Multi-Agent Reinforcement Learning for Cooperative Platoon Lane-Change Decision in CARLA

Alvin Tan Yee Jia
National Cheng Kung University
Tainan, Taiwan
e94115011@gs.ncku.edu.tw

Advisor:
Hung-Chang Hsiao
National Cheng Kung University
Tainan, Taiwan

Motivation

Autonomous driving is **not only about controlling one car.**
In platoon driving, vehicles must cooperate when unexpected road events occur.
If the **leader suddenly stops**, the lane becomes blocked, and the followers must decide whether to **wait, overtake, return, or regroup.**

Problem Statement

This project focuses on a **stopped-leader overtaking task.**
When the leader blocks the lane, the followers must learn safe decisions for:
Follow → Detect Stop → Overtake → Return → Regroup

Why Is This Difficult

Even with only three high-level actions, the agent must make decisions over many time steps.

$$N\_sequence = |A|^T = 3^T$$

Action space |A| = 3
Decision steps T

Decision steps T	Possible action sequences 3^T
50	7.18×10^23
100	5.15×10^47
200	2.66×10^95

The difficulty is not the **number of actions**, but the **timing of decisions.**
For two followers, the joint action space becomes: $(3^2)^T = 9^T$
This makes platoon overtaking a multi-agent sequential decision-making problem.

System Architecture

MARL Flow:

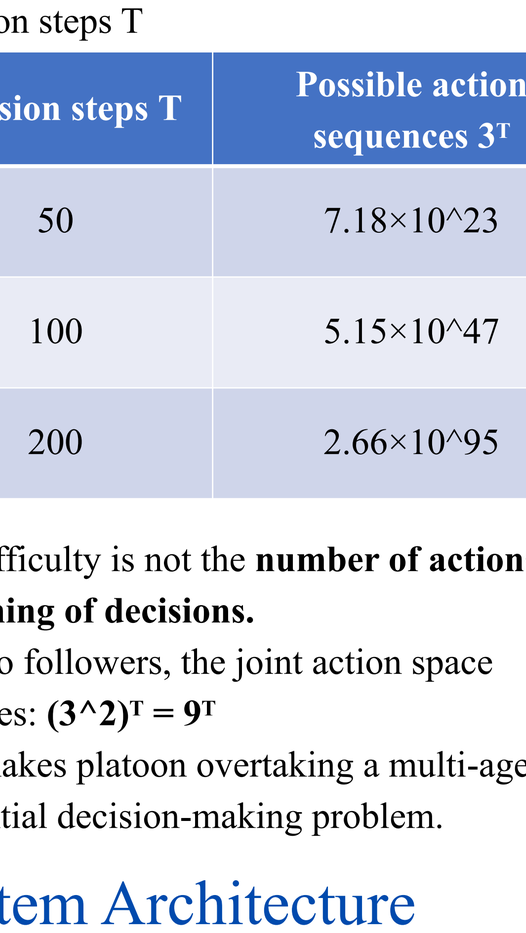

Fig. MARL Platoon Control with CTDE.

CTDE Core:

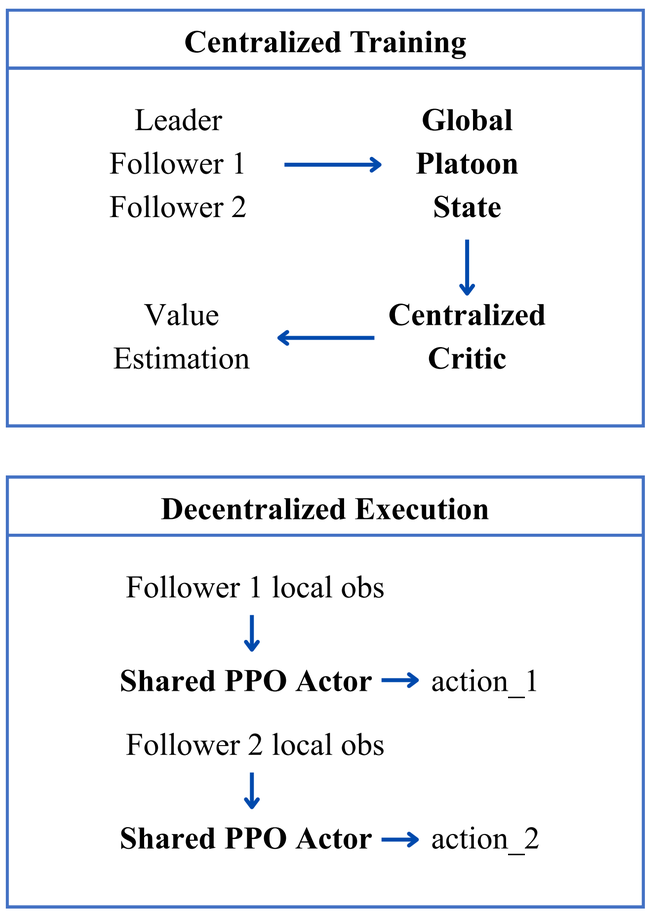

Fig. CTDE design: the critic learns from the full platoon state during training, while each follower executes independently using local observations.

