

# Supplementary Materials

## LORD: Large Models based Opposite Reward Design for Autonomous Driving

Xin Ye\* Feng Tao\* Abhirup Mallik Burhaneddin Yaman† Liu Ren  
Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)  
{xin.ye3, feng.tao2, abhirup.mallik, burhaneddin.yaman, liu.ren}@us.bosch.com

### A. Implementation Details

#### A.1. Detailed Setup of Highway-env

We follow the code repository <sup>1</sup> of GRAD baseline to setup Highway-env. More particularly, we define the state space  $S$  of the ego vehicle as its kinematic observation which is a  $V \times F$  array provided by the environment that describes a list of  $V$  nearby vehicles by a set of features of size  $F$ , including the vehicles' positions, speeds and orientations. We adopt the discrete meta-actions as the ego vehicle's action space  $A$  that consists of lane and speed change. Table 1 shows our customized configurations and we use default values for other parameters. During training, we set `lane_count` as 4 and `vehicles_density` as 2 to train all methods in `lane-4-density-2` setting. In addition, we set `duration` as 60 to train the agent to address long-horizon tasks. During testing, we set `lane_count` and `vehicles_density` accordingly and change `duration` to 30 to evaluate the success rate of all methods in `lane-4-density-2`, `lane-5-density-2.5` and `lane-5-density-3` settings.

#### A.2. Highway-env Visual Modification

As shown in Fig. 1a and 1b, Highway-env provides a simplified visualization. All vehicles are depicted as rectangles where the ego vehicle is colored in green and npc vehicles are colored in blue (see Fig. 1a). When a collision happens, the victim vehicles are colored in red as Fig. 1b illustrates. These rendered images are likely out of the training distribution of the large pretrained models. In consequence, our large model based rewards may not work well for the settings with image or video based observations. To remedy this issue, we modify the graphics of the Highway-env by replacing the rectangle textures with more photo-realistic car images. Besides, we also remove the useless background of the images. Fig. 1c and Fig. 1d show the snapshots of our modified Highway-env in which the white

Table 1. Configurations of Highway-env for training.

Parameter	Value
observation	
-type	Kinematics
-features	[presence, x, y, vx, vy, cos_h, sin_h, heading]
-absolute	True
-normalize	True
-vehicles_count	33
-see_behind	True
action	
-type	DiscreteMetaAction
-target_speeds	[20, 25, 30, 35, 40]
duration	60
ego_spacing	4
lane_count	4
vehicles_density	2

car denotes the ego vehicle and blue cars are npc vehicles.

#### A.3. Observation Design

To enable a more efficient use of large pretrained models as zero-shot reward models, we empirically adopt the following observation designs as inputs to the large pretrained models. (1) For image based observation, we adopt the simulated image rendered by Highway-env with the parameter `scaling` being set to 10. We then replace the rectangles used to represent the ego and npc vehicles with more photorealistic car images. We further remove the image background and we crop the image to the size of  $224 \times 224$  centered on the ego vehicle. (2) For video based observation, we stack the latest 30 consecutive image based observations with a 15Hz frequency. (3) For text based observation, we only pay attention to the nearby vehicles that are within  $5 \times \text{ego\_speed}$  meters of the ego vehicle and drive on the same, left or right lane of the ego vehicle. We then calculate the ego vehicle's time to collision (ttc) to each of these attended vehicles. If a vehicle drives on the same lane

\*Equal contributions. † Corresponding author.

<sup>1</sup><https://github.com/zerongxi/graph-sdc>

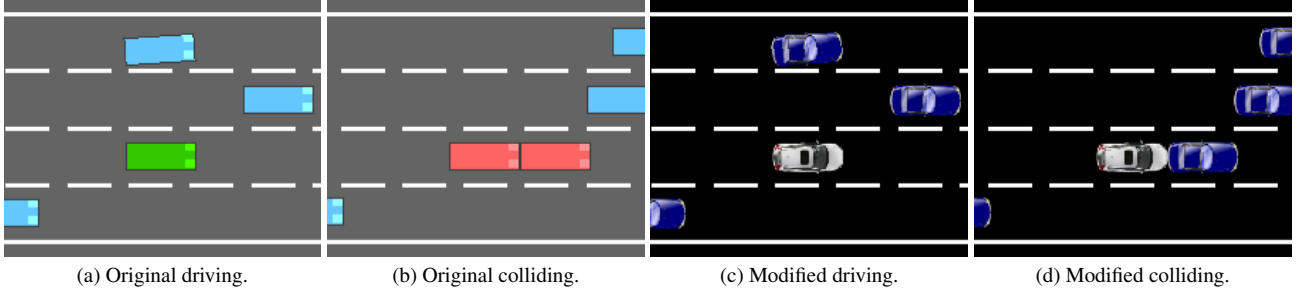


Figure 1. Illustrations of the original and the modified Highway-env. In the modified environment, white car denotes the ego vehicle and blue cars depict the npc vehicles.

