

DriverGaze360: OmniDirectional Driver Attention with Object-Level Guidance

Supplementary Material

7. Traffic Scenarios

7.1. Goal-Direction Navigation

In goal-directed navigation driver participants follow a pre-planned route using audio navigation cues (e.g., go straight, turn left, merge right), while interacting with regular city traffic and adhering to all traffic rules. Each session begins with randomized environmental conditions and progresses through diverse road types—including urban streets, highways, and multilane roundabouts. Session durations range from 7–15 minutes, totaling to ~ 370 minutes for all sessions. Figure 10 shows one example of a goal-directed navigation session, consisting of 2 scripted sub-scenarios embedded within naturalistic driving. The range of scenarios varies from 2–10 depending on the route.

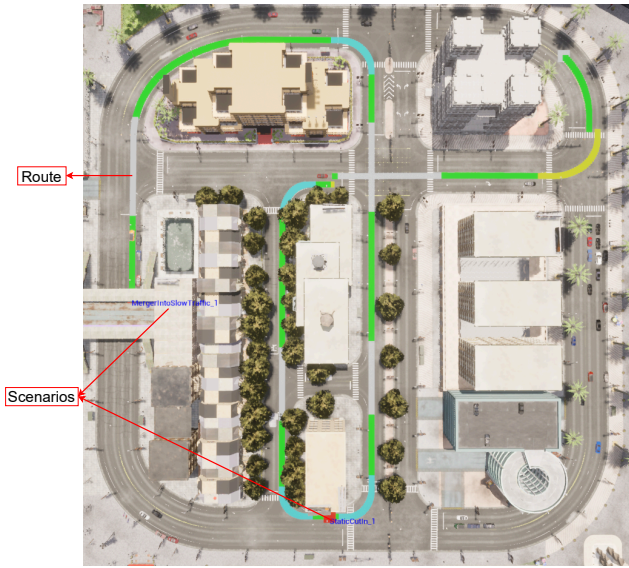


Figure 10. Example of Goal-Directed Navigation with two sub-scenarios.

7.2. Safety-Critical Events

Figure 11 summarizes the safety-critical scenarios integrated into the data collection pipeline. These events are implemented in SCENIC [13] and parameterized by weather, map selection, traffic density, spawning locations, and actor configurations. Each event is executed as a short (≤ 60 s) self-contained clip. After every run, the simulator resets with a newly sampled parameter set, providing broad coverage across conditions. For each safety-critical scenario, we collect roughly 5–6 samples for each event from each driver.

The distribution of collected examples across events is shown in Figure 12.

We define these scenarios as:

- **Highway emergency braking:** sudden deceleration from leading vehicle with blocked adjacent lanes in a highway.
- **Highway merging:** Highway merging with dense traffic from the right lane.
- **Urban pedestrian crossing:** Pedestrian crossing with partial occlusion in urban roadway.
- **Signalized left turn:** Turning left on a signalized intersection with oncoming vehicles.
- **Highway cut-in:** Vehicle pull in front with a short headway on a highway.

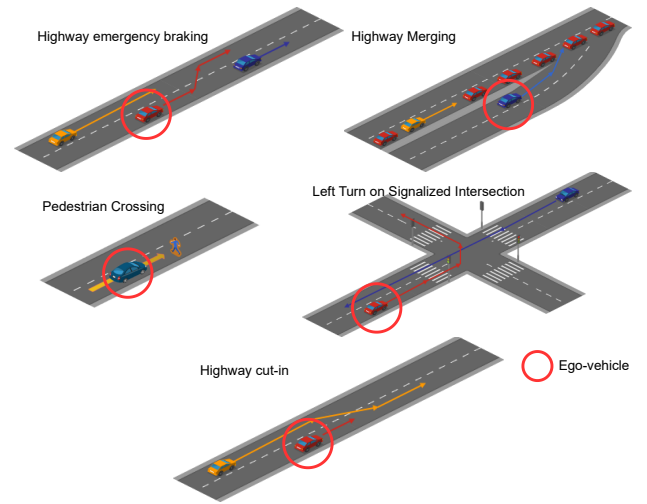


Figure 11. Safety-critical events.

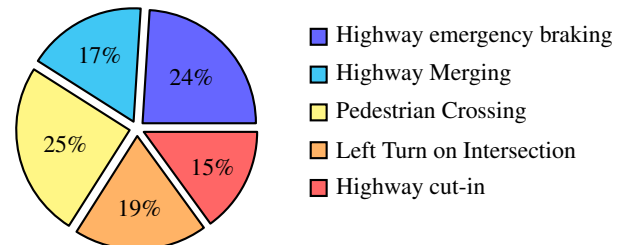


Figure 12. Safety-critical event distribution. Total collected driving time in the dataset for safety-critical events is 85 minutes.