Technical Stack

- Simulator: **CARLA 0.9.15**, Town01_Opt
- Language: **Python 3.7**
- MARL Environment: **PettingZoo ParallelEnv**
- Training: **Ray RLlib PPO**
- **CTDE**: shared-policy actor + centralized critic
- Baseline: **Stable-Baselines3 PPO**
- Control: **rule-based waypoint controller**
- Safety: lane availability check + action filtering

Single-Agent PPO Result

A single PPO-controlled follower is trained in a **1 leader + 1 follower** stopped-leader scenario to learn **when to overtake** and **when to return safely**.

Test Condition	Success Rate	Collision Rate	Mean Reward
Fixed stop step = 40, speed = 15	100%	0%	133.69
Random stop timing	100%	0%	130.66
Fixed stop step = 40, speed = 12	100%	0%	133.68

Result highlight:
The policy succeeds under both **fixed and random leader stop timing**, showing that it does not simply memorize one stop time.

MARL Training Result

The **MARL-CTDE** setup shows preliminary cooperative learning with two RL followers.

- 1 leader + 2 RL followers
- **Shared-policy PPO** with **centralized critic**
- Mean episode reward: **−299.03 → +212.31**
- Training scale: **52k environment steps**, **13 PPO iterations**
- Coordination conflicts: 0 recorded

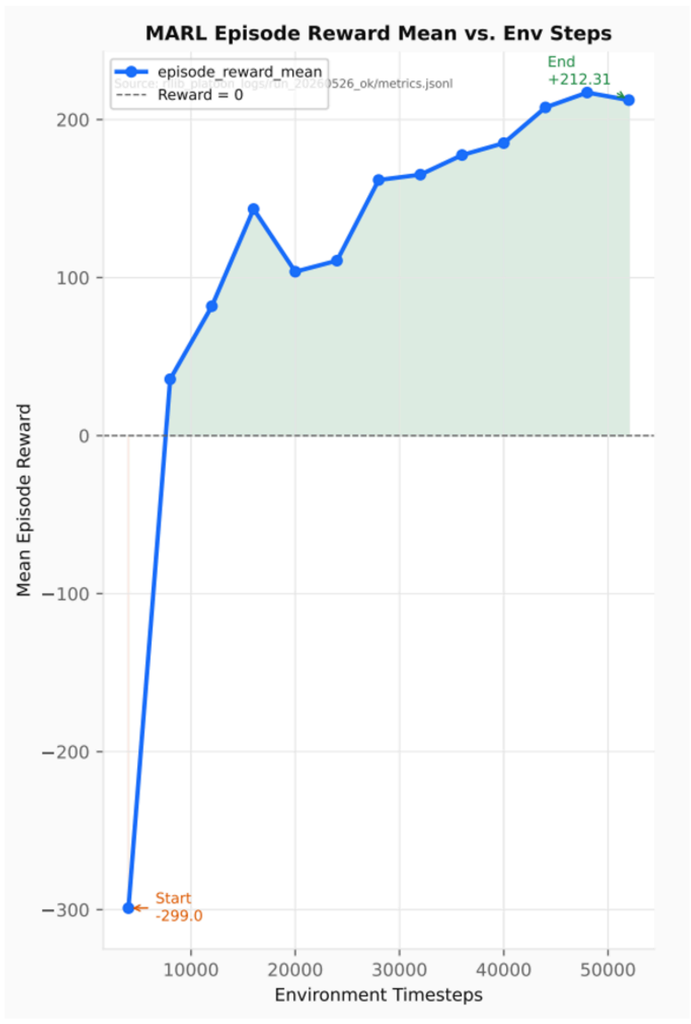

Fig. MARL training curve.

Conclusion

This project builds a **CARLA-based stopped-leader overtaking task** for autonomous platoons.

The single-agent PPO baseline verifies that the **core overtaking decision is learnable.**

The MARL-CTDE framework further extends the task toward **cooperative platoon decision-making** with shared-policy PPO and a centralized critic.

Future Work

- Improve robustness across different **CARLA roads, spawn points, and waypoint topologies.**
- Stabilize **MARL regrouping behavior.**
- Scale from **2 followers to n followers.**

Demo Video

Scan the QR code to view:
- Single-Agent PPO stopped-leader overtaking
- MARL platoon two-follower cooperative overtaking