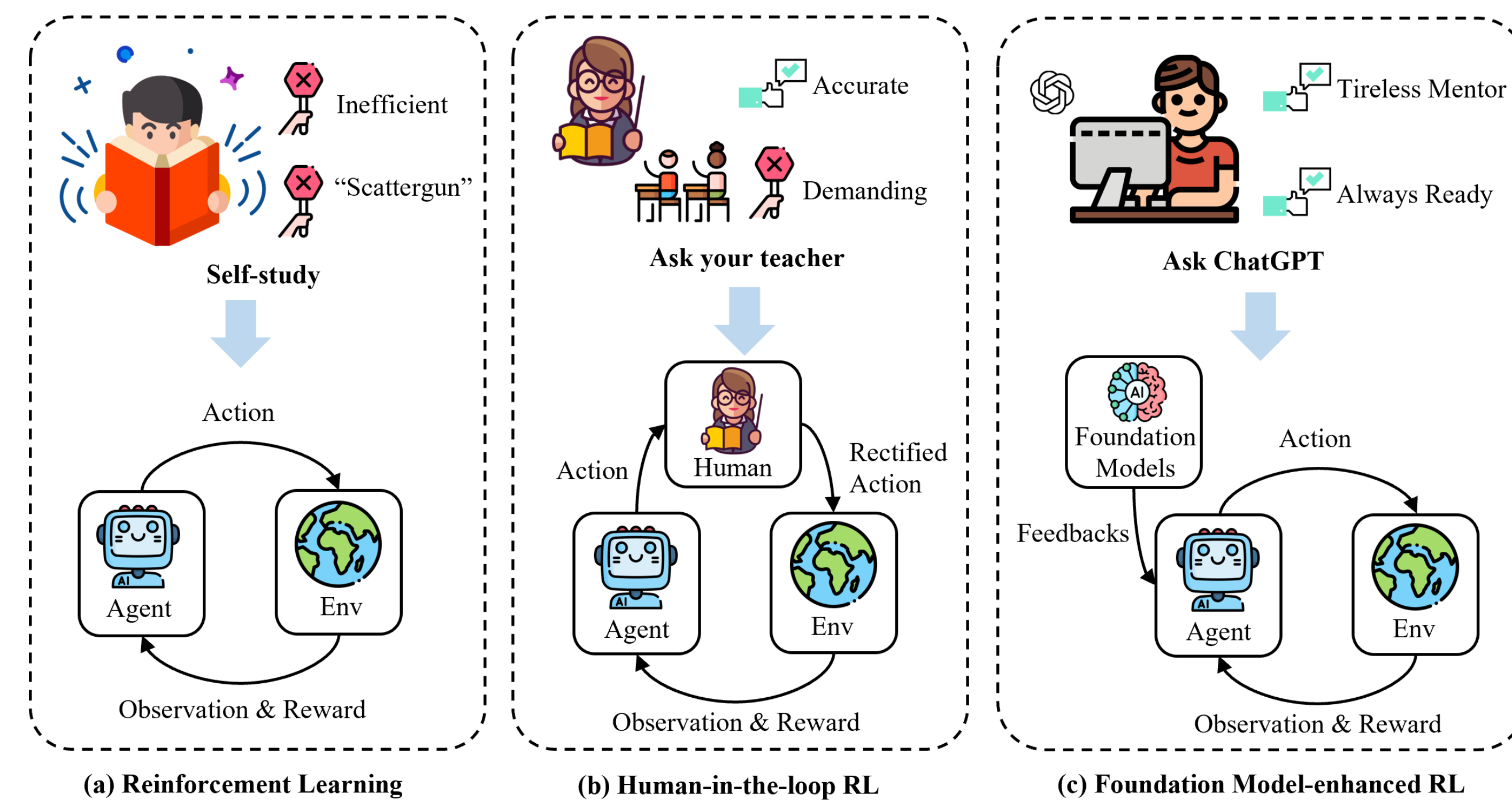
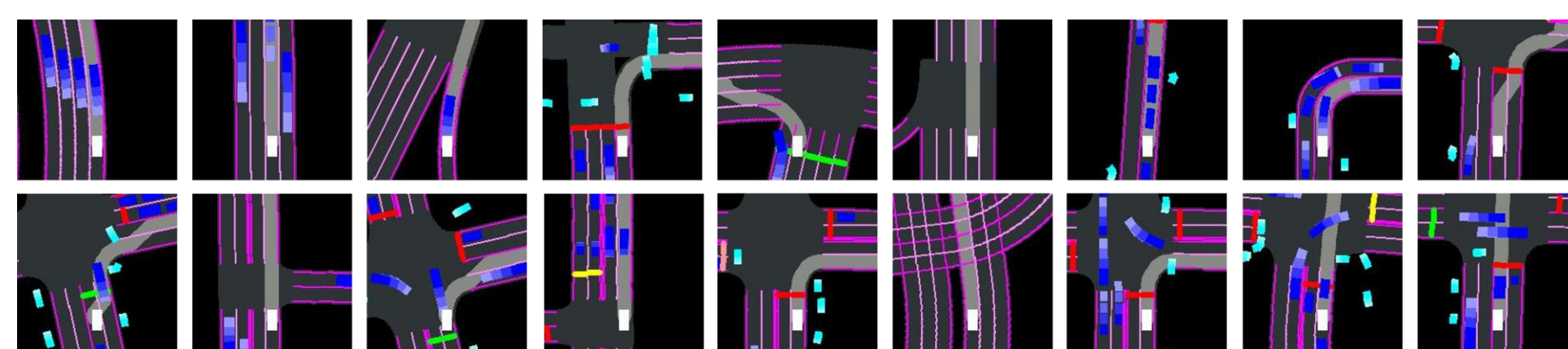