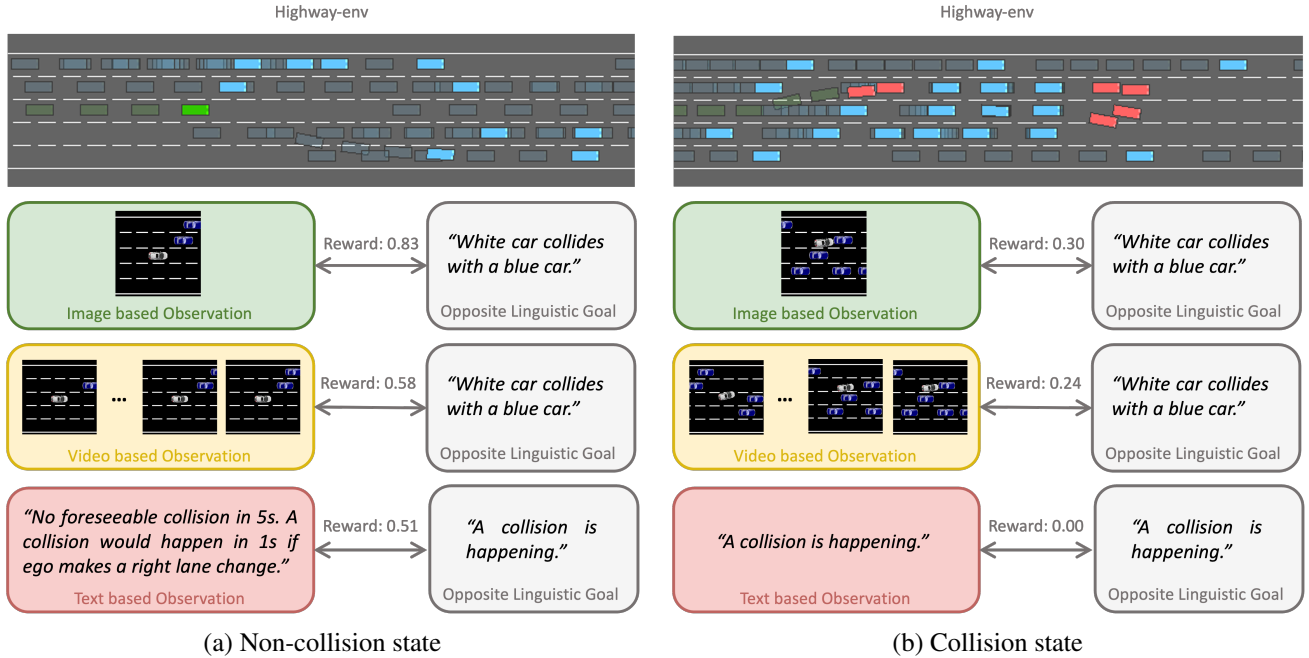


Figure 2. An example of our image, video and text based observations and the corresponding rewards for a non-collision state (a) and a collision state (b).

of the ego vehicle and the ttc is smaller than 5s, we describe it in our text based observation by *"A collision will be happening in {ttc}s."* Otherwise, we give a description of *"No foreseeable collision in 5s."* We also describe conditional collisions by *"A collision would happen in {ttc}s if ego makes a left/right lane change."* Examples of the three types of observations can be found in Fig. 2.

## B. Case Study

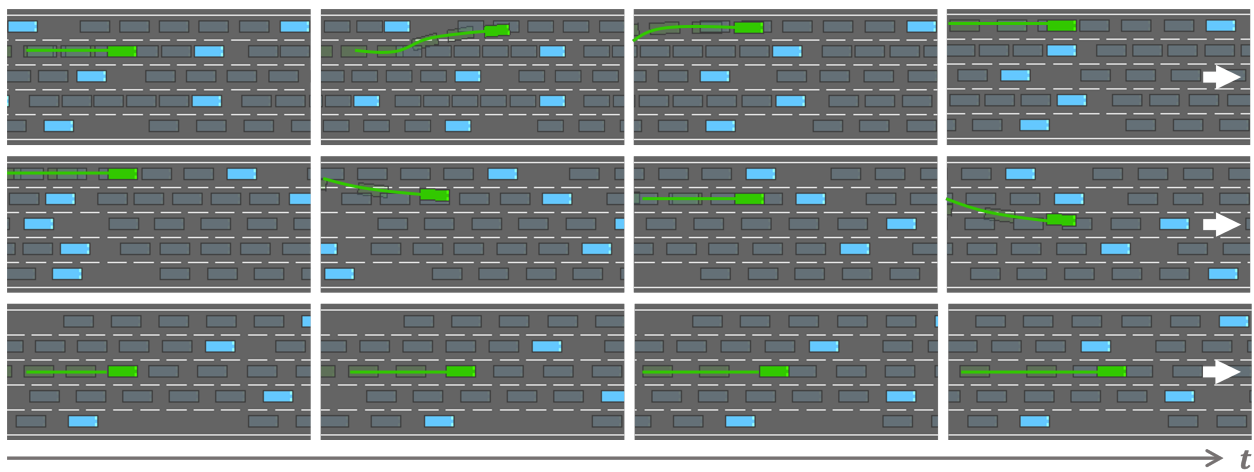
### B.1. Rewards from Different Observations

