Group 17

Curriculum Based RL for Autonomous Driving

Investigating structured progression from simple to complex driving scenarios to improve sample efficiency and generalization in simulated highway environments.

Motivation and Challenge

Complex & Safety-Critical

Autonomous driving demands robust agents adaptable to diverse, dynamic environments.

Sample Inefficiency

Training agents for complex traffic scenarios is often sample-inefficient and unstable.

Curriculum Learning

Structuring training from simpler to more complex tasks can improve efficiency and generalization.

Related Work: Curriculum RL & Simulators

Curriculum RL Frameworks

Ordered sequences of MDPs improve sample efficiency and overcome local optima.

Driving Simulators

HighwayEnv and MetaDrive provide diverse environments for RL research.

Curriculum for Autonomous Driving

Approaches like CuRLA vary environment complexity or reward shaping over training stages.

Reinforcement Learning Algorithms

1

Deep Q-Networks (DQN)

Value-based, off-policy method for discrete actions, extending Q-learning to high-dimensional settings.

2

Proximal Policy Optimization (PPO)

On-policy actor-critic method, directly parameterizing stochastic policy with clipped objective.

3

SimpleDQN

Our own implementation of DQN, used for comparative analysis.

Experimental Setup: HighwayEnv

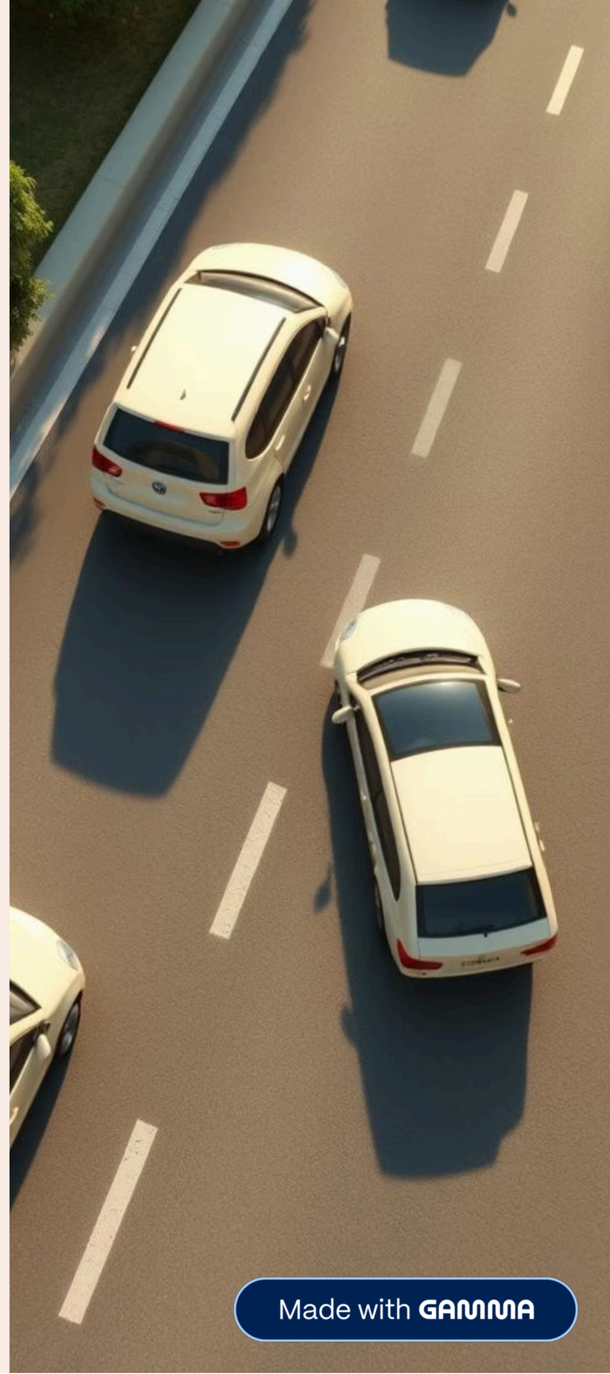
Tasks & Configuration

- Multi-lane highway driving (highway-v0)
- Highway merging (merge-v0)
- Signalized intersections (intersection-v0)
- Circular roundabouts (roundabout-v0)

Fixed episode duration (40s), configurable traffic density.

