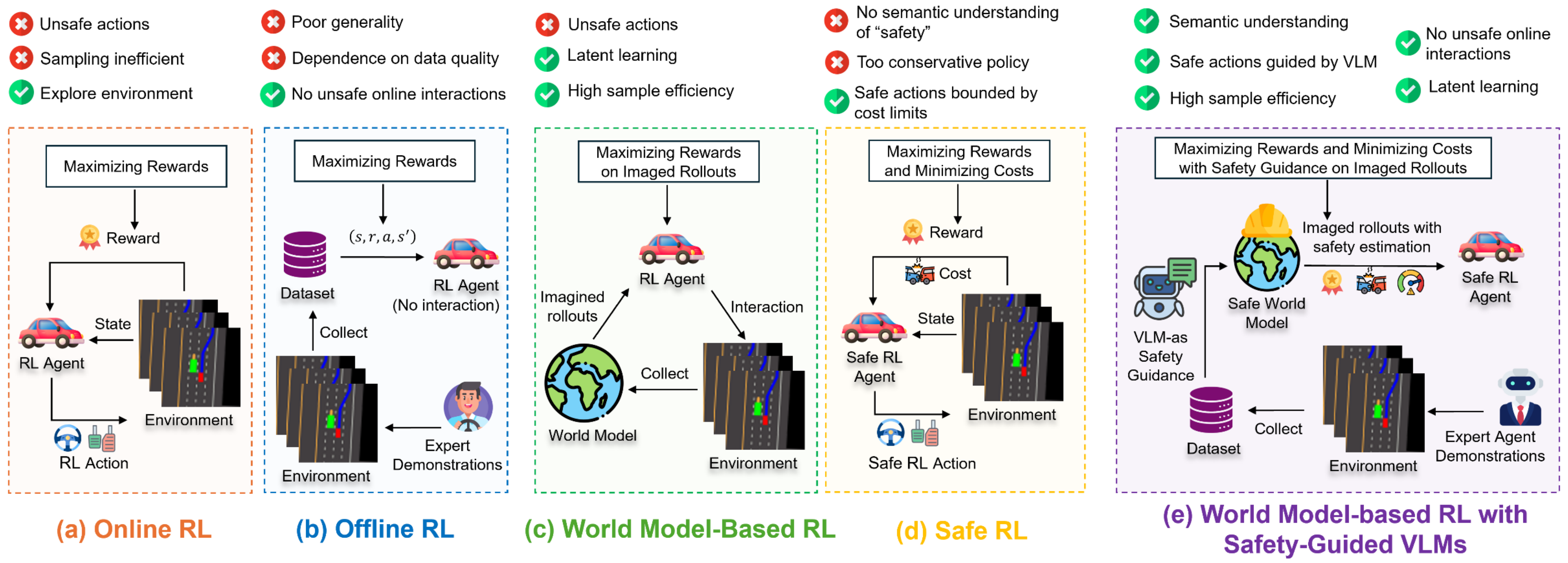
VL-SAFE: Vision-Language Guided Safety-Aware Reinforcement Learning with World Models for Autonomous Driving