Fig. 2 (a) and (b) present the rewards we get from different observations for a non-collision and a collision state respectively. While the reward values are nonidentical across different observations, they are all higher for the non-

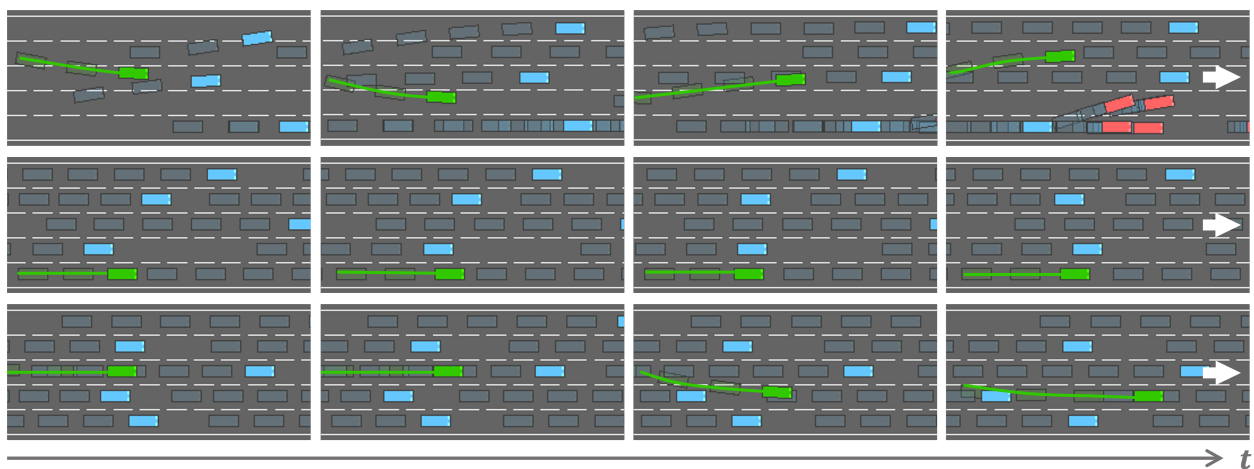
collision state compared to the collision one. In this way, the ego vehicle can distinguish the dangerous states and learn a safe driving policy.

### B.2. Qualitative Results

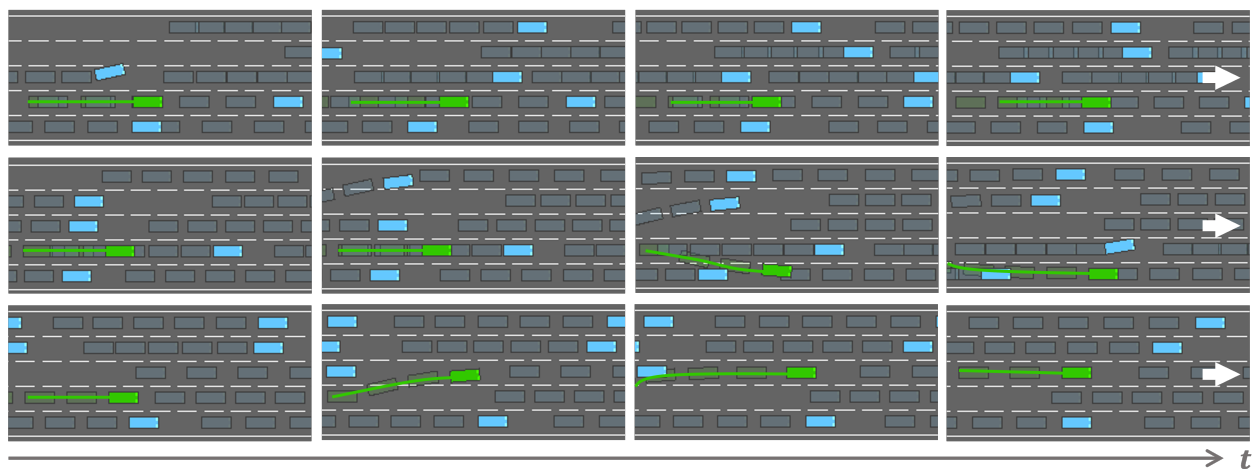
Fig. 3 shows more examples of how the driving policy learned by our LORD with image, video and text based observation performs in the lane-5-density-3 setting of Highway-env. We can observe that the ego vehicle learns diverse ways to avoid collisions in congested traffic scenarios. The results shall be better viewed in the supplementary videos.



(a) LORD with image based observation.



(b) LORD with video based observation.



(c) LORD with text based observation.

Figure 3. The driving policy learned by our LORD with image, video and text based observation.