Vision-Language Models-Enhanced Reinforcement Learning for Autonomous Driving

PURDUE UNIVERSITY | Department of Computer Science

Jiancong Chen, Rohit Manjunath, Saahil Mathur, Eashan Bajaj

CS 593 Reinforcement Learning Group Project, Department of Computer Science, Purdue University

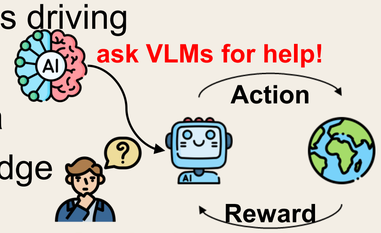
INTRODUCTION

Reinforcement Learning (RL)

- Lightweight and achieves real-time control
- Wide applications in the fields of autonomous driving

Limitation

- **Sample Efficiency**: it requires massive data
- **Learn by self-exploration**: no priori knowledge



➡ **Is there anything that has a large amount of priori knowledge, and we can easily get guidance from it? – like ChatGPT**

- **VLMs**: Vision-Language Models, it can process and understand text and visual input.
- **A pretrained / finetuned VLM** can provide multi-functional feedbacks, including expert actions, scene understanding, and safety scoring.

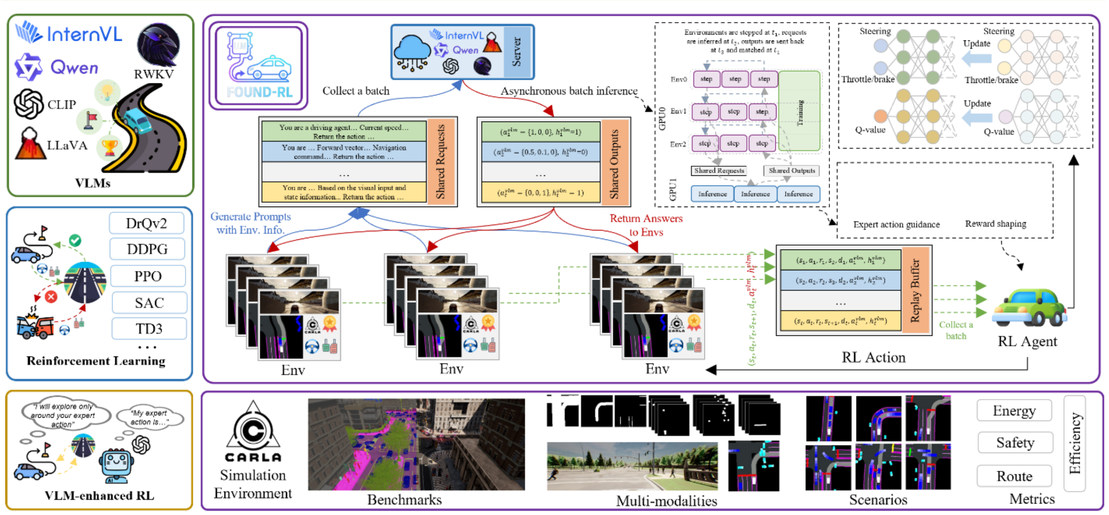
PROBLEM STATEMENT

The application of VLMs faces three questions:

(1) How to leverage VLMs in the training of RL? A general VLM-enhanced RL framework (real-time inference and on-policy decision-making)

(2) How to utilize VLM feedback for RL agents? In methodology section

(3) How to evaluate VLM guidance & Benchmarking VLMs and RLs? In results section



Overall Framework

METHODOLOGY

How to utilize VLM feedback for RL agents?