Observation & Action Spaces

- Kinematics observation: position, velocity, acceleration, lane info, distances.
- DiscreteMetaAction: keep lane, change lane (left/right), accelerate, brake.



Experimental Setup: MetaDrive

Reward Shaping

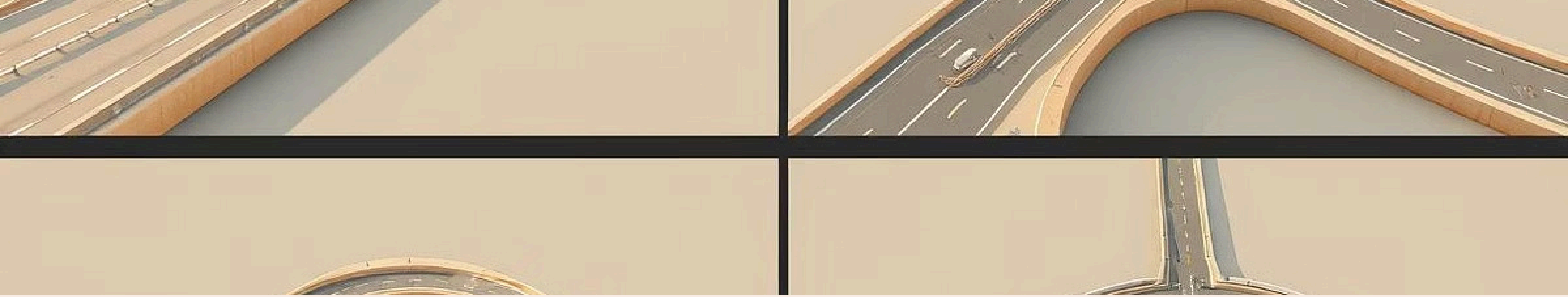
- **Base term** proportional to the original environment reward.
- **Speed term** rewarding the agent for driving up to a **stage-specific speed limit**.
- **Penalties** for collisions, off-road events, and traffic violations.
- **Terminal success bonus** for safe arrival at the destination.
- **Small per-step penalty** to discourage unnecessarily long episodes.

State & Action Spaces

- **DQN:** Uses a discrete-action wrapper, mapping a small set of steering-throttle pairs to discrete actions.
- **PPO:** Operates directly in the continuous action space.
- Meta-actions include lane changes (left, right, maintain lane) and speed adjustments (accelerate, decelerate, maintain speed) with values ranging from [-1,1] for both discrete and continuous spaces.

Evaluation

- For each **stage Ck in non-curriculum**, a separate model is trained from scratch using the stage's budget.
- A **held-out scenario** on an **unseen 6-block map** (denoted "SCrRXO") - consists of straight, circular, ramp, intersection, and roundabout segments with **30% traffic.**
- **Evaluation metrics** like average return and episode length are logged to analyze performance in new, structured environments.



Curriculum Regimen: HighwayEnv

01

Stage 1: Highway Low

Single block highway-v0, low traffic (ρ=0.20). Focus: cruising, lane discipline.

02

Stage 2: Highway Merge

Highway-v0 (ρ=0.25), merge-v0 (ρ=0.30). Focus: merging maneuvers.

03

Stage 3: Highway Merge Intersection

Highway-v0 (ρ=0.30), merge-v0 (ρ=0.35), intersection-v0 (ρ=0.35). Focus: dense highway, merging, intersections.

04

Stage 4: All Blocks

Highway-v0 (ρ=0.35), merge-v0 (ρ=0.40), intersection-v0 (ρ=0.40), roundabout-v0 (ρ=0.45). Focus: order-invariant generalization.

Curriculum Regimen: MetaDrive

01

Stage C0: Straight, no traffic ('S')

- Focus on basic control (road navigation without obstacles).
- Penalties for going off-road and step-penalty to reach destination faster.

02

Stage C1: Roundabout ('O')

- More complex road layout with no traffic to navigate turns with proper throttle and speed.
- Greater success bonus and step-penalty.

03

Stage C2: Light Traffic

- Introduces traffic of 5% intensity.
- 10-block map consisting of different types of randomly initialized map pieces.
- Penalizes collisions with traffic and obstacles.