INTRODUCTION

- **Motivation: RL is fast at inference but hard to train well in complex driving scenes.** End-to-end reinforcement learning is a strong paradigm for autonomous driving, yet it typically suffers from low sample efficiency and limited semantic interpretability, especially in long-tail and safety-critical scenarios.
- **Gap: VLM supervision is powerful, but hard to use in closed-loop RL training.** Vision-language models can act like scalable “tireless mentors” by providing context-aware guidance, but their heavy inference cost creates major latency bottlenecks in high-frequency, multi-environment RL pipelines.
- **Our solution: Found-RL enables practical foundation-model-enhanced RL.** We introduce Found-RL with an asynchronous batch inference design that decouples VLM reasoning from the simulation loop, enabling real-time/near-real-time training while supporting online action guidance (VMR/AWAG) and CLIP-based dense reward shaping for better policy learning.

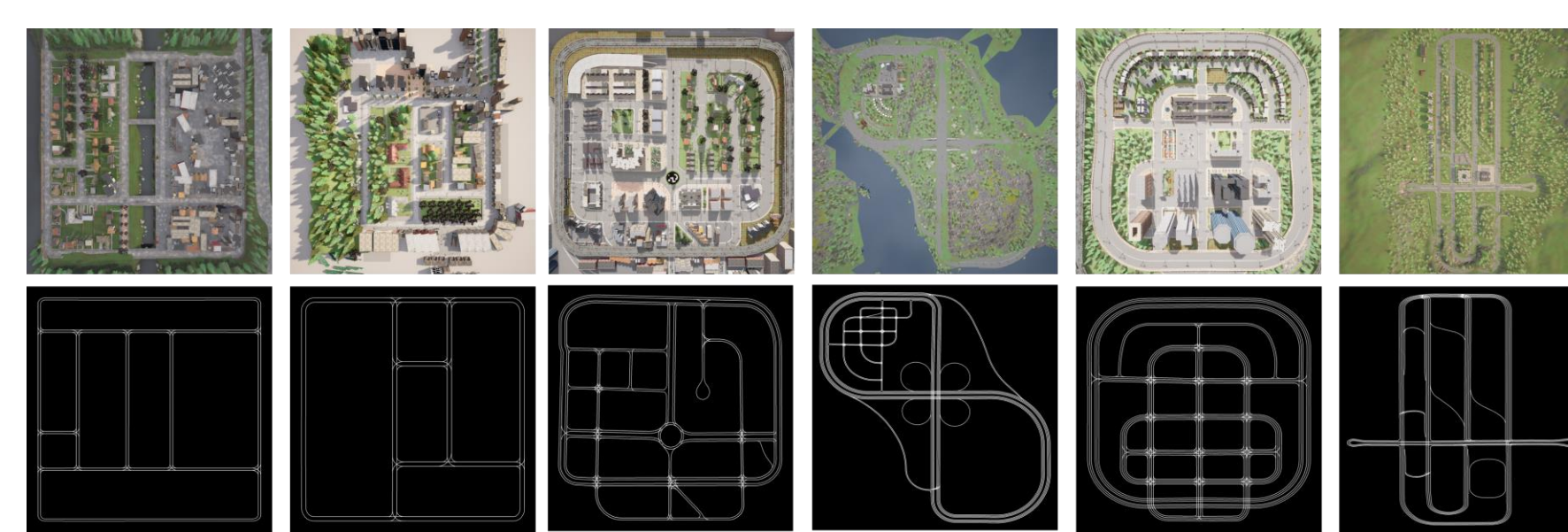


EXPERIMENT SETTING

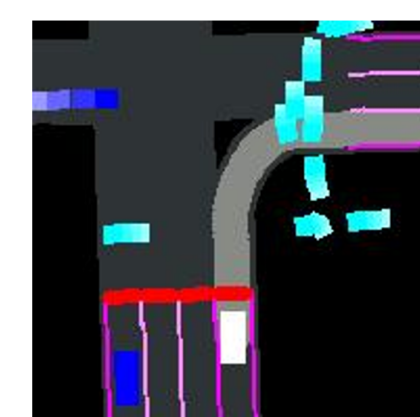


Driving scenarios.

CARLA-based comprehensive driving scenarios



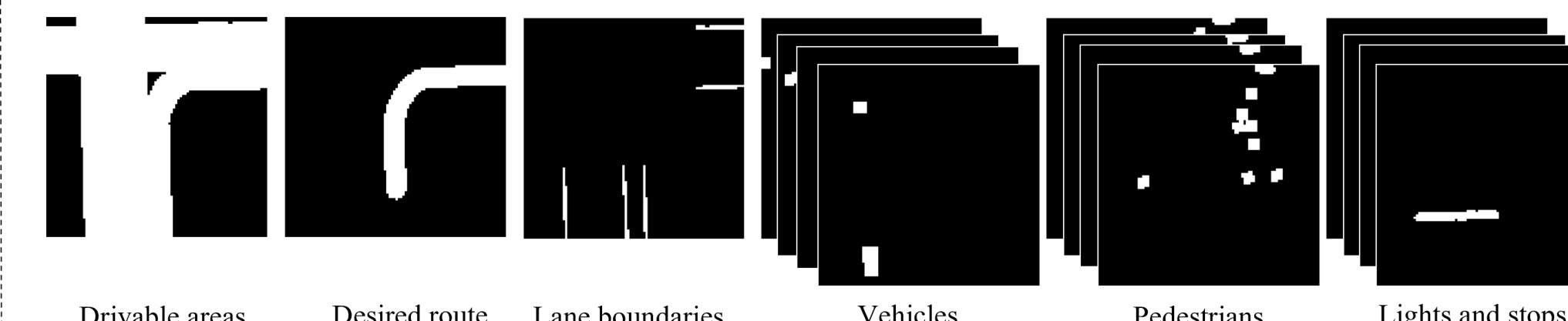
Map layout.