(1) Value-Margin Regularization (VMR): Assuming the guidance from VLMs is positive

On top of TD loss, add a **small extra loss** that pushes the critic to rate the VLM action slightly higher than the action stored in the replay buffer. This VMR term is weighted by a coefficient that **starts relatively large and gradually decays during training**. For critic Q_θ, the VMR loss is:

$$L_{\text{VMR}} = \mathbb{E}_D\left[m_t\left(Q_\theta(o_t, a_t^{\text{vlm}}) - \left(\text{stopgrad}(Q_\theta(o_t, a_t)) + \Delta\right)\right)^2\right] \Rightarrow L_{\text{critic}} = L_{\text{TD}} + \lambda(\text{step})L_{\text{VMR}}$$

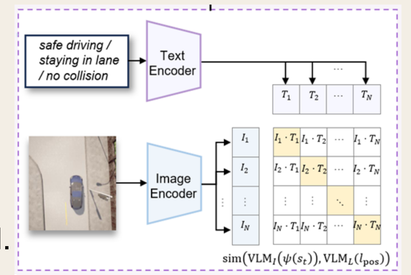
(2) Advantage-weighting Action Guidance (AWAG): A way of selectively imitating VLM actions

Compare **the critic's value of the policy's own action** with **the value of the VLM action** and turn it into a weight if the advantage is positive (higher Q). This AWAG term will be added to the normal actor loss with a decaying coefficient.

$$A_t = q_{\text{vlm}} - q_\pi \Rightarrow \mathcal{L}_{\text{AWAC}}(\phi) = -\mathbb{E}_D\left[m_t \cdot g_t \cdot \text{stopgrad}(w_t) \cdot \log \pi_\phi\left(a_t^{\text{vlm}} \mid o_t\right)\right] \Rightarrow \mathcal{L}_{\text{actor}}(\phi) = \mathcal{L}_{\text{base}}(\phi) + \iota(\text{step})\mathcal{L}_{\text{AWAC}}(\phi)$$

(3) VLM-based reward shaping: A simple reward shaping term based on CLIP

For each frame, we take the RGB image from the agent's observation and compare it with a **safety-related prompt**. CLIP gives us **a similarity score** between the image and the prompt.



$$u_t = VLM_{\text{I}}(s_t), v_t = VLM_{\text{L}}(l_{pos}) \Rightarrow c_t = \left\langle \frac{u_t}{\|u_t\|_2 + \varepsilon}, \frac{v_t}{\|v_t\|_2 + \varepsilon} \right\rangle \in [-1,1]$$

Normalize this score into a value between 0-1: **safety score**

If the safety score is high, we add **a small positive bonus** to the reward.
If the score is low, we add nothing.

$$b_t = \text{ReLU}\left(s_t^{\text{clip}} - \tau\right), \tau = 0.5, \tilde{r}_t = r_t + b_t$$

This reward shaping encourages the agent to move toward visually safer states.

RESULTS

Evaluate VLM guidance for RL agents

Compared to online-RL, VLM-enhanced RL achieves lower energy consumption and higher route completion.

Evaluate Benchmarking RLs and VLMs

After introduced VLM guidance, the gains between pure RL and VLM agents (DrQv2) have significantly narrowed.

➡ **Proposed approach is effective**

Experiment Setting

Simulation platform: CARLA simulator

Simulation tasks:

Four lanes, Left turn, Navigation, Right turn

		Energy	Safety		Comprehensive & Route	Efficiency				
	Algo	Icell	Coll. of ped.	Col. of veh.	Driving score	Route compl.	Acc.	Speed	Vehicle blocked	Success rate
Online RL	SAC	0.133	0.006	17.112	0.143	0.168	3.935	3.755	3.135	0.000
	DrQv2	0.121	0.000	1.115	0.573	0.626	3.186	3.936	0.031	0.167
VLM-enhaced RL	SAC-AWAG	0.087	0.008	1.290	0.219	0.261	1.907	5.226	0.047	0.006
	SAC-VMR	0.076	0.000	2.515	0.542	0.600	2.094	3.864	0.000	0.036
	DrQv2-CLIP	0.071	0.000	0.827	0.664	0.725	1.998	3.968	0.036	0.095
	DrQv2-VMR	0.075	0.000	12.524	0.776	0.842	2.160	3.459	0.197	0.143
	DrQv2-AWAG	0.049	0.000	13.112	0.733	0.769	1.550	3.405	0.075	0.250