Yansong Qu†, Zilin Huang†, Zihao Sheng†, Jiancong Chen, Samuel Labi, Sikai Chen*

Purdue University-West Lafayette, University of Wisconsin-Madison

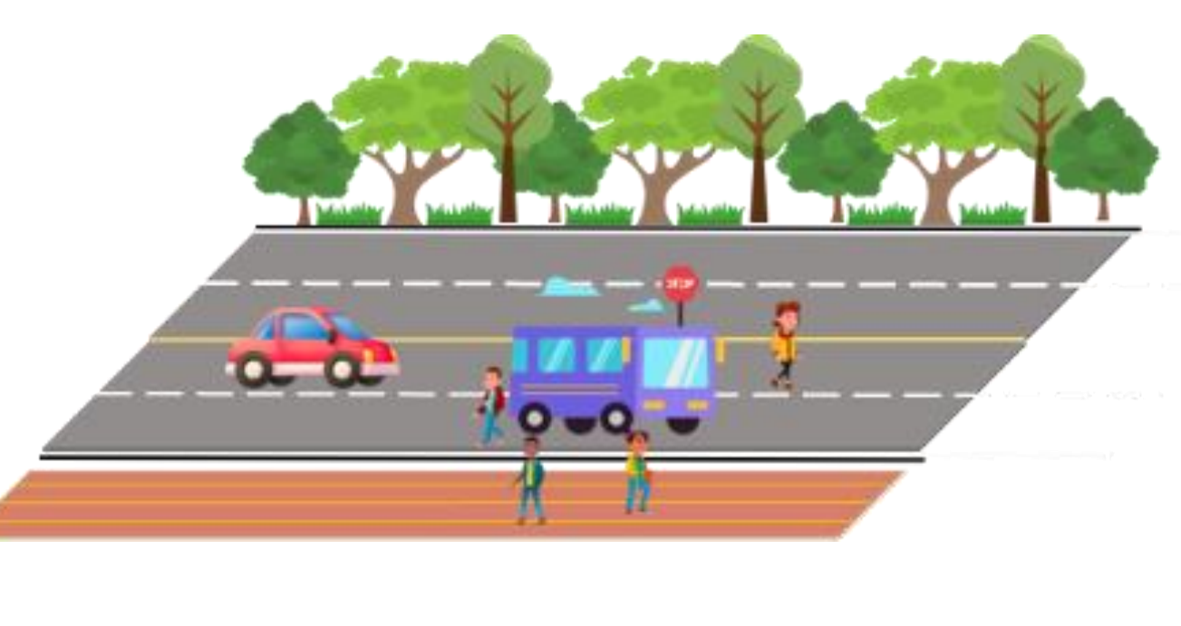
Introduction

- To make the Reinforcement learning (RL)-based autonomous driving safer, we combine world model and safe RL under offline setting to complement each other's strengths and weaknesses.
- More importantly, to achieve a safer RL, a fundamental question is: how can we identify risky states, semantically understand “safety”, and guide policy learning accordingly?
- We propose a VLM-as-safety-guidance paradigm to provide intuitive and generalizable safety signals. These signals can guide the RL agent to determine when it is appropriate to maximize rewards and when to minimize, leading to more balanced and safe policy learning.



(a) Online RL (b) Offline RL (c) World Model-Based RL (d) Safe RL (e) World Model-based RL with Safety-Guided VLMs