Figure 13. Fixation calibration between CARLA (left) and eye-tracker (right).

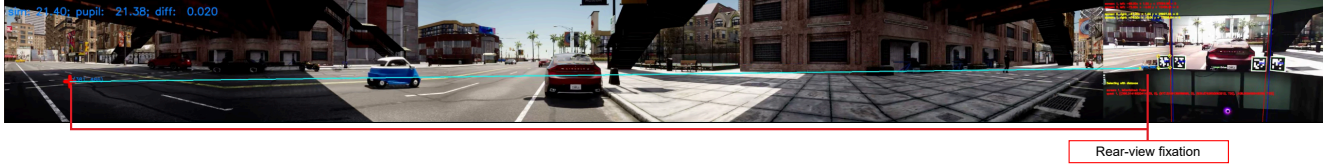


Figure 14. Fixation calibration for rear-view mirror. Driver checks the mirror before pulling out of parking.

8. Eye-Tracker Fixation Alignment

8.1. Fixation Calibration

We map fixation points from the eye-tracker frame to the CARLA coordinate frame using AprilTags [30] and homography transformations. Figure 13 illustrates the calibration procedure for the forward view, while Figure 14 shows the corresponding setup for the rear-view mirror.

We begin by detecting all AprilTags present in the image. Using the coordinates of each detected tag, we construct vertical lines from the left and right edges of the tag. These parallel lines allow us to determine which screen the user’s gaze currently falls on. Once the active screen is identified, we compute the homography between that screen and the CARLA simulator. The full procedure is detailed in Algorithm 2.

Algorithm 2: Fixation Calibration

Input: Fixation coordinate in eye-tracker frame X ,
Egocentric image I , Simulator frame F

Output: Fixation coordinate in simulator frame Y

- 1 Extract AprilTags from I : $T \leftarrow \text{getAprilTags}(I)$;
 - 2 Get fixated screen S using X and T :
 $S \leftarrow \text{getCurrentScreen}(X, T)$;
 - 3 Calculate homography H between S and F :
 $H \leftarrow \text{getHomography}(S, F)$;
 - 4 Compute Y : $Y \leftarrow HX$;
 - 5 **return** Y ;
-

8.2. Attention Map Generation

The driver attention map S_t for a frame at time t is built by accumulating projected fixation points in a temporal sliding window of $k = 30$ frames, centered at t . For each time step

$t + i$ in the window, where:

$$i \in \left\{ -\frac{k}{2}, -\frac{k}{2} + 1, \dots, \frac{k}{2} - 1, \frac{k}{2} \right\},$$

fixation point projections on Y_{t+i} are estimated through the homography transformation as discussed in Algorithm 2. A continuous fixation map is obtained from the projected fixations by centering on each of them a multivariate Gaussian having a diagonal covariance matrix σ :

$$S_t(x, y) = \frac{1}{k} \sum_{i=-\frac{k}{2}}^{\frac{k}{2}} \mathcal{N}((x, y) | Y_{t+i}, \sigma) \quad (1)$$

Eventually, each map S_t is normalized to sum to 1, so that it forms a probability distribution of fixation points.

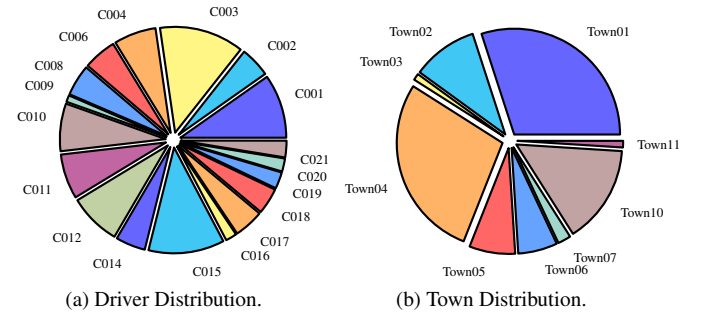
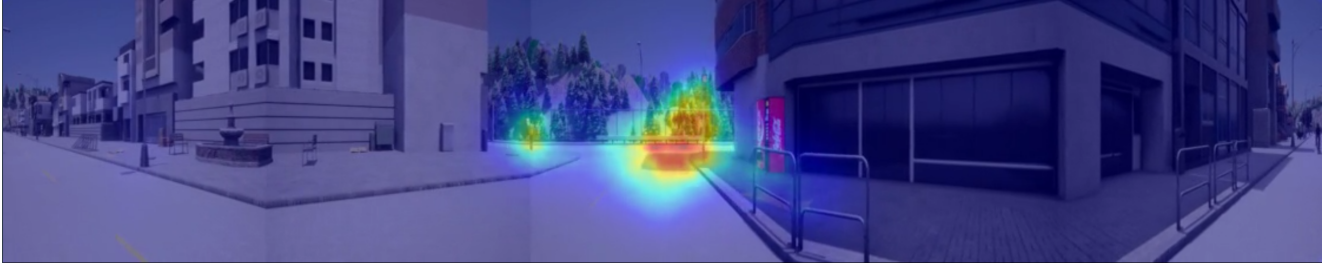