04

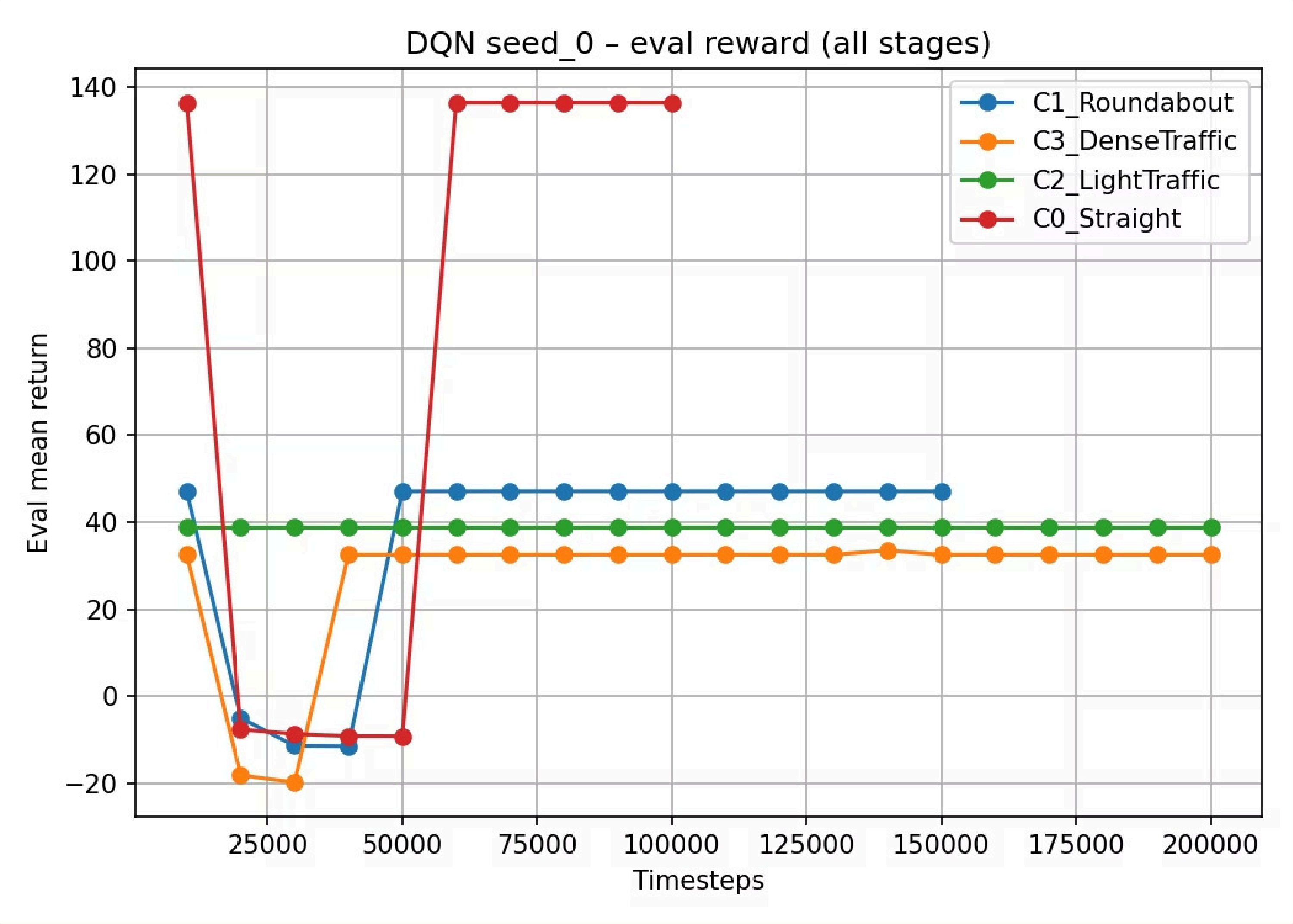
Stage C3: Dense Traffic

- Introduces traffic of 30% intensity.
- 20-block map.
- Penalizes collisions with traffic and obstacles on a greater level with increased bonus as well.