Motivation

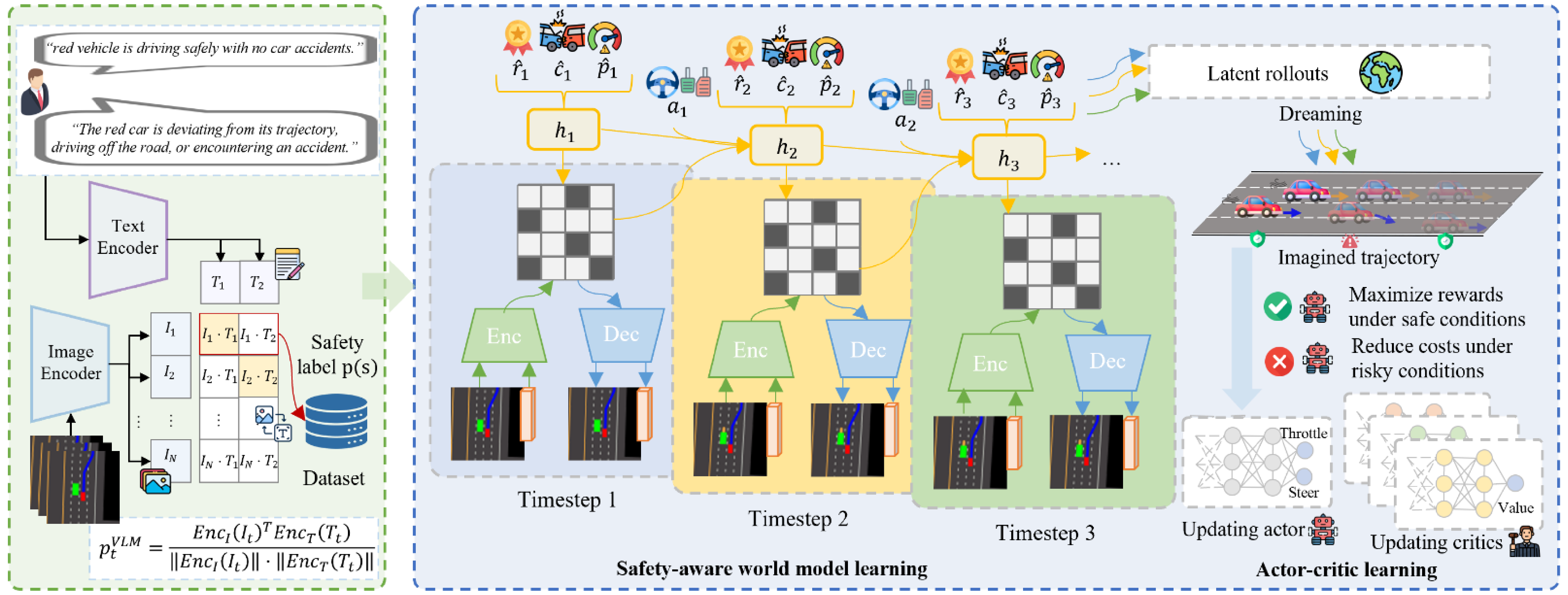


An example from life: Imagine an autonomous vehicle approaching a stopped school bus on a multi-lane road, where several children are walking across the street toward the sidewalk. The school bus has its STOP sign extended, indicating that all surrounding vehicles must come to a complete stop.

1. A classical RL-based or offline-trained agent may continue driving, as no collision has occurred and the reward function still favors progress.
2. A world model-based agent may also ignore STOP sign during imagined rollouts, failing to distinguish the risk embedded in this visual context.
3. A traditional safe RL policy may behave unpredictable, it may stop in some cases, but often fails to recognize the raised STOP sign and the presence of children, especially if such visual cues were not explicitly encoded in the cost function.
4. In contrast, a VLM-guided policy can semantically understand the scene: the raised STOP sign on the school bus and the crossing children clearly indicate a high-risk situation. Even in the absence of a collision, the VLM assigns a low safety score based on visual semantics, guiding the agent to stop proactively. It enables agents to make early, context-aware, and proportionate decisions, ultimately allowing them to act safely across varied scenarios

Methodology (VL-SAFE)

Learning Optimal Policy: $\pi^*(\boldsymbol{a} \mid \boldsymbol{s}) = \frac{1}{Z} w \cdot \pi_b(\boldsymbol{a} \mid \boldsymbol{s})$ $\quad w = p(s) \cdot \exp(\beta_1 A^r(\boldsymbol{s}, \boldsymbol{a})) + (1 - p(s)) \cdot \exp(-\beta_2 A^c(\boldsymbol{s}, \boldsymbol{a}))$


“red vehicle is driving safely with no car accidents.”
“The red car is deviating from its trajectory, driving off the road, or encountering an accident.”
Text Encoder
Image Encoder
T_1
T_2
I_1
I_2
I_N
$I_1 \cdot T_1$
$I_1 \cdot T_2$
$I_2 \cdot T_1$
$I_2 \cdot T_2$
$I_N \cdot T_1$
$I_N \cdot T_2$
Safety label p(s)
Dataset
$p_t^{VLM} = \frac{Enc_I(I_t)^T Enc_T(T_t)}{\|Enc_I(I_t)\| \cdot \|Enc_T(T_t)\|}$
$\hat{r}_1$ $\hat{c}_1$ $\hat{p}_1$
a_1
$\hat{r}_2$ $\hat{c}_2$ $\hat{p}_2$
a_2
$\hat{r}_3$ $\hat{c}_3$ $\hat{p}_3$
h_1
h_2
h_3
…
Enc
Dec
Timestep 1
Timestep 2
Timestep 3
Safety-aware world model learning
Latent rollouts
Dreaming
Imagined trajectory
Maximize rewards under safe conditions
Reduce costs under risky conditions
Throttle
Steer
Value
Updating actor
Updating critics
Actor-critic learning