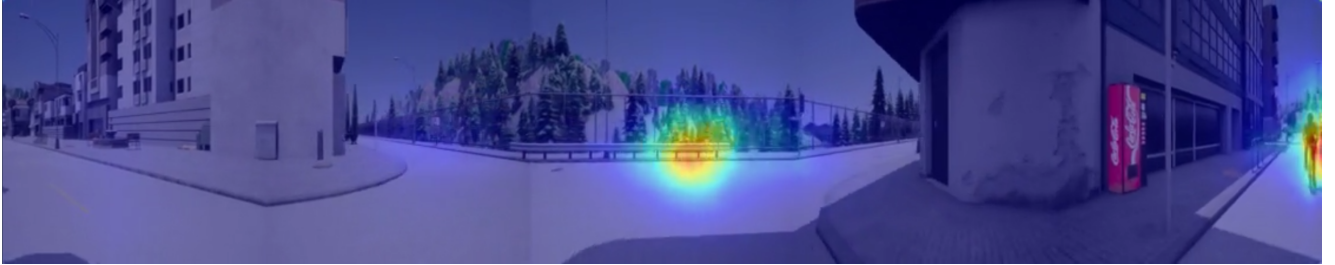
Figure 15. DriverGaze360 Statistics.

9. DriverGaze360 Statistics

We present summary statistics from DriverGaze360 in Figure 15. Driver contributions are fairly evenly distributed across the 19 participants (C001–C021, excluding C005 and



(a) Simultaneous focus on traffic light and pedestrian.



(b) Focus on the cyclist in the rear-view mirror while turning.

Figure 16. DriverGaze360 Inference Results.

C013). The contribution of each CARLA Town in the dataset is illustrated in Figure 15b. As described in Section 3.5, we partition the towns based on their geographic characteristics to ensure that no town appears in both the training and testing sets. After splitting, we balance the partitions so that the training set contains 303 minutes of footage (Towns 2, 3, 4, 7, 10, 11) and the validation set contains 234 minutes (Towns 1, 5, 6).

10. Additional Qualitative Results

Our method is simultaneously able to attend to the frontal view, traffic lights, and car in the rear-view. We demonstrate our method’s ability to predict driver attention towards critical regions—such as pedestrians, cars, and traffic lights in Figure 16a. Moreover, it can predict attention to rear-view areas during turning maneuvers; for example, in Figure 16b, the model correctly focuses on the cyclists while making a right-turn.

11. Comparison to Panoramic Methods

Uniquely to our setup, we use five rectilinear cameras (with no spherical distortion), not equirectangular projections typical of 360° saliency work. Nevertheless, we adapt two representative 360° saliency models [29, 38] for our rectilinear input. As shown in Table 7, our method outperforms these adapted baselines.

12. Limitations

The primary limitation of DriverGaze360 is the sim-to-real gap inherent to simulation-based data collection. A simula-

Table 7. Comparison to Panoramic Methods.

Model	KLD ↓	CC ↑	SIM ↑	NSS ↑
PanoConv [29]	1.450	0.599	0.425	5.540
PAVER [38]	3.375	0.089	0.070	0.236
DriverGaze360-Net (ours)	1.067	0.667	0.515	6.309

tor cannot perfectly replicate real-world driving conditions, which may influence participant behavior and gaze patterns relative to on-road driving. That said, simulation offers significant advantages: precise control over traffic and environmental conditions, and the ability to safely capture rare, high-risk events that are difficult to record in the real world. We therefore advise that claims about real-world generalization be interpreted cautiously, and encourage future work to investigate domain adaptation strategies to bridge this gap.