Multi-Agent Reinforcement Learning for Safe Autonomous Driving Under Pedestrian Behavioral Uncertainty

Prakash Aryan[1], Kaushik Raghupathruni[1], Timo Kehrer[1], Sebastiano Panichella[1,2]

[1]University of Bern, Switzerland [2]AI4I, Italian Institute of Artificial Intelligence, Turin, Italy

ICRA 2026 VIENNA — IEEE International Conference on Robotics and Automation

8th Long-term Human Motion Prediction Workshop

TL;DR **Joint training of the SDC with 12 pedestrians driven by hidden personality traits reaches 78% goal completion at 14% collisions, with the SDC-pedestrian speed differential exposing where jaywalking risk concentrates.**

1. Problem: Scripted Pedestrians Are Too Easy

Most SDC simulators rely on scripted pedestrians governed by fixed crossing rules, so the vehicle is seldom evaluated against the behaviors most responsible for real-world collisions, such as jaywalking and last-second hesitation. Because the personality traits underlying these decisions are not directly observable, a meaningful test environment requires pedestrians that respond to the vehicle rather than execute predetermined scripts.

- Scripted policies underrepresent jaywalking and hesitation, which keeps reported collision rates artificially low.
- Pedestrians that adapt to the vehicle expose failure modes that fixed scripts cannot reproduce.

Research Question

Can co-training yield pedestrian behavior realistic enough to expose safety failures while the SDC still learns a policy that reaches its goal?

2. Co-Training Design

Pedestrians learn a high-level go/wait policy on top of scripted Dijkstra locomotion. The SDC is modeled as a kinematic bicycle (max 8.33 m/s, wheelbase 2.5 m). A latent personality trait $\tau_j \in [0, 1]$, sampled per episode and unobserved by the SDC, determines each pedestrian's jaywalking probability:

$$P(\text{jaywalk}) = \tau_j \times 0.25$$

Training follows the CTDE paradigm: a centralized critic combined with decentralized actors, where pedestrian actors share parameters. Since τ_j is not observable, the SDC must infer crossing risk from kinematics and scene geometry alone.

Interaction Loop

Pedestrians produce crossing demand of varying difficulty, and the SDC learns to complete its route while responding to that demand.

3. Training Setup

Map 120×120 m urban layout
Episode 500 steps, 10 Hz
Parallelism 512 envs
Total 6.55×10^8 steps
PPO $\epsilon = 0.2$, 4 epochs
Agents 12 pedestrians + SDC
Rollout 256 steps
Training 5,000 updates
Speed ∼558k FPS
GAE $\gamma = 0.995$, $\lambda = 0.95$

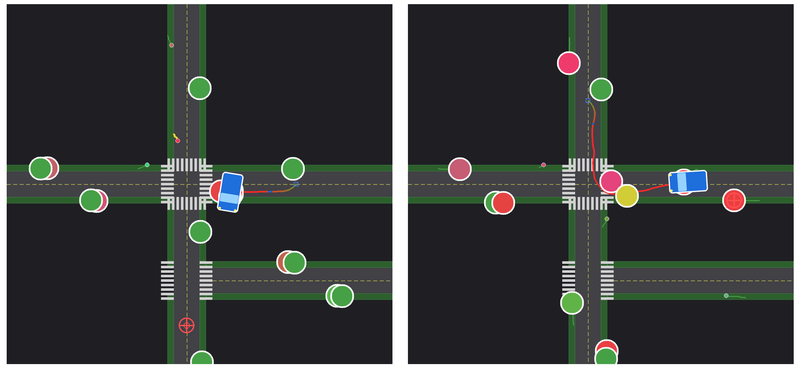
Reward Design

- **Pedestrians:** waypoint progress, waypoint completion, smart waiting, collision penalty.
- **SDC:** goal arrival, collision, speeding-near-crossing, off-lane, and heading penalties.

Baselines

Six controllers: random, constant-speed, rule-based, rule-based+braking, single-agent RL, and MARL co-training.

4. Urban Scenario



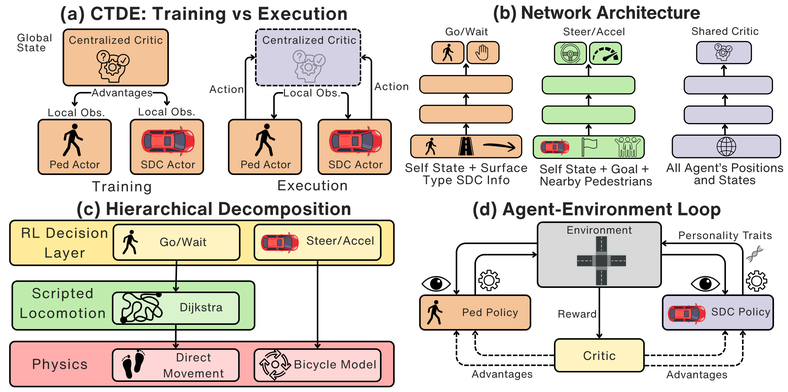
(a) Collision with jaywalker *(b) Successful avoidance*

Original paper simulator images. Blue rectangle = SDC; dots = pedestrians; red crosshair = goal.