(a) VLM as safety guidance

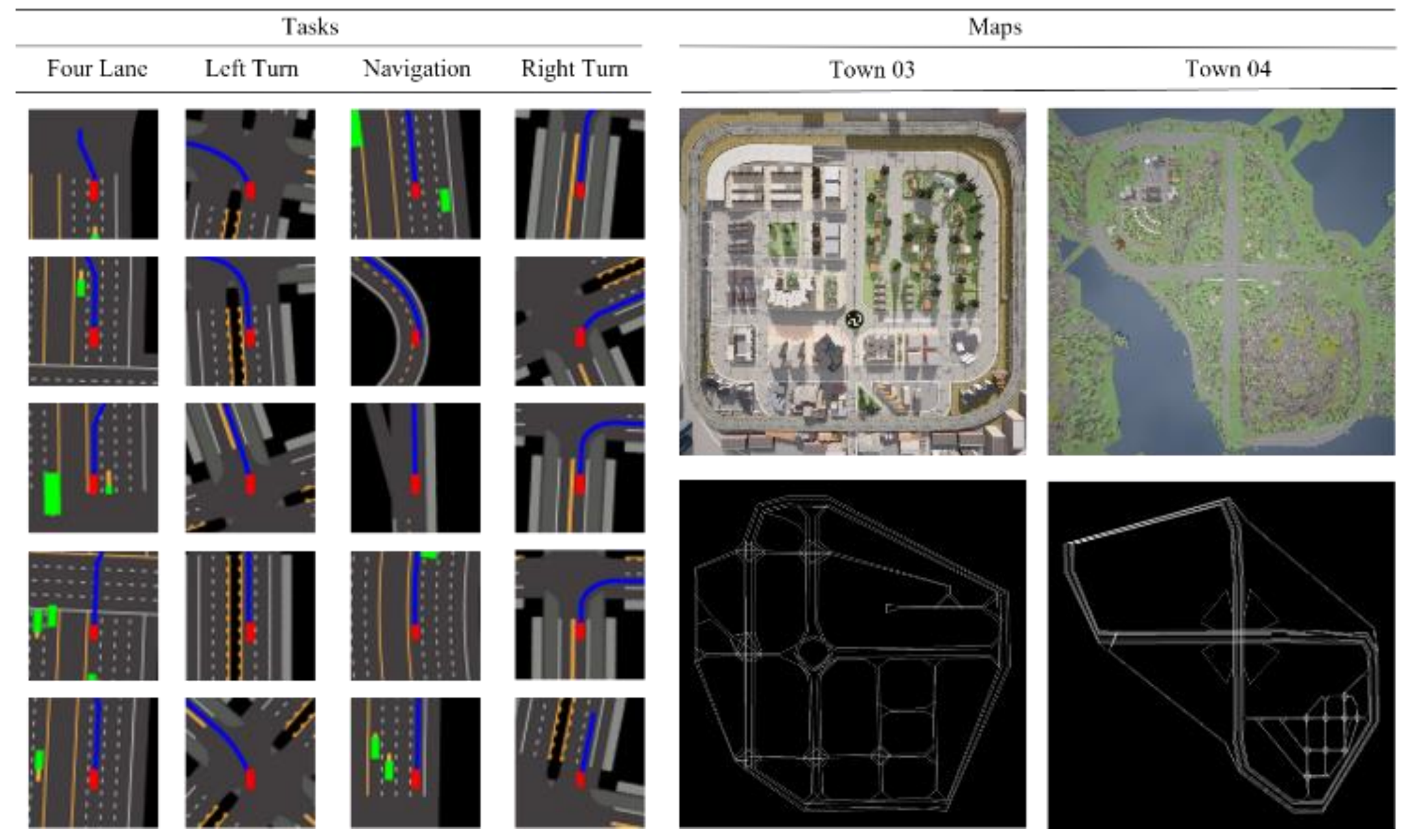
(b) World model-based safe reinforcement learning

Details:

1. The first phase generates ground truth safety estimations using CLIPs for each state in offline dataset collected by expert agent.
2. The second stage will learn a safety-aware world model to generate imagined rollouts along with predicted safety estimations for actor-critic learning.
3. Achieves a balance between reward maximization and cost minimization under the VLM-based safety guidance.
4. Provides a semantic understanding of "safety" in complex driving contexts, enabling an effective safe policy learning.

.

Experiment Results



Driving scenarios: Four tasks in the CarDreamer simulation platform on Town 03 and Town 04 maps under CARLA.

Observation space: BEV image.

Action space: discrete acceleration = [-2.0, 0.0, 2.0] and discrete steering = [-0.6, -0.2, 0.0, 0.2, 0.6].

Evaluation metrics: Optimization goals: Return, Cost; mobility: Average speed (AS), Waypoint completion (WC), travel distance (TD); safety: Arrive rate (AR), Collision rate (CR), time-based collision frequency (TCF), distance-based collision frequency (DCF), Collision speed (CS)

Compared to traditional methods, our framework yields superior performance in terms of driving safety, sampling efficiency, and generalizability.

TABLE I
Comparison of Baselines and VL-SAFE under CarDreamer

Four Lane

Models	Return ↑	Cost ↓	AS ↑	WC ↑	TD ↑	AR ↑	CR ↓	TCF ↓	DCF ↓	CS ↓
PPO	356.48	79.32	2.15	71.5	75.64	0.24	0.09	0.2	1.1	3.04
SAC	96.73	28.35	0.36	20.18	17.9	0	0	0	0	0
DQN	612.33	57.79	4.69	151.54	149.96	0.81	0.11	0.33	0.73	4.25
BC	-85.39	156.46	8.76	112.35	137.13	0.59	0.2	1.29	1.48	9.51
CQL	110.42	101	5.81	104.15	120.55	0.54	0.13	0.45	1.1	7
IQL	-43.38	67.35	0.08	2.55	3.41	0	0.01	0.01	0	1.22
DreamerV3	326.95	71.81	6.3	135.6	149.49	0.67	0.2	0.89	1.46	8.37
FSOP	429.18	73.15	5.6	137.7	144.58	0.52	0.15	0.25	0.92	5.6
VL-SAFE	552.43	69.17	4.25	146.7	132.55	0.7	0.05	0.34	0.46	5.22

Left Turn

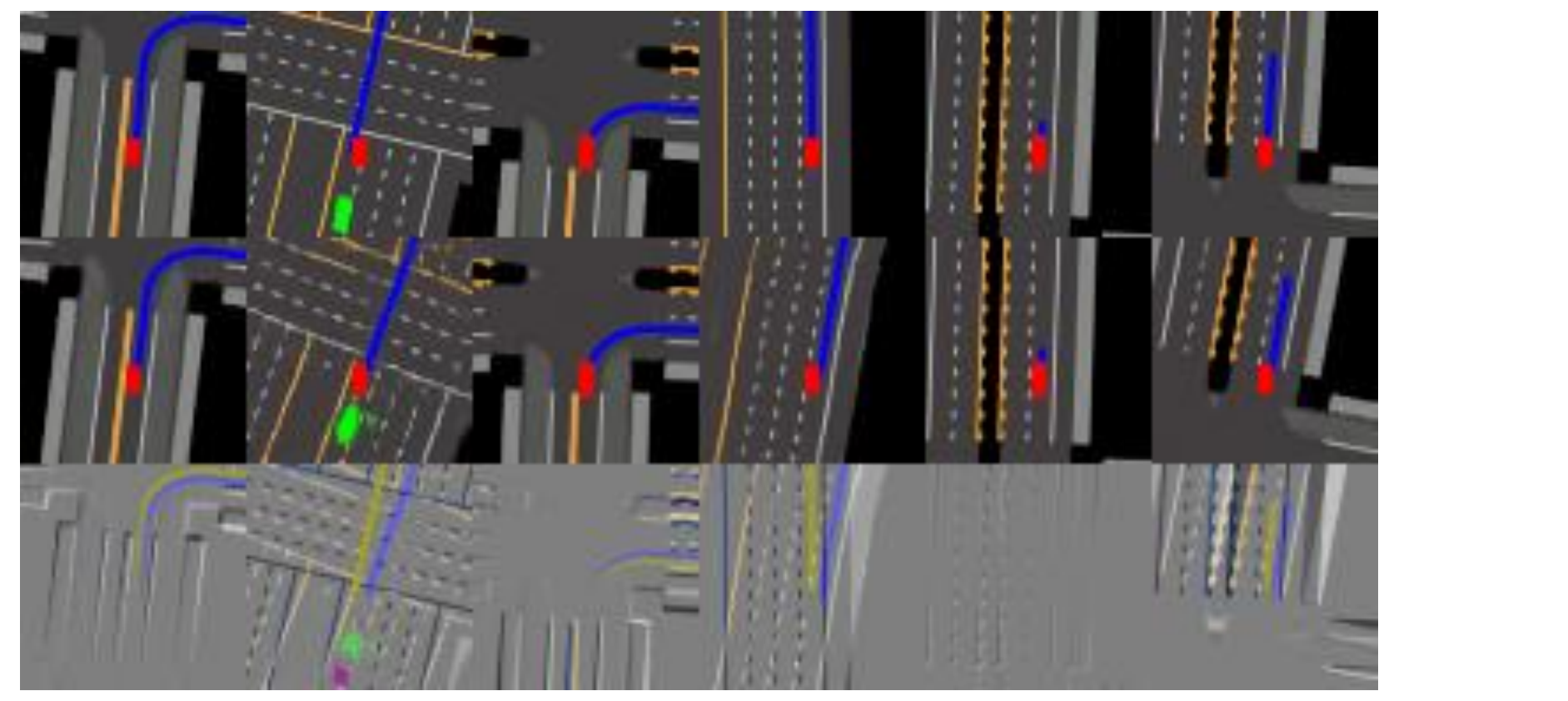
Models	Return ↑	Cost ↓	AS ↑	WC ↑	TD ↑	AR ↑	CR ↓	TCF ↓	DCF ↓	CS ↓
PPO	208.75	35.59	1.63	37.94	34.83	0	0	0	0	0
SAC	110.2	0.28	0.31	18.86	15.5	0	0	0	0	0
DQN	214.11	44.18	3.23	45.88	61.17	0.9	0	0	0	0
BC	112.56	8.04	6.88	57.01	61.67	0.97	0	0	0	0
CQL	129.96	3.53	4.48	47.88	47.02	0.67	0	0	0	0
IQL	-1.47	0	0	0	0	0	0	0	0	0
DreamerV3	305.42	17.63	4.14	63	61.54	1	0	0	0	0
FSOP	281.87	0.22	5.12	70.9	61	1	0	0	0	0
VL-SAFE	358.7	1.68	2.18	71.5	60.63	0.94	0	0	0	0

Navigation

Models	Return ↑	Cost ↓	AS ↑	WC ↑	TD ↑	AR ↑	CR ↓	TCF ↓	DCF ↓	CS ↓
PPO	434.51	68.5	1.62	82.95	70.13	0	0.02	0.04	0.14	0.18
SAC	-11.03	12.38	0	0.13	0.14	0	0	0	0	0
DQN	344.74	40.93	4.47	163.59	156.4	0	0.07	0.16	0.41	0.97
BC	-90.54	124.89	8.84	262.16	302.86	0	0.16	0.46	0.68	5.27
CQL	231.76	110.19	6.01	235.41	239.74	0	0.13	0.29	0.57	5.12
IQL	-177.51	217.15	0.25	3.54	5.36	0	0.02	0.03	3.67	0.03
DreamerV3	528.86	48.41	7.08	297.55	321.43	0	0.1	0.39	0.36	3.04
FSOP	775.17	54.4	6.73	359	305.04	0	0.08	0.37	0.61	3.96
VL-SAFE	935.63	0.55	3.57	246.6	178.62	0	0.01	0.04	0	0

Right Turn

Models	Return ↑	Cost ↓	AS ↑	WC ↑	TD ↑	AR ↑	CR ↓	TCF ↓	DCF ↓	CS ↓
PPO	288.4	10.59	1.96	52.58	42.36	0.65	0	0	0	0
SAC	-8.32	96.57	0.31	14.51	13.97	0	0	0	0	0
DQN	46.04	36.46	0.49	10.96	15.11	0.14	0	0	0	0
BC	107.67	7.76	6.28	47.62	45.26	0.83	0	0	0	0
CQL	67.12	77.04	4.91	44.12	45.35	0.77	0	0	0	0
IQL	-3.88	31.43	0.14	3.78	4.11	0	0	0	0	0
DreamerV3	318.44	2.67	3.5	59.5	48.37	0.85	0	0	0	0
FSOP	258.61	0.72	4.49	60.9	48.01	0.93	0	0	0	0
VL-SAFE	315.17	0	2.53	58.1	43.5	0.89	0	0	0	0



Comparison between ground-truth observations and predictions generated by the world model.

LinkedIn

SCAN ME

Paper Video and Code

SCAN ME