Question: You are a driving agent in a simulated environment.
 Current speed: 0.0 m/s.
 Forward vector: (4.51, 0.13).
 Navigation command: "follow the current lane".
 Based on the visual input and state information, predict the next action.
 Return the action in the following JSON format: {"throttle": float, "steer": float, "brake": float}


Answer: {"throttle": 1.0, "steer": 0.0, "brake": 0.0}

(a) Observation space of VLM-based driving agents



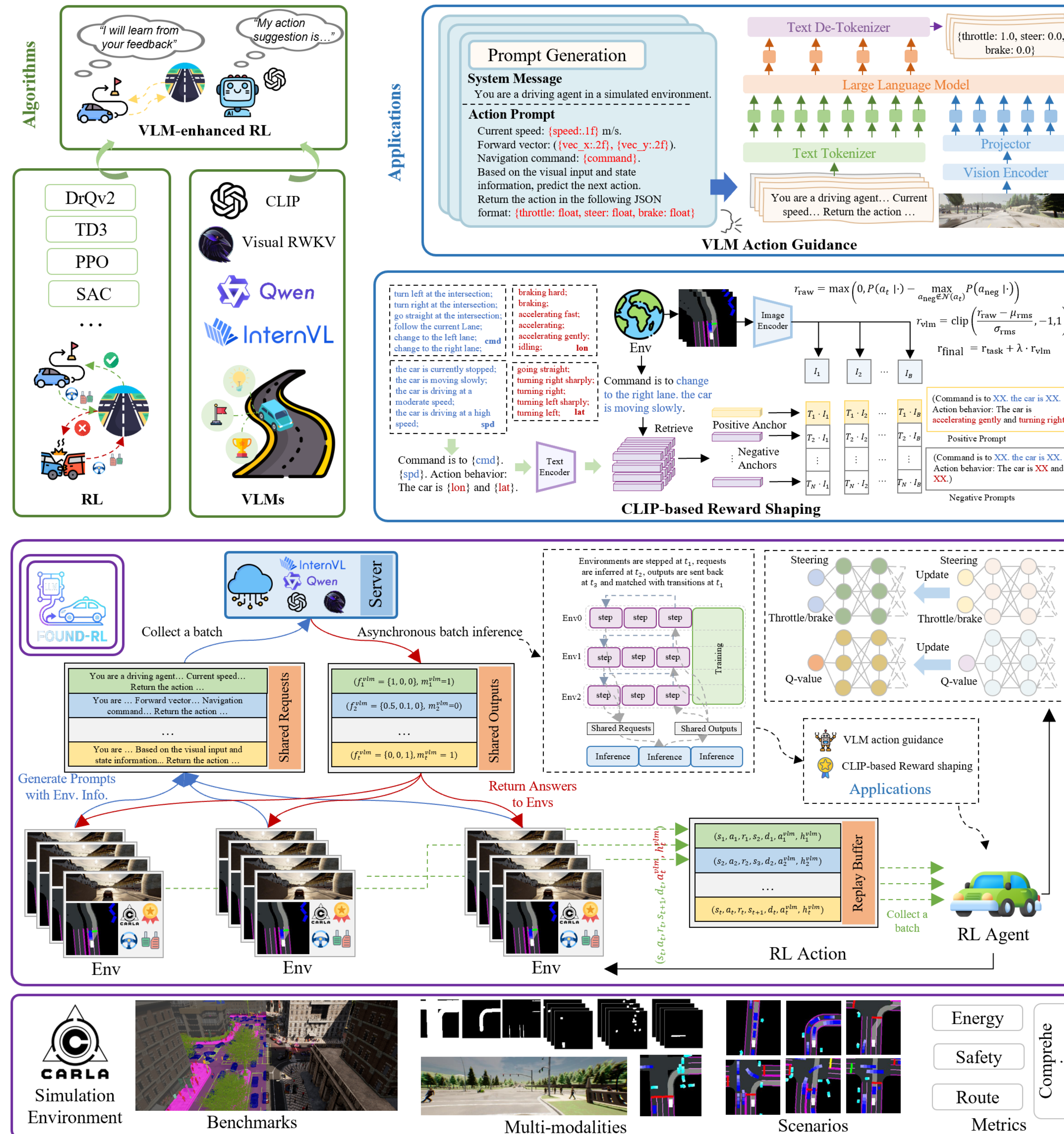
 Control context is encoded as the last-step throttle, steer, brake and gear values.