Scenario Reading

- Collision cases correspond to high-speed SDC approaches near an uncertain crossing.
- Avoidance cases illustrate the desired behavior: continued progress toward the goal while decelerating around high-risk pedestrians.

5. MAPPO System Architecture



Training/execution, network architecture, hierarchical control, and the agent-environment loop.

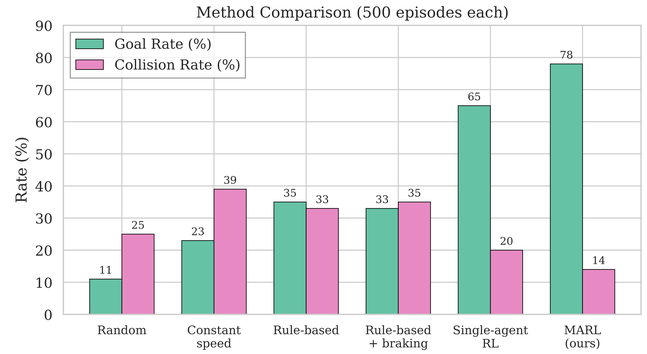
Design Reading

- A centralized critic improves learning, while decentralized actors remain executable without privileged training state.
- Pedestrian locomotion stays scripted at the route-following layer; RL controls the risky go/wait decisions.

6. Experimental Results

MARL co-training is the only configuration in which goal completion increases and collision rate decreases simultaneously.

- MARL (ours): **78% goal completion** at **14% collisions**, the strongest joint result across all six baselines.
- Collisions are reduced by **30%** relative to single-agent RL with no loss in goal completion.



Goal and collision rates across methods, 500 episodes each.

7. Discussion

Co-trained pedestrians produce closer and less predictable encounters than scripted partners, yet the resulting SDC policy transfers back to scripted evaluation with goal completion improving from 65% to 76%. The jaywalking-rate ablation shows the policy is robust under moderate uncertainty, but performance degrades sharply when jaywalking dominates the scene.

Interpreting the Results

The 78% goal completion is meaningful because it coincides with the lowest collision rate. High completion alone can reflect aggression, while low collisions alone can reflect over-conservatism. At 50% jaywalking, goal completion falls to 64% and collisions rise to 28%, making latent crossing intent the central stressor rather than background noise.

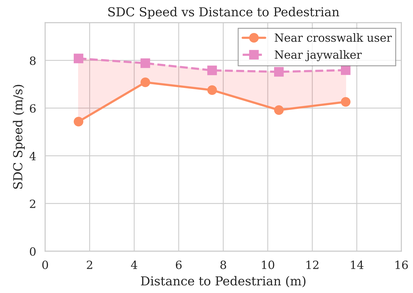
Stress-Test Use

The framework also serves as a stress-test instrument: controllers are trained, rollouts are mined for failures, and high Δv encounters are prioritized for replay and controller revision. This makes the simulator useful for ranking policies by where they fail, not only by aggregate success rates. Overall, the results argue for evaluating autonomous-driving policies with encounter-level behavioral uncertainty, not just end-of-episode scores. Co-training is therefore best read as realistic scenario generation, not as a claim that human pedestrians optimize the same joint objective as the SDC.

8. Uncertainty Quantification

Speed Differential: $\bar{v}_c = \frac{1}{|T_c|} \sum_{t \in T_c} v_{\text{sdc}}(t)$, with $\Delta v = \bar{v}_{\text{jw}} - \bar{v}_{\text{cw}}$.

- At 0–3 m the SDC travels **2.65 m/s faster** near jaywalkers.
- Jaywalking accounts for **13% of crossings** but **62% of collisions**.



SDC speed vs. distance to pedestrian by crossing type.

9. Future Work

Jaywalker on road

Jaywalker on crosswalk

VR participant study

- **CARLA/UE5 heterogeneous traffic:** scale beyond the compact map to richer urban layouts with mixed vehicles, pedestrians, and controllable crossing intent.
- **VR human-participant validation:** compare immersive participant decisions with learned pedestrian policies and the speed-differential risk signal.
- **Generalization checks:** test unseen maps, varied traffic densities, stronger planning baselines, and real-world validation of the speed-differential metric.
- **Policy diagnostics:** add speed-based collision criteria, per-agent-type critics, and safety layers around learned SDC controllers.

10. Takeaways and Resources

Best Configuration: 78% goal completion at 14% collisions with learned pedestrians; the same controller reaches 76% goal completion against scripted pedestrians.
Stress-Test Signal: raising the jaywalking prior from 13% to 50% reduces goal completion 76%→64% and increases collisions 16%→28%.

Contact: Prakash Aryan, University of Bern
prakash.aryan@unibe.ch

Paper Code Demo