MetaDrive Results - Non-Curriculum

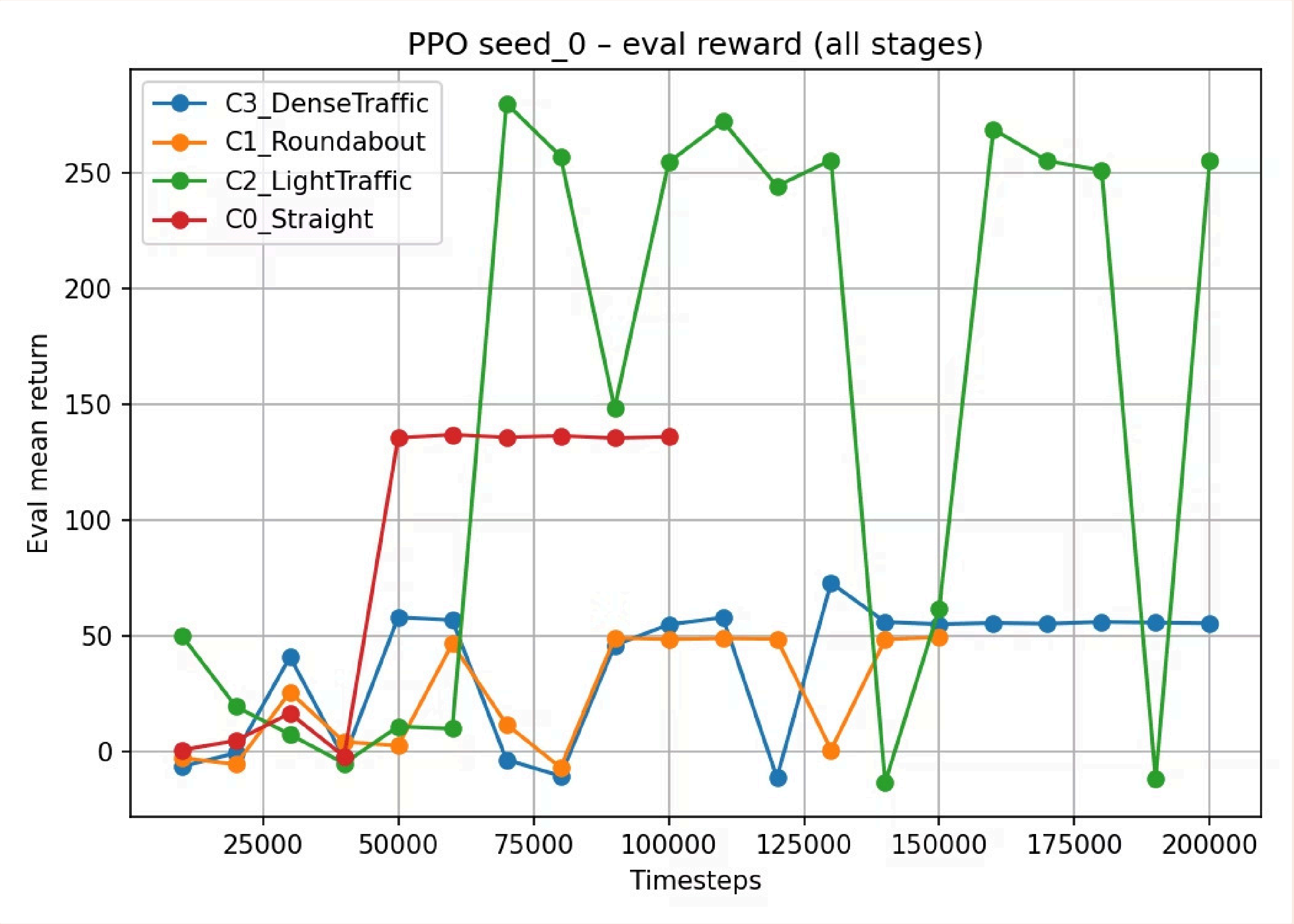
Stages	DQN Mean Training Reward ± Std	PPO Mean Training Reward ± Std
C0 (Straight)	136.2 ± 5.1	128.5 ± 4.2
C1 (Roundabout)	47.3 ± 0.0	52.1 ± 8.6
C2 (Light Traffic)	38.9 ± 0.0	43.5 ± 6.8
C3 (Dense Traffic)	32.5 ± 0.0	41.2 ± 7.5