 Traffic light context is encoded as three normalized scalars: the discrete light phase (none/red/yellow/green), a clipped distance-to-red/yellow/light ratio, and the remaining time-to-phase-end ratio (set to -1 when not applicable).

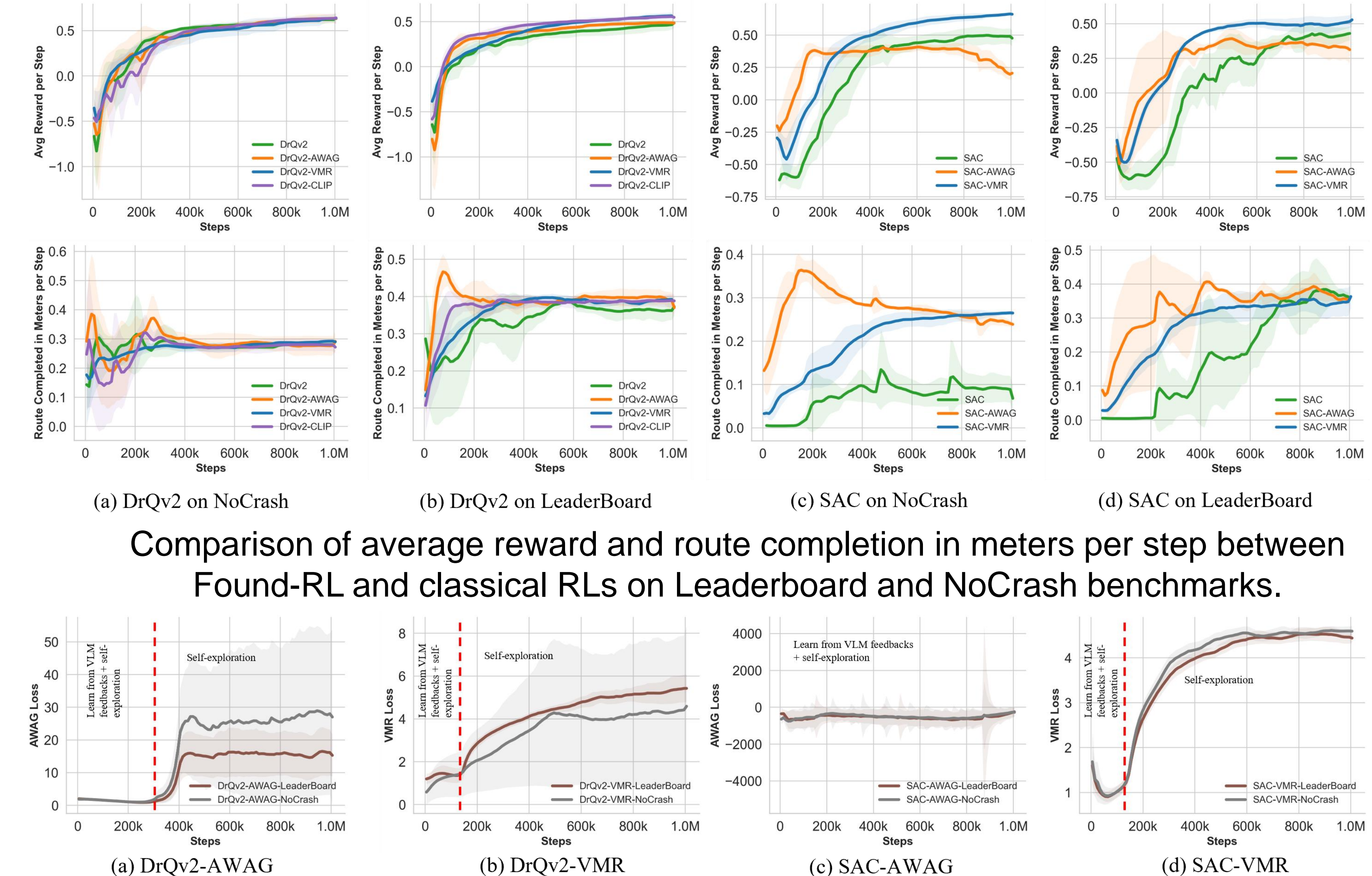
 Stop sign context is encoded as a single binary scalar that equals 1 when the ego vehicle is within a preset distance threshold of a target stop sign, and 0 otherwise.

