Interpretable Decision-Making for End-to-End Autonomous Driving

Supplementary Material

Mona Mirzaie Bodo Rosenhahn

Institute for Information Processing
Leibniz University Hannover, Germany

mirzaie@tnt.uni-hannover.de

In this supplementary document, we first analyze failure cases in challenging scenarios and discuss the underlying causes (Sec. 1). Next, we present a detailed analysis of our model's behavior through visualizations, including activation maps that highlight key decision-making regions for critical traffic participants (Sec. 2). We then describe how interpretability metrics and mask generation are applied in our evaluations (Sec. 3), and subsequently provide a quantitative analysis of concept-level interpretability (Sec. 4). Finally, we conduct an ablation study to assess the impact of hyperparameter choices on model performance (Sec. 5).

1. Failure Case Analysis

This section presents examples from Fig. 1 demonstrating scenarios that lead to lower Intersection over Union (IoU) and Ground Truth Coverage (GTC) scores. Although our model successfully identifies relevant regions, the activation maps do not always fully overlap with ground-truth bounding boxes, resulting in lower scores. This partial coverage results in lower metric values, making it challenging to fully capture the interpretability advantages of our model based solely on these metrics.

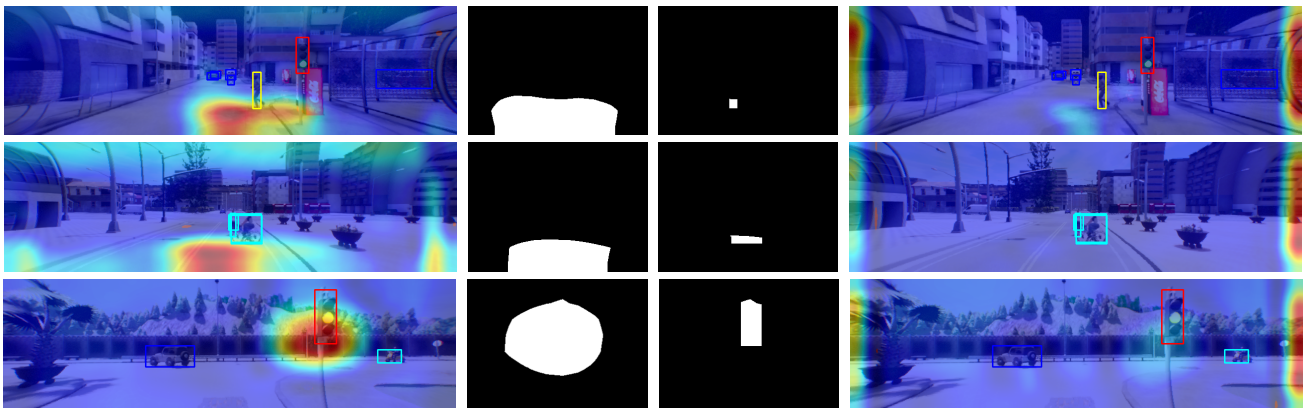

Figure 1. **Example of Failure Cases**. From left to right: heatmap of our model, heatmap binary mask, intersection of the bounding box and heatmap binary mask, and heatmap of the reproduced TCP.

2. Comparative Analysis of Activation Maps

In this section, we present additional qualitative analyses of activation maps generated using EigenCam [4] for our proposed model DTCP compared to the baseline TCP [5]. As depicted in Fig. 2 **(a-g)**, DTCP effectively localizes attention on regions critical for driving decisions, significantly enhancing interpretability. Detecting traffic lights, especially those situated at higher and farther positions, is challenging due to their relatively small size in camera images. However, as illustrated in Fig. 2 **(a)**, DTCP effectively addresses this difficulty by accurately focusing its activation map within the bounding box of distant traffic lights. Fig. 2 **(b-c)** illustrates that our model assigns significant attention to yellow traffic lights. Specifically, in **(c)**, upon recognizing the yellow traffic light, the model effectively identifies vehicles within the intersection in our driving path as potential collision risks. As depicted in Fig. 2 **(d)**, our model carefully monitors the surrounding environment during turning maneuvers, proactively identifying potential risks from dynamic objects, such as cyclists (marked by a cyan bounding box) or pedestrians, that could unexpectedly enter our driving path. Fig. 2 **(e-g)** demonstrates that our model remains highly

vigilant in detecting unexpected situations, identifying hazards such as vehicles and pedestrians that violate traffic rules or run red lights, thereby reducing potential collision risks.