MetaDrive Results - Non-Curriculum



- DQN exhibits a high-speed, aggressive approach in the simpler stages (C0), exploiting distance-based rewards to maximize speed.
- leads to collisions as traffic density increases.
- In more complex stages (C2 and C3), DQN's deterministic policy fails: agent continues to maximize speed, causing crashing, which results in poor success rates and reward exploitation.
- Limited discrete action space

MetaDrive Results - Non-Curriculum (PPO)

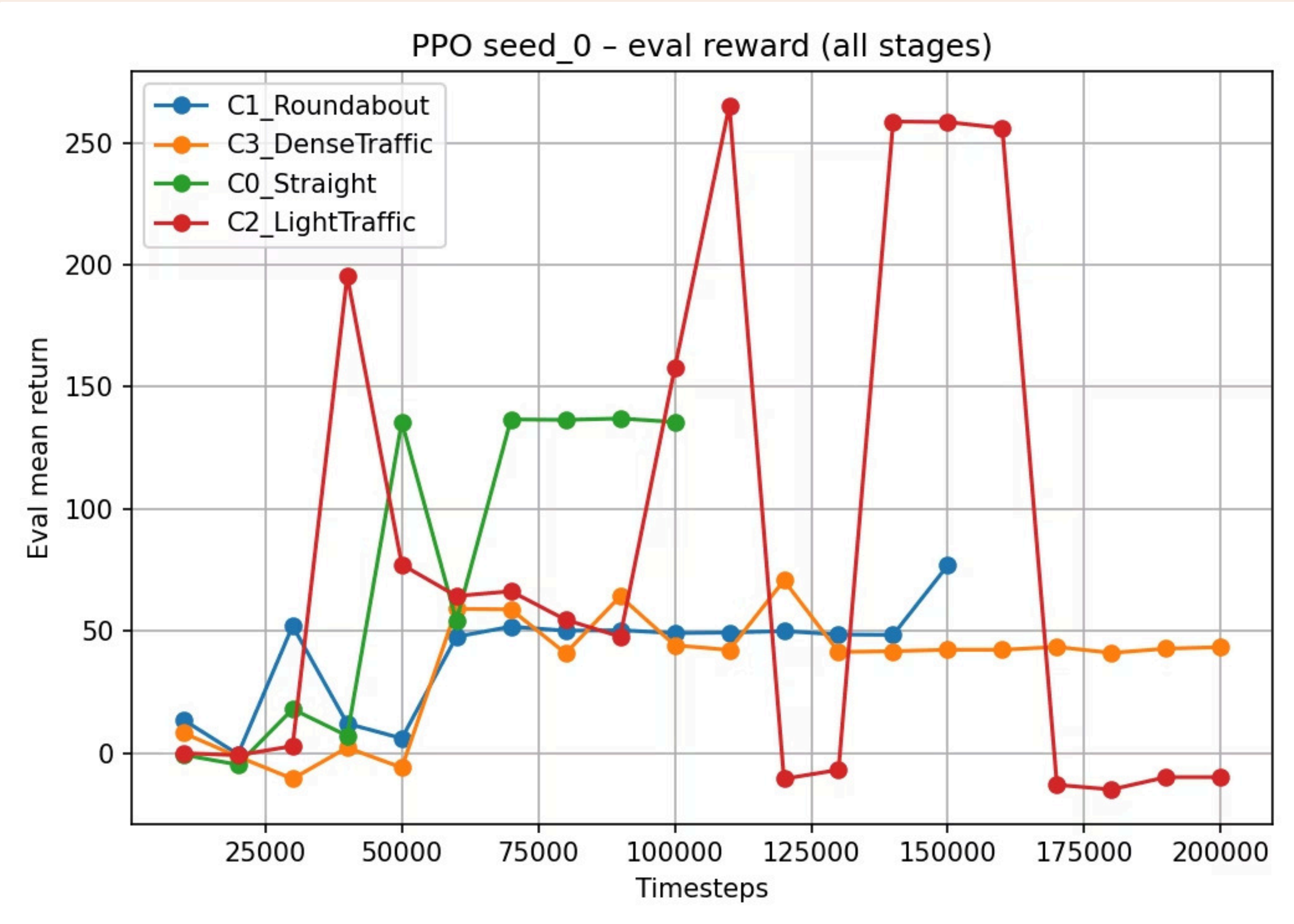


- PPO maintains a safe but conservative driving policy.
- In simpler stages, PPO performs well with relatively stable rewards.
- The agent occasionally sacrifices efficiency for safety.
- Reward fluctuations in Stage C2 suggest that PPO struggles to balance speed and safety in dense traffic
- Task difficulty increase too much - success rate drops in later stages

