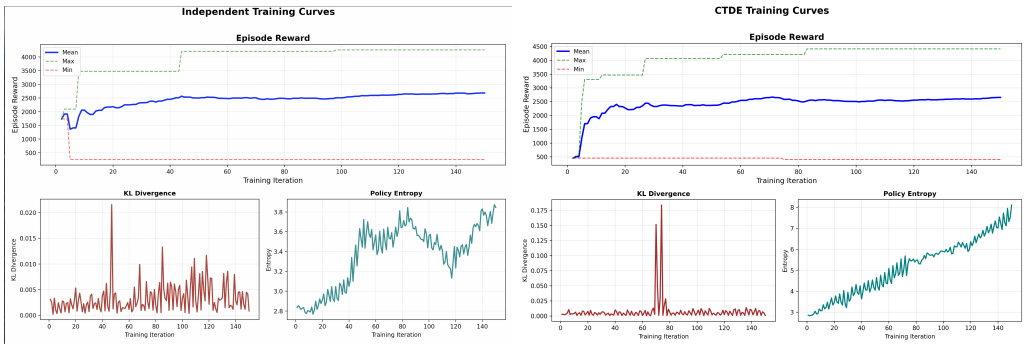

Figure 2: Training curves for Independent PPO (left) and CTDE PPO (right). CTDE achieves faster reward growth and sustains higher exploration.

Figure 2 shows the learning dynamics of Independent PPO and CTDE PPO. IL PPO improves slowly with low entropy and small KL divergence, indicating limited exploration and incremental updates based solely on local observations. CTDE converges faster, with higher entropy and sharper—but stable—KL adjustments driven by centralized value estimates. These patterns show that global information during training substantially accelerates the discovery of coordinated behaviors.

5.2 Stability, Coordination, and Safety Outcomes

CTDE outperforms Independent PPO across nearly all stability and cooperation metrics (Table 2). Spacing variance drops by more than $5\times$, oscillation amplitude decreases, and the damping ratio doubles, indicating significantly improved suppression of longitudinal disturbances. CTDE also achieves higher synchronization and greater policy divergence, suggesting more coherent and specialized behaviors among agents.

Metric	**CTDE PPO**	**Independent PPO**
Spacing Variance	2349.6	13771.9
Osc. Amplitude	145.2	178.1
Damping Ratio	3.13	1.60
Avg. Velocity (m/s)	13.00	10.82
Speed Var.	6.43	22.19
Throughput (veh/s)	0.077	0.102
Near Collisions	0	5
Min TTC (s)	2.68	1.08
Sync. Index	0.867	0.701

Table 2: Performance of CTDE vs. Independent PPO across stability, efficiency, and safety metrics.

Safety differences are particularly pronounced: Independent PPO produces five near-collisions and a lower minimum TTC, while CTDE avoids all unsafe events despite identical reward structures. Although Independent PPO achieves slightly higher throughput, this comes with substantially higher speed variance and degraded stability.

Both methods remain string-unstable (> 1 amplification), but CTDE reduces disturbance growth and maintains smoother behavior overall. These results indicate that centralized learning signals are highly beneficial for cooperative longitudinal control, while purely decentralized learning struggles to acquire safe and stable policies.

6 Conclusion and Discussion

This study presents the first controlled comparison of Independent PPO and Centralized Training with Decentralized Execution (CTDE) in a multi-agent vehicle platooning setting. Our evaluation focuses on disturbance-propagation and stability metrics, which are key elements of platoon safety that are underrepresented in existing MARL work.

CTDE consistently outperforms independent learning across nearly all coordination and safety dimensions. It achieves over a 5× reduction in spacing variance and doubles the damping ratio, indicating stronger suppression of velocity fluctuations and shockwave disturbances. Moreover, CTDE policies exhibit higher synchronization indices and policy divergence, suggesting both tighter coordination and emergent role specialization among agents. Safety outcomes are also more favorable: CTDE eliminates near-collisions and maintains a safer minimum time-to-collision, while Independent PPO results in multiple risky interactions despite slightly higher throughput.

These findings highlight a fundamental advantage of centralized critics in tightly coupled multi-agent control. Even under decentralized execution and partial observability, CTDE yields more robust and stable behaviors. This suggests that global context during training is sufficient to instill cooperation and disturbance mitigation in agent policies, without requiring runtime communication.

However, both algorithms remain string-unstable. This reveals a key shortcoming of existing MARL approaches: while CTDE enhances coordination and local smoothness, it does not yet meet the formal stability guarantees of classical control theory.

Future work should (1) extend evaluations to larger platoons and higher vehicle densities, (2) incorporate explicit V2V communication to test whether runtime information exchange improves disturbance damping, and (3) compare against recent graph-based and value-decomposition MARL architectures designed for cooperative control. Additionally, integrating analytical tools from classical control like $\mathcal{H}_\infty$ stability margins or frequency-domain transfer functions could provide deeper insight into the limits and failure modes of learned policies. Bridging the gap between MARL and provable stability remains a compelling and necessary direction for safe multi-agent driving systems.

7 Contributions

- **Taylor Tam**: Implemented CTDE PPO; integrated lane-changing into both IL and CTDE agents; designed and refined shared reward function; created and configured the Flow–SUMO environment; ran experiments, collected and analyzed evaluation data; led the writing and revision of the paper.
- **Vansh Gadhia**: Implemented the Independent PPO baseline; contributed to the lane-changing extensions; assisted with experiment setup, results analysis, code cleanup, and debugging; contributed substantially to writing and polishing the report.
- **Leon Liu**: Setup initial Independent PPO training pipeline; refined the reward functions for both IL and CTDE settings; assisted with code maintenance, tuning, and debugging; conducted research on existing literature; edited and refined the final manuscript.

References

Tom Heskes, Sai Wang, and Jens Kober. 2020. Cooperative multi-agent reinforcement learning for highway platooning. *IEEE Transactions on Intelligent Transportation Systems*.

Jinlong Li, Wei Xu, and Jian Wang. 2021. Multi-agent reinforcement learning-based cooperative control for vehicle platooning: A review. *IEEE Transactions on Intelligent Transportation Systems*. Survey of MARL methods for vehicle platooning and coordination; good conceptual support.

Pablo Álvarez López, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. Microscopic traffic simulation using sumo. *IEEE Intelligent Transportation Systems Conference*.

Eugene Vinitsky, Abdul Rahman Kreidieh, Luc Le Flem, Cathy Wu, and Alexandre M. Bayen. 2018. Benchmarks for reinforcement learning in mixed-autonomy traffic. In *Conference on Robot Learning (CoRL)*. Extends Flow to benchmark multi-agent RL in mixed-autonomy settings; directly relevant to platooning.

Cathy Wu, Alexandre Gomes, and Aleksandr M Bayen. 2017. Flow: Architecture and benchmarking for reinforcement learning in traffic control. In *Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE.

Chao Yu et al. 2020. Multi-agent ppo: Stable multi-agent learning with policy clipping. In *NeurIPS Deep RL Workshop*.