Understanding and Modeling the Effects of Task and Context on Driver's Attention Allocation Supplementary Materials

Iuliia Kotseruba and John K. Tsotsos

I. DR(EYE)VE GROUND TRUTH

Here we provide additional qualitative examples to illustrate some of the shortcomings of the original ground truth in DR(eye)VE dataset. Figure 1 shows several common scenarios where fixations are incorrectly projected from the driver's eye-tracking glasses (ETG) view to the rooftop Garmin camera view (GAR): a) when the driver looks down, the scene is not visible in the ETG view and the homography cannot be established, thus driver's gaze is projected incorrectly onto the hood of the car; b) gaze towards the rearview mirror may be incorrectly projected onto elements of the scene behind the mirror; c) when making turns, the driver may look at the areas that are outside of the GAR camera view, however, these fixations are still mapped onto the scene in the original ground truth; d) since the GAR camera is mounted outside, the raindrops may cause issues with transformation even though the driver's gaze is within the GAR camera field of view.

Figure 2 demonstrates the results of removing saccade to the speedometer from the ground truth. Figure 3 shows an example of ground truth containing driver's fixations that are outside of the GAR camera field of view.

II. TASK AND CONTEXT REPRESENTATION

This section details how task and context labels were used for training the proposed SCOUT model.

Table I lists all task and context information grouped into three sets: *global context* — represents global weather, time of day, and location, *local context* — information about the upcoming intersections and actions to be taken along the route, *current action* — information about current longitudinal action (speed and acceleration) and lateral action (turns, lane changes, and driving straight).

Global context labels were provided with the dataset and are unchanged. Local context information relies on the new manual annotations and GPS data provided with the dataset. The DR(eye)VE dataset description in [1] does not specify whether the drivers knew the complete route in advance or were given step-by-step instructions. In the latter case, it is also unknown how well in advance such instructions were given and in what form. Therefore, we use guidelines designed for navigation systems that provide voice navigation commands to drivers [2]. We first compute the current distance in meters to the nearest intersection on the route using available GPS coordinates of the ego-vehicle. If that distance is larger than the max lead distance defined as $(\mathrm{speed}(\mathrm{km/h}) * 2.22) + 37.144$, we set it to *inf* to indicate that the intersection is too far, otherwise, we use the distance directly.

TABLE I: Task and context representation

Category	Features	Values
Global context	Weather	text label: sunny, cloudy, or rainy
	Time of day	text label: morning, evening, night
	Location	text label: highway, urban, suburban
Local context	Distance to intersection (m)	numeric value
	Priority	right-of-way, yield
	Next action	text label: turn right, turn left, drive straight
Current action	Speed (km/h)	numeric array
	Acceleration (m/s2)	numeric array
	Action	text label: turn right/left, lane change right/left, drive straight

To represent current longitudinal action, we use the sequence of speed and acceleration values. Raw speed values provided with the dataset are interpolated, passed through the median filter to remove outliers, and then used to compute acceleration. For lateral action, we use manually assigned text labels since heading angle is too coarse and inaccurate and yaw information is not available. If there is more than one label in the sample, the most frequently occurring label is chosen.

III. BENCHMARK RESULTS

A. Implementation of Kullback-Leibler divergence

Kullback-Leibler divergence (KLD) is the key dissimilarity metric commonly reported when evaluating saliency algorithms [3].

KLD is defined as $P\|Q = \sum p_i \ln\frac{p_i}{q_i}$, where P and Q are two distributions being compared. Since saliency maps are usually sparse, cases when non-zero value in one image ($p_i > 0$) corresponds to zero value in another ($q_i = 0$) occur frequently. To prevent division by zero, a small constant ϵ is added to the denominator in the equation above. Consequently, if the saliency values are small, the result can be dominated by this parameter. Depending on the implementation, ϵ may be set differently. Below are three typical values used:

- MATLAB machine epsilon: $\epsilon = 1.1920929\mathrm{e}-07$;
- `numpy` machine epsilon: $\epsilon = 2.2204\mathrm{e}-16$[1] (default in Python implementations);
- A small constant: $\epsilon = 0.0001$.

[1] `np.finfo(np.float32).eps`

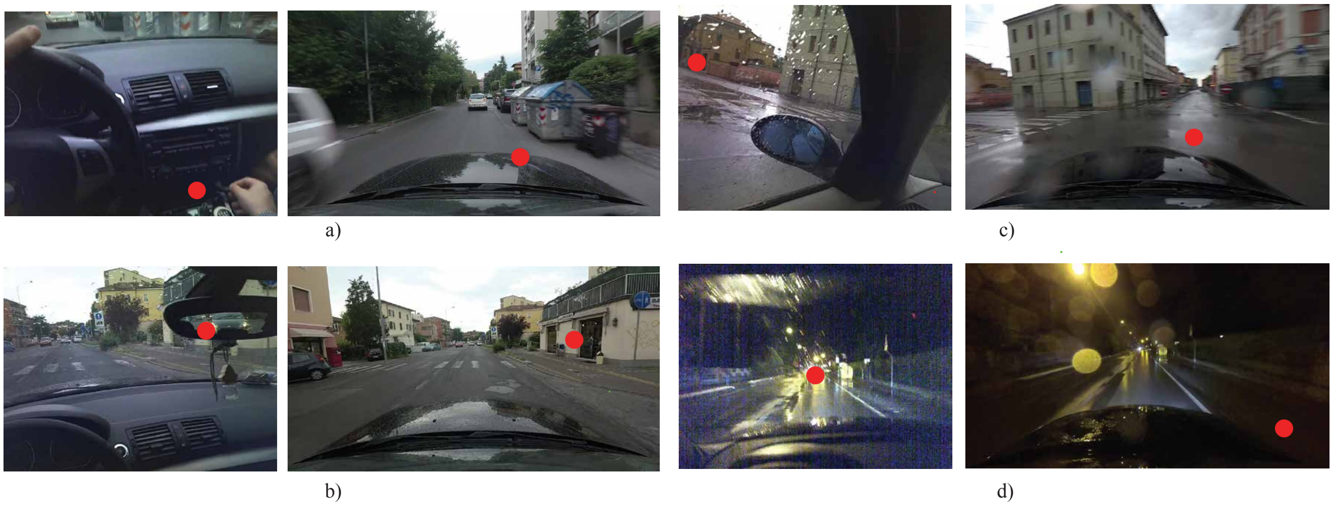

Fig. 1: Common scenarios where driver's gaze (shown as a red circle) is incorrectly projected onto the scene. Left column shows the view from the driver's eye-tracking glasses (ETG) and the right column shows the view of the rooftop Garmin camera (GAR). The following scenarios are demonstrated: a) looking down; b) checking rearview mirror; c) looking outside the GAR camera view during turns; d) raindrops on the windshield.