MetaDrive Results - Curriculum

Stages	Training Mean Reward ± Std	Held-out (SCrRXO) Reward
C0 (Straight)	135.6 ± 5.1	32.1
C1 (Roundabout)	76.8 ± 8.2	64.3
C2 (Light Traffic)	265.3 ± 12.4	-49.9
C3 (Dense Traffic)	43.3 ± 6.5	45.5

MetaDrive Results - Curriculum



- Curriculum benefits from adjusting from a easier task to a mildly harder task → success rate remained high in C0 & C1.
- High rewards during training in C2 did not translate into good **held-out performance**.
- The agent **recovered** in the final stage by leveraging the skills acquired in earlier stages.
- Due to the harsh increase in task difficulty from C1 to C2, higher episode lengths were observed in Stage C2 → needed more time.

MetaDrive Results - Curriculum vs Non-Curriculum

Metric	Non-Curriculum PPO	Curriculum PPO	Change
Mean Training Reward (C3)	41.2	43.3	+5.1%
Held-out SCrRXO Reward	43.5	45.5	+4.6%
Total Training Time	≈6 hrs	≈4.5 hrs	−25%

MetaDrive Analysis

Non-Curriculum

DQN:

- The **flatline behavior** and **zero standard deviation** across stages C1-C3 indicate reward hacking.
- **high collision rates** and **zero success rate** - converged to a deterministic strategy.
- The **discrete action space** of **DQN** made it harder for the agent to apply **fine-grained control**.

PPO:

- **PPO** converged to a **safe but slow** local optimum, showing **persistent variance** in episode lengths, suggesting more cautious driving to avoid crashes.
- While **PPO** had a **non-zero survival rate**, its **safety-first approach** resulted in **lower variance** in episode rewards and **higher survival rates** compared to **DQN**, often failing to reach the destination on time.

Curriculum (PPO)

- **Catastrophic Forgetting:** drop in held-out reward at **Stage C2**
- **Recovery and Generalization:** recovery seen in **Stage C3** demonstrates that the curriculum helped the agent re-integrate learned behaviors from simpler tasks.
- Curriculum learning in **MetaDrive** was beneficial in terms of **stabilizing training** and allowing the agent to better generalize across environments.
- Not enough to guarantee high success rates and low collision rates in complex scenarios.
- **Training Efficiency:** Curriculum PPO required **25% less training time** - indicates the **efficiency** of curriculum learning in providing **progressive exposure** to complex tasks.
- The **improvement in generalization** was moderate but noticeable, indicating the importance of structured exposure to increasing complexity.

HighwayEnv Results: Reward Comparison

Curriculum learning's impact varied significantly across algorithms in HighwayEnv.

+44.96%

PPO Mean Reward

Strongest gains with curriculum learning, also improved success and completion rates.

+11.25%

DQN Mean Reward

Moderate but consistent improvements, better completion rate but no change in success rate.

+5.71%

SimpleDQN Mean Reward

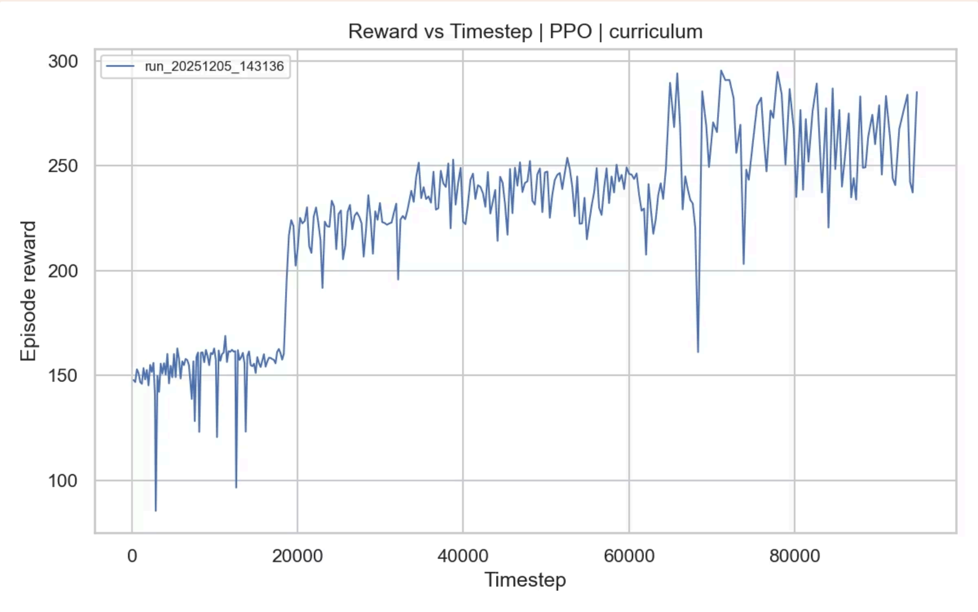
Mixed effects: increased reward and completion, but sharp drop in success rate.