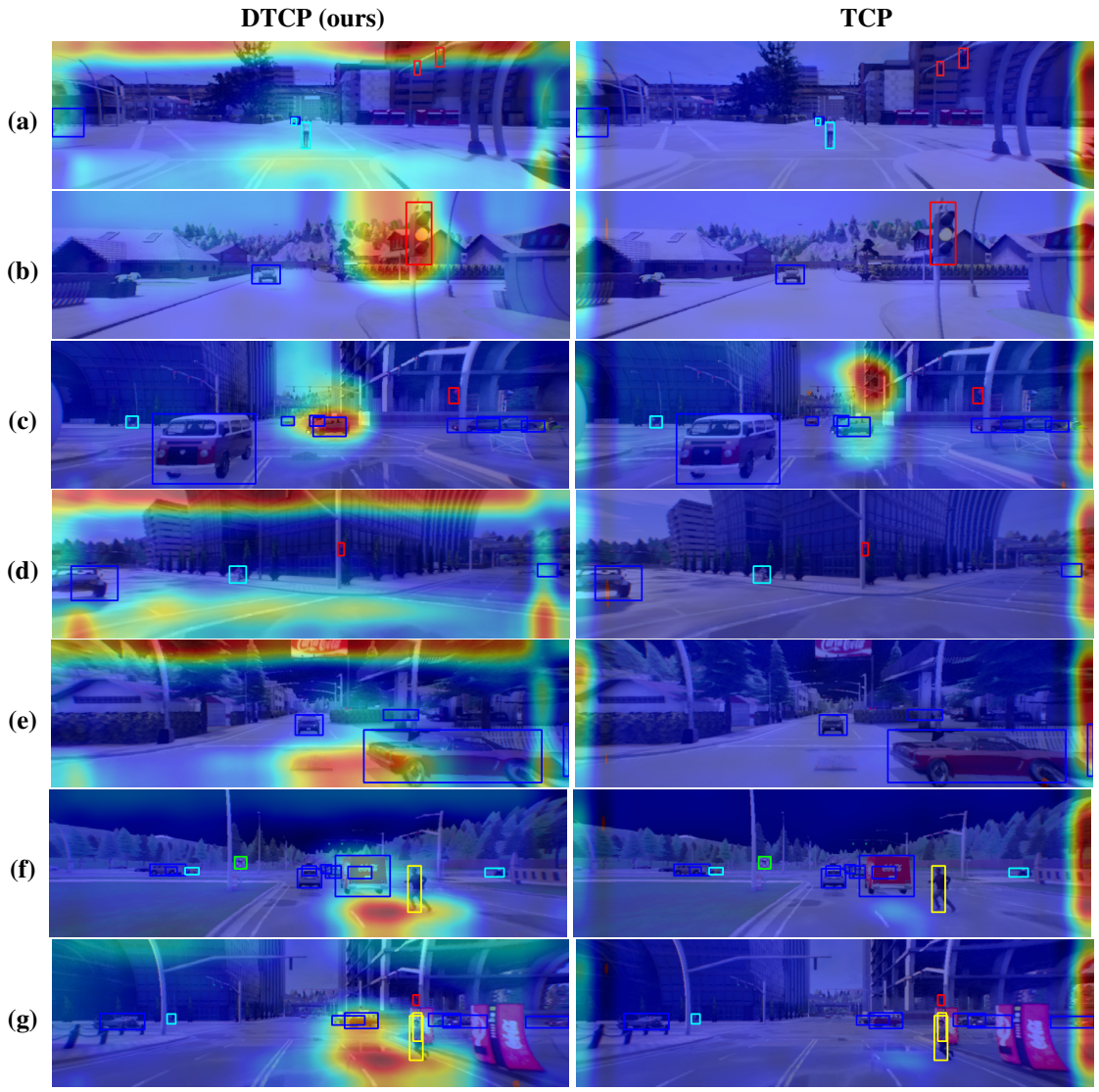

Figure 2. **Comparison of activation maps**. Left: Heatmap of our model DTCP; right: heatmap of the reproduced TCP. Our model achieves superior performance by activating key regions relevant to driving decisions, thus enhancing interpretability.

3. Mask Generation and Interpretability Metrics

For measuring interpretability using *Shared Interest* metrics (IoU, GTC, SC) [1], we generate binary masks from object bounding boxes and our model's heatmap, as illustrated in Fig. 3. Ground-truth bounding boxes are created for categories most likely involved in infractions—pedestrians, cyclists, vehicles, and traffic lights.
For a ground-truth set ($\mathcal{G}$) and a saliency set ($\mathcal{S}$), *Shared Interest* metrics are defined as:

$$\text{IoU} = \frac{|G \cap S|}{|G \cup S|} \tag{1}$$

$$\text{GTC} = \frac{|G \cap S|}{|G|} \tag{2}$$

Figure 3. **Example of Generated Binary Mask**. From left to right: heatmap with overlapping bounding boxes, binary mask derived from the heatmap, and binary mask of the bounding boxes.

$$\mathrm{SC} = \frac{|G \cap S|}{|S|} \tag{3}$$

4. Concept-Level Interpretability

To further demonstrate that DTCP attends more effectively to human-understandable concepts compared to the reproduced model, we compute driving-related semantic IoU, as reported in Tab. 1. Specifically, we measure IoU between model-generated saliency maps and key semantic classes: *road*, *roadline*, and *sidewalk*. The advantage of DTCP in allocating greater attention to dynamic traffic participants and traffic lights is already illustrated in Tab. 3 of our main paper. Results in Tab. 1 further confirm DTCP's superior ability to attend to driving-critical regions, reflecting human-like behavior essential for safe driving.

Table 1. **Interpretability evaluation for driving-related concepts.** We report the average IoU scores. Our method allocates greater attention to decision-critical areas in driving scenarios.

Routes	**Method**	**IoU ↑**		
		Roadline	**Road**	**Sidewalk**
Town01	TCP [5]	0.00	0.01	0.04
	DTCP (**ours**)	**0.01**	**0.30**	**0.21**
Town02	TCP [5]	0.00	0.02	0.03
	DTCP (**ours**)	**0.01**	**0.31**	**0.14**
Town03	TCP [5]	0.00	0.02	0.08
	DTCP (**ours**)	**0.02**	**0.41**	**0.21**
Town05	TCP [5]	0.00	0.02	0.06
	DTCP (**ours**)	**0.02**	**0.40**	**0.11**

5. Diversity Loss Weight Effect

Tab. 2 presents the impact of different loss weights (λ_{div}) for our proposed diversity loss, as introduced in Sec. 3.3 of the main paper. We report the mean and standard deviation of the driving score over three evaluation runs in the CARLA simulator [3] on the LAV benchmark [2]. As shown, setting $\lambda_{\text{div}} = 0.00005$ yields the best driving performance with the lowest standard deviation.

Table 2. **Effects of Diversity Loss Weight**. An empirical analysis to determine the optimal value of λ_{div} in our model on the LAV Routes benchmark [2].

λ_{div}	5×10^{-1}	5×10^{-2}	5×10^{-3}	5×10^{-4}	5×10^{-5}	5×10^{-6}
Driving Score	42.87 ± 3.25	55.58 ± 6.44	48.22 ± 4.01	44.01 ± 1.29	$\mathbf{60.42 \pm 1.27}$	52.65 ± 1.78

References

[1] Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, and Hendrik Strobelt. Shared interest: Measuring human-ai alignment to identify recurring patterns in model behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022. 2

[2] Dian Chen and Philipp Krähenbühl. Learning from all vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17222–17231, 2022. 3

[3] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator. In *Conference on robot learning*, pages 1–16. PMLR, 2017. 3

[4] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–7. IEEE, 2020. 1

[5] Penghao Wu, Xiaosong Jia, Li Chen, Junchi Yan, Hongyang Li, and Yu Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. *Advances in Neural Information Processing Systems*, 35:6119–6132, 2022. 1, 3