Obs space.

METHODOLOGY (FOUND-RL)



EXPERIMENT RESULTS



Comparison of average reward and route completion in meters per step between Found-RL and classical RLs on Leaderboard and NoCrash benchmarks.

Comparison of AWAG and VMR losses between Found-RL and classical RLs on Leaderboard and NoCrash benchmarks.

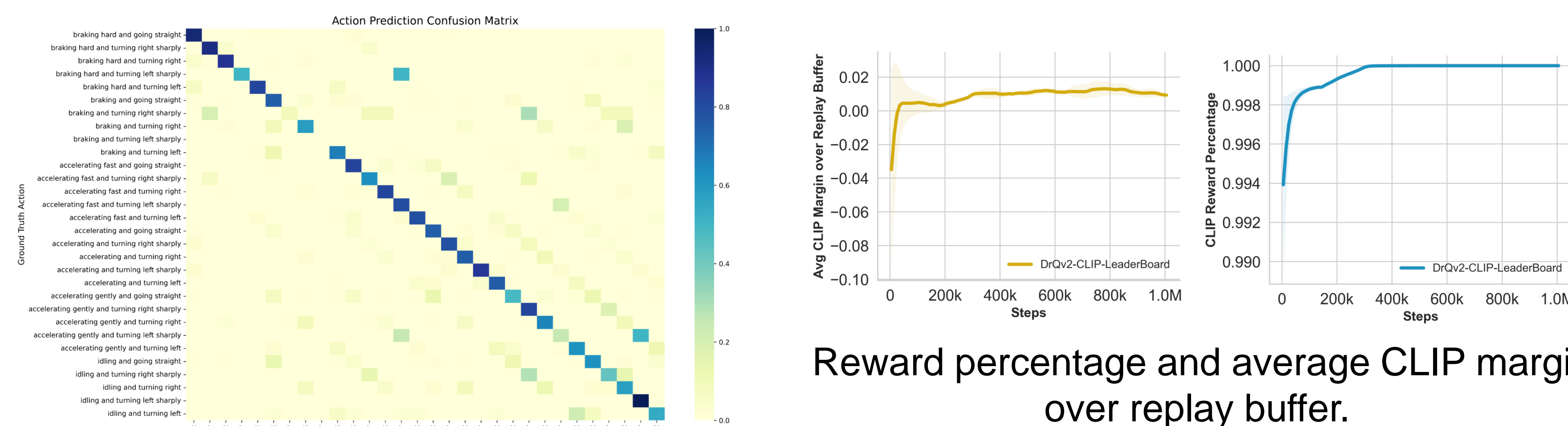
Table 1. Compare with RLs on Leaderboard benchmark

Algorithm	Comprehensive				Route		Energy			Safety	
	Return ↑	Driving Score ↑	Infra. Penalty ↓	Success Rate ↑	Route Compl. ↑	Speed ↑	Cell ↑	Fuel Rate ↑	Collisions		
									Pol. ↓	Vel. ↓	Red Light ↑
TDS	1703.83±843.48	0.53±0.24	0.89±0.03	0.37±0.30	0.28±0.26	3.48±0.17	0.18±0.01	0.004±0.000	0.00±0.00	0.21±0.07	
SAC	1270.91±102.66	0.21±0.06	0.90±0.01	0.00±0.02	0.53±0.05	3.73±0.09	0.15±0.02	0.003±0.001	0.01±0.02	0.90±0.33	
SAC+AWG	556.33±152.82	0.28±0.05	0.82±0.01	0.08±0.04	0.33±0.06	5.21±0.01	0.08±0.01	0.002±0.000	0.01±0.01	0.63±0.09	
D2Q	1621.34±263.96	0.53±0.10	0.89±0.03	0.31±0.14	0.58±0.11	3.63±0.27	0.07±0.01	0.002±0.000	0.00±0.00	0.20±0.05	
D2Q+VME	1800.36±517.97	0.56±0.11	0.88±0.01	0.38±0.13	0.61±0.12	3.72±0.20	0.09±0.03	0.002±0.001	0.01±0.01	0.19±0.05	
D2Q+AWG	1626.35±497.37	0.61±0.10	0.90±0.02	0.46±0.15	0.66±0.10	3.68±0.27	0.06±0.01	0.001±0.000	0.01±0.01	0.13±0.03	
D2Q+V2-MVR	2237.31±173.89	0.72±0.04	0.89±0.03	0.60±0.04	0.78±0.05	3.56±0.09	0.09±0.05	0.002±0.001	0.00±0.00	0.16±0.04	
D2Q+CLIP	2188.54±204.03	0.77±0.05	0.90±0.02	0.57±0.06	0.77±0.05	3.60±0.10	0.13±0.03	0.002±0.002	0.06±0.01	0.14±0.05	

Table 1. Compare with VLMs on Leaderboard benchmark

Algorithm	Model Size	Input Modality	Comprehensive				Roast		Energy		Safety		
			Return [†]	Diverging		Sec. Rate	Comp [†]	Led [†]	Fuel Rate	Fuel Vol.	Red. Vol.	Red. Lp/L [†]	
				Score [†]	Prsby [†]								
Vivid FOWB (1)	0.1B	BEV (Q2×Q2×Q3)	2044.01	0.86	0.52	0.70	3.63	0.05	0.01	0.15	0.18		
Q2×2.5- 3s	7B	BEV (Q2×Q2×Q3)	2702.49	0.76	0.31	0.65	0.80	3.50	0.04	0.01	0.10	0.09	0.05
Q2×2.5- 3s	7B	BEV (Q2×Q2×Q3)	2763.70	0.74	0.30	0.79	0.79	3.54	0.03	0.051	0.10	0.13	0.05
Roast-2.5B	7B	BEV (Q2×Q2×Q3)	2385.72	0.85	0.52	0.75	0.89	3.40	0.03	0.041	0.08	0.04	0.04
Roast-2.5B	7B	BEV (Q2×Q2×Q3)	2337.31	0.82	0.49	0.68	3.83	0.04	0.001	0.19	0.19	0.07	
DQ2×V3MR	3.82M	$\times 5\%$ × Stair (10L)	2237.31	0.72	0.89	0.62	0.78	3.56	0.09	0.002	0.16	0.16	0.02
DQ2×CLR	3.82M	$\times 5\%$ × Stair (10L)	2188.54	0.77	0.90	0.57	0.77	3.60	0.13	0.002	0.00	0.14	0.03

Compare found-rl between RLs and VLMs, **our found-rl can achieve near-VLM performance with real-time inference speed**



Reward percentage and average CLIP margin over replay buffer.



Personal Website

Action prediction confusion matrix.