HighwayEnv Results: Other Metrics

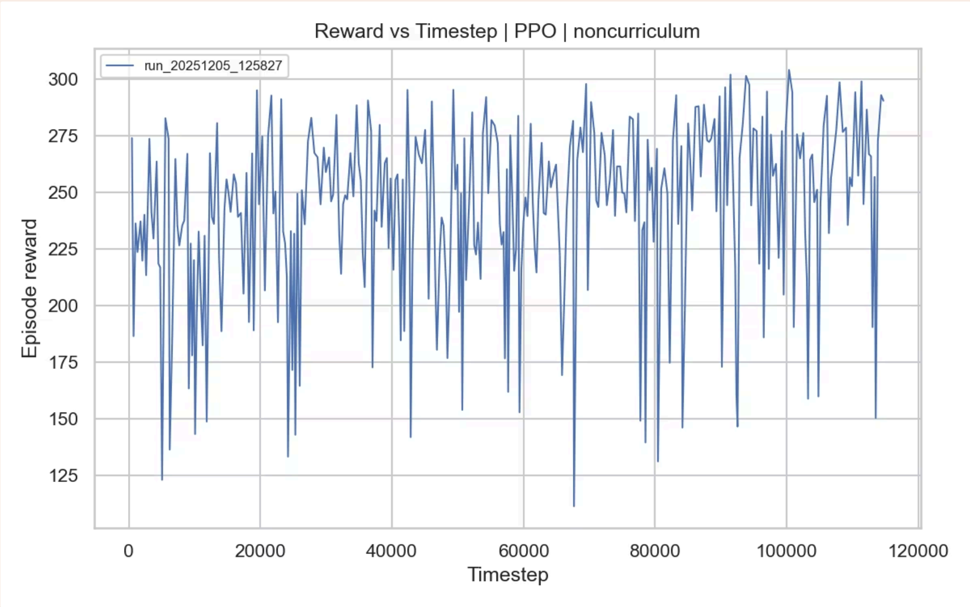
Algorithm	Δ Success Rate	Δ Completion Rate	Δ Training Time (%)	Δ Training Steps*
PPO	+21.43%	+22.00%	-28.92%	-20000
DQN	0.00%	+15.92%	-12.58%	-5000
SimpleDQN	-20.59%	+24.57%	+12.95%	0

* Note that all non-curriculum regimens trained for 100000 steps.

An Interesting Comparison



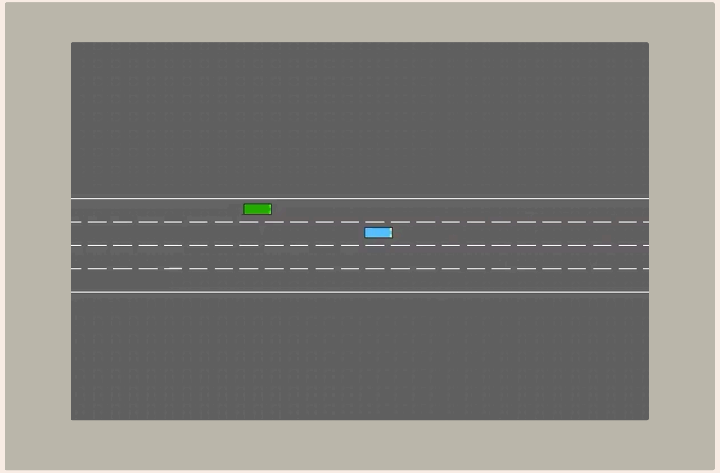
Reward vs. Timestep for PPO Curriculum Regimen

