Physics-Informed Reward Framework for Vision-Language Driven Safe Autonomous Driving

Supplementary Material

A. Analysis of CBF

Lemma 1 (Continuous Differentiability of CBF) *Given a CBF h, the function $t \mapsto h(\bar{x}_{\mathcal{N}(t)}(t))$ remains continuously differentiable, even though $\bar{x}_{\mathcal{N}(t)}(t)$ has discontinuities whenever the agent set $\mathcal{N}(t)$ changes.*

Proof 1 *Let $\tilde{h} : R \to R$ be defined as:*

$$\tilde{h}(t) \coloneqq h(\bar{x}_{\mathcal{N}(t)}(t)). \tag{1}$$

We consider two cases: when the agent set $\mathcal{N}(t)$ remains unchanged and when it changes due to the addition or removal of agents. First, assume that at time $t = t_0$, the agent set $\mathcal{N}(t)$ remains unchanged, i.e., $\lim_{t\uparrow t_0} \mathcal{N}(t) = \lim_{t\downarrow t_0} \mathcal{N}(t) = \mathcal{N}(t_0)$. Given that both h and $\bar{x}_{\mathcal{N}}$ are continuously differentiable at t_0, it follows that $\tilde{h}$ is also continuously differentiable at t_0. Next, we consider the case when $\mathcal{N}(t)$ changes at time t_0. For clarity, we denote the set $\mathcal{N}(t)$ immediately before and after the change as $\mathcal{N}^-$ and $\mathcal{N}^+$, respectively. By applying Eq. 12b of the main paper on $t < t_0$ and $t \geq t_0$ we obtain the corresponding one-sided limits:

$$\lim_{t\uparrow t_0} \frac{h(\bar{x}_{\mathcal{N}^-}(t)) - h(\bar{x}_{\mathcal{N}^+}(t_0))}{t - t_0} = 0, \tag{2}$$

and

$$\lim_{t\downarrow t_0} \frac{h(\bar{x}_{\mathcal{N}^+}(t)) - h(\bar{x}_{\mathcal{N}^+}(t_0))}{t - t_0} = 0. \tag{3}$$

Together with the condition in Eq. 12a of the main paper, these limits ensure the existence and continuity of the derivative of $\tilde{h}$ at t_0. Therefore, $\tilde{h}$ is continuously differentiable for all t.

Theorem 1 (Safety Forward Invariance of CBF) *Let h be a CBF, and define its 0-superlevel set as $\mathcal{C} \subset \mathcal{X}$, i.e., $\mathcal{C} = \{x \in \mathcal{X} : h(x) \geq 0\}$. Suppose $\mathcal{C}$ is contained within a predefined safe region $\mathcal{S} \subset \mathcal{X}$, and the initial state satisfies $\bar{x}(0) \in \mathcal{C}$. Then, under any locally Lipschitz continuous control input $u : \mathcal{X} \to \mathcal{U}_{\text{safe}}$, where*

$$\mathcal{U}_{\text{safe}} \coloneqq \left\{ u \in \mathcal{U} \;\middle|\; \dot{h}(\bar{x}_{\mathcal{N}}) + \alpha(h(\bar{x}_{\mathcal{N}})) \geq 0 \right\}, \tag{4}$$

the resulting system trajectories satisfy $\bar{x}(t) \in \mathcal{S}$ for all $t \geq 0$.

Proof 2 *Define t_k for $k \in N$ with $t_0 = 0$, such that $\mathcal{N}$ remains constant on each time segment $[t_k, t_{k+1})$. For any time segment $[t_k, t_{k+1})$, suppose $h(\bar{x}_{\mathcal{N}(t_k)}(t_k)) \geq 0$. Applying Eq. 13c of the main paper to the function $t \mapsto h(\bar{x}_{\mathcal{N}(t)}(t))$ yields:*

$$h(\bar{x}_{\mathcal{N}(t)}(t)) \geq 0, \quad \forall t \in [t_k, t_{k+1}). \tag{5}$$

Consequently, we have

$$\bar{x}(t_k) \in \mathcal{C} \subset \mathcal{S} \implies \bar{x}(t) \in \mathcal{C} \subset \mathcal{S}, \quad \forall t \in [t_k, t_{k+1}). \tag{6}$$

From Lemma 1, we obtain that the function $t \mapsto h(\bar{x}_{\mathcal{N}(t)}(t))$ is continuously differentiable, indicating the existence of a right limit at t_{k+1}. Applying this limit to Eq. 5, we have $h(\bar{x}_{\mathcal{N}(t_{k+1})}(t_{k+1})) \geq 0$. Since $h(\bar{x}_{\mathcal{N}(0)}(0)) \geq 0$ by assumption, an inductive argument over the time segment $[t_k, t_{k+1})$ implies that $h(\bar{x}_{\mathcal{N}(t)}(t)) \geq 0$ for all $t \geq 0$. Therefore, the system states satisfy $\bar{x} \in \mathcal{C} \subset \mathcal{S}$ for all $t \geq 0$.

Figure 1. Bird's eye view of Towns and their drivable routes in the CARLA simulator.

Figure 2. A driving scenario and its corresponding BEV image in the CARLA simulator.

B. Environment Setup

Driving Scenarios: We conduct experiments across multiple CARLA town maps with varying sizes and structural complexities to simulate diverse driving environments. As illustrated in Fig. 1, these maps encompass T-intersections, highways, roundabouts, tunnels, and other complex scenarios that require agents to master both fundamental driving skills and high-level decision-making. Unless otherwise stated, all training and evaluation are performed in Town a, a relatively compact map. To emulate realistic and challenging traffic conditions, we introduce background vehicles operating in autopilot mode-20 vehicles for regular traffic and 40 for dense traffic-thereby introducing dynamic interactions that compel reinforcement learning agents to develop robust and adaptive driving strategies. Fig. 2 presents an example driving scenario along with its corresponding bird's-eye view (BEV) representation in CARLA.

Navigation Routes: Following the protocol of VLM-RL [20], we initialize 101 predefined spawning points and randomly select a start-end pair to define each navigation task. The shortest path between them is computed using the A* search algorithm and serves as the target trajectory. To improve training continuity, episodes are not terminated upon reaching an endpoint; instead, a new endpoint is randomly selected, and the process is repeated until the vehicle traverses a total distance of 3000 meters. During evaluation, however, each episode concludes upon reaching the assigned endpoint to ensure consistency across the test scenarios.

Evaluation Metrics: For driving efficiency assessment, we adopt average speed (AS), route completion (RC), and total traveled distance (TD) as our evaluation metrics. The route completion is defined as the number of routes successfully completed within a single episode, while total traveled distance refers to the cumulative distance covered by the vehicle during that episode. For safety performance assessment, we measure the collision rate (CR) and success rate (SR). The Collision Rate measures the percentage of episodes that involve collision events, while the success rate evaluates the model's ability to successfully reach the destination across a set of 10 predefined routes. During training, safety performance is evaluated based on the collision rate over the entire training process, while during testing, the success rate of completing the preset routes serves as the primary metric.

C. Implement Details

To simulate the acquisition of Bird's Eye View (BEV) images, we mount a virtual camera at a fixed height of 6.4 meters above the vehicle in the simulator. The camera maintains a constant relative pose to the vehicle during its motion. The original BEV images have a resolution of 517×517 pixels. We adopt OpenCLIP's ViT-bigG-14 as our foundational VLM to compute semantic similarity scores. The model takes a 224×224 image, divided into non-overlapping 14×14 pixel patches as input, which we obtain by resizing BEV images. To ensure stable and consistent semantic reward generation, all CLIP components remain frozen throughout our experiments. All models are trained for 1,000,000 steps on the same setting using two NVIDIA GeForce RTX 3090 GPUs, with a batch size of 64 and a replay buffer size of 100,000. Hyper-parameters used in our experiments are shown in Tab. 1.

Hyper-parameter	value
Bounding box threshold θ	0.65
IDM acceleration exponent δ	4
CBF factor α	0.25
Comfortable deceleration b	$3m/s^2$
Comfortable acceleration c	$2m/s^2$
Driver reaction time T	1s
Minimum static distance d_{min}	2m
Sensing distance D	7m
Penalty factors β	0.5
Similarity positive weighting factor λ_1	0.2
Similarity negative weighting factor λ_2	0.2
Similarity lower threshold θ_{min}	-0.03
Similarity upper threshold θ_{max}	0.0

Table 1. Hyper-parameters used in our experiments.

α	AS ↑	RC ↑	TD ↑	SR ↑
0.1	18.0±0.11	0.97±0.01	2023.1±3.0	0.91±0.01
0.25	**18.5±0.13**	**0.98±0.01**	**2054.3±21.6**	**0.94±0.02**
0.5	17.6±0.79	0.96±0.01	2016.0±22.5	0.90±0.01
1	17.7±0.24	0.96±0.01	2011.1±27.8	0.89±0.01
10	17.9±0.20	0.94±0.01	1985.4±22.7	0.87±0.02

Table 2. Performance results across different hyper-parameters.

Text Prompts. We employ the CLG text prompts and target semantic query to achieve global visual semantic assessment and real-time semantic localization, respectively. Specifically, we use "The road is clear with no car accidents" and "Two cars have collided with each other on the road" as the positive and negative language goals to assess potential collision risks in the scene. In parallel, the semantic query "car" is adopted as the target concept for vehicle localization, enabling the agent to identify surrounding vehicles and avoid potential collisions.

Vehicle Positioning. Since the camera is fixed to maintain a constant relative position and orientation during driving, the ego vehicle consistently appears at the same location with a bottom-up driving direction in the BEV image. To compute the distance to the preceding vehicle, we only consider

those within a certain forward range, i.e., vehicles located in specific pixel regions of the BEV image.

Pixel-Adaptive Mask Refinement: Our implementation of the Pixel-Adaptive Mask Refinement is based on Araslanov et.al [3]. It consists of three convolution-based components: LocalAffinityAbs, which captures absolute differences between each pixel and its neighbors; LocalAffinityCopy, which extracts raw neighborhood values; and LocalStDev, which computes local standard deviations for normalization. The affinity weights are obtained by applying softmax over the normalized differences and are used to update the mask over 10 iterations and 6 dilation rates [1,2,4,8,12,24].

Ablation Configuration: To examine the contribution of each reward component, we conduct an ablation study by progressively incorporating different elements into the reward design, as shown in Tab. 1 of the main paper. In the first configuration, only the VLM-based semantic score is used to evaluate visual alignment between driving states and safe-driving semantics, where the vehicle speed is directly adjusted according to the semantic score. In the second configuration, we detect the real-time inter-vehicle distance and incorporate IDM guidance. The third configuration further integrates CBF constraints on top of IDM guidance. In the fourth configuration, we introduce the safe distance reward defined in Eq. 17 of the main paper to explicitly penalize violations of distance constraints. Finally, the complete configuration integrates all components, yielding a unified reward formulation that couples visual semantics with physics-based safety constraints for robust safe-driving policy learning.

Different Traffic Scenarios Settings: As shown in Tab. 4 of the main paper, we set two active traffic scenarios, obstacle negotiation and unstructured intersection negotiation, to evaluate the scalability of our framework. The performance is measured by AS and SR, representing driving efficiency and safety, respectively. In the obstacle negotiation scenario, the ego vehicle drives along ten predefined routes and must overtake various static obstacles, including cars, trucks, and bicycles, while safely reaching the target point. In the unstructured intersection negotiation scenario, the ego vehicle drives through a T-shaped intersection without traffic lights under dense traffic flow, where it must make autonomous decisions to avoid potential collisions and reach the goal safely. Experiments for unstructured intersection negotiation are conducted at the same intersection with different initial traffic conditions.

D. Additonal Ablation Studies

Sensitivity to Hyper-Parameters. We perform a sensitivity analysis of our proposed method under dense traffic conditions to investigate the impact of the critical hyperparameter α associated with the CBF component. Specifi-

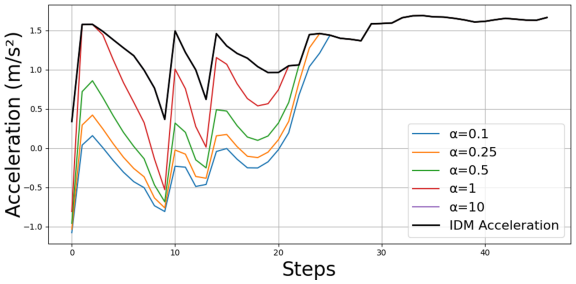

Figure 3. Acceleration under Different α Parameters.

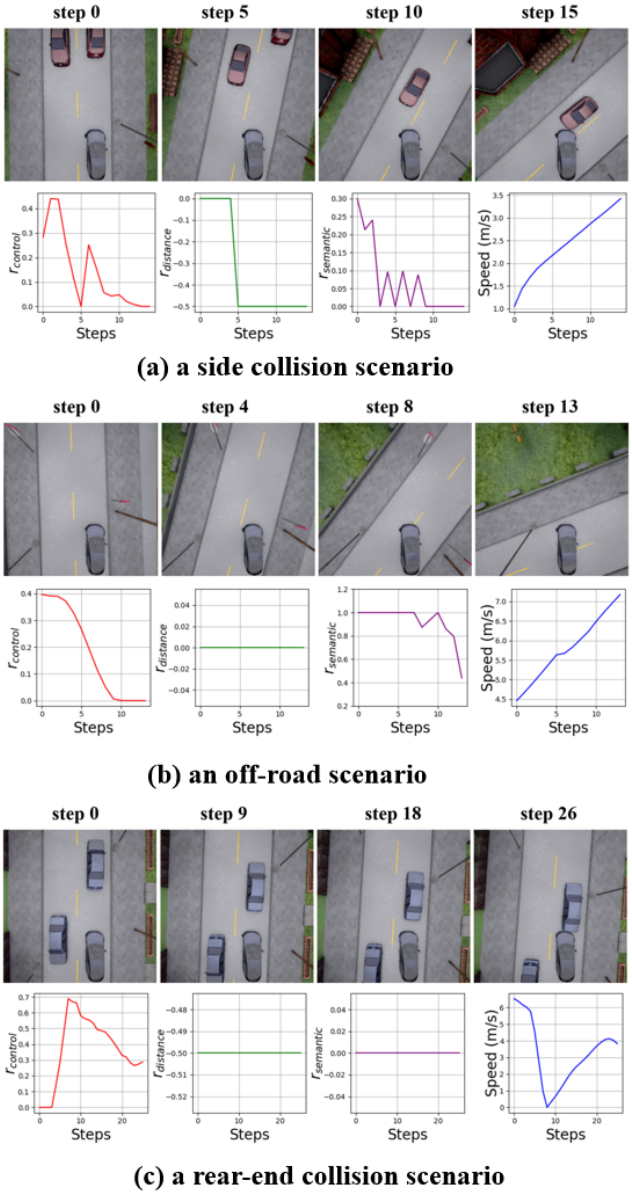

Figure 4. Visualization of the reward dynamics and system states in different driving scenarios.

cally, α is used to define the CBF derivative condition in Eq. 13c. We utilize the same car-following scenario to compute the theoretical acceleration profiles under various values of the CBF parameter α, alongside the baseline acceleration curve generated by the IDM. As shown in Fig. 3, when α is small, the system exhibits more conservative driving behavior, characterized by lower acceleration to satisfy safety constraints strictly. As α increases, the influence of the CBF condition gradually diminishes, leading the system to adopt more aggressive control strategies. Notably, when α becomes sufficiently large (e.g., $\alpha = 10$), the optimized acceleration profile closely aligns with that of the IDM, indicating that the effect of the CBF constraint has vanished. We next evaluate the model's performance under varying values of the CBF parameter α in the range $[0.1, 10]$. As shown in Tab. 2, larger values of α (e.g., α = 1,10) lead to a decline in success rate, suggesting that overly relaxed safety constraints may impair decision quality. However, within the range of $[0.1, 1]$, the model maintains relatively close performance, demonstrating a degree of robustness with respect to this parameter. This robustness means that extensive fine-tuning of α is not necessary to achieve desirable results.

E. More Qualitative Results

To further investigate our method, we provide a detailed visualization analysis of the reward dynamics and system states in different driving scenarios. Fig. 4 shows image sequences and their corresponding reward dynamics for three driving scenarios. In the side collision scenario (Fig. 4(a)), as the vehicle gradually approaches the oncoming vehicle, the semantic reward $r_{semantic}$ decreases, reflecting an increasing collision risk. When the inter-vehicle distance becomes critically small at step 5, the safety distance reward $r_{distance}$ imposes a penalty, signaling imminent danger. Meanwhile, the control reward $r_{control}$ continues to decline, encouraging deceleration to mitigate the risk of collision. In the off-road scenario (Fig. 4(b)), since no vehicles are present ahead, $r_{distance}$ remains zero. However, as the ego vehicle drifts laterally away from the lane, $r_{control}$ decreases, penalizing unsafe deviations. When the deviation reaches a critical threshold with potential for collision, $r_{semantic}$ drops sharply, indicating elevated risk and reduced semantic consistency with safe driving behavior. In the rear-end collision scenario (Fig. 4(c)), when multiple vehicles are in close proximity, $r_{distance}$ and $r_{semantic}$ respectively remain at -0.5 and 0, indicating limited sensitivity to dynamic risk changes. In contrast, $r_{control}$ adapts dynamically to the deviation between the current and desired speed. Around step 9, as the ego vehicle decelerates to a stop while maintaining a safe gap, $r_{control}$ rises, signifying a low-risk state and demonstrating its effectiveness in dynamic risk assessment and safe decision-making. Collectively, these results illustrate how the complementary reward components respond to diverse hazards across scenarios, guiding the agent toward safer and more adaptive driving behavior.