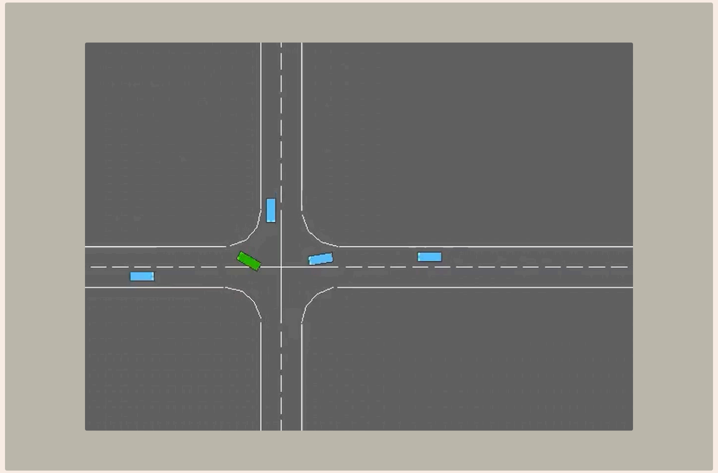


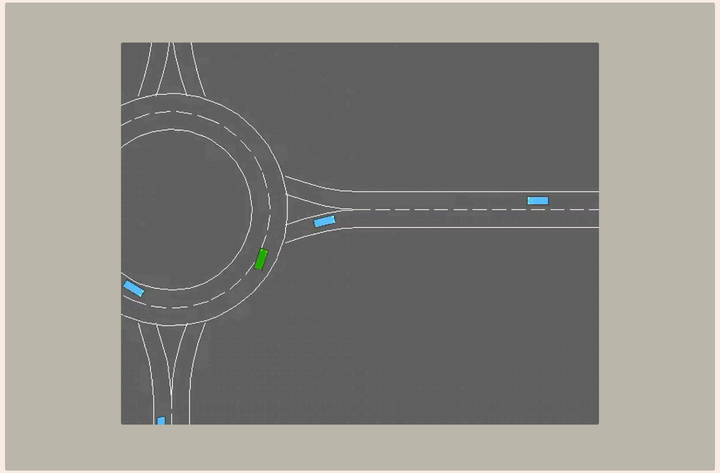
Reward vs. Timestep for PPO Non-Curriculum Regimen

- Compare timestep 80,000 onwards. Proof of learning stability.
- Greater learning stability indicates robust policy convergence for curriculum instead of oscillation for non-curriculum.

Simulations







Let's proceed!

HighwayEnv Analysis

Behavioral Dynamics

PPO

- "Throttle-as-knob" control strategy across both Curriculum and Non-Curriculum.
- Deceleration as soon as vehicle enters field of view (observation space).
- Curriculum introduces adaptive lane changing.
- Catastrophic forgetting: complex behavioral patterns exhibited on basic maps as exemplified by highway-v0.

DQN

- High speed collision-prone policies adopted early on.
- Indicative of rapid decay of epsilon-greedy exploration schedule which causes committal to suboptimal Q-values.
- Curriculum learning improves mean reward and completion rates indicating that staged exposure allows for learning of more robust Q-values in early stages.
- DQN's discrete action space makes it harder to apply fine-grained control.

HighwayEnv Analysis

Behavioral Dynamics

SimpleDQN

- Success rate decline might be because of aggressive hyperparameter configuration such as:
  - smaller buffer size
  - more frequent updates
  - higher learning rate
  - more frequent target network update.
- Simulations give a little hope as curriculum regimen shows complex policy component of adaptive overtaking.
- Possibly plagued with high speed collision-prone policy.

Conclusion: Context-Dependent Efficacy

Curriculum learning's impact is highly context-dependent, with varying effects on different algorithms and environments.

Algorithm-Specific Design

Value-based agents need careful reward shaping; policy-gradient agents benefit from smoothing mechanisms.

Catastrophic Forgetting

Simple sequential training is insufficient for robust multi-task retention, highlighting the need for improved stage transitions.

Future Work

Integrate dynamic reward shaping and explore hybrid architectures to balance safety and efficiency.

BUT, there is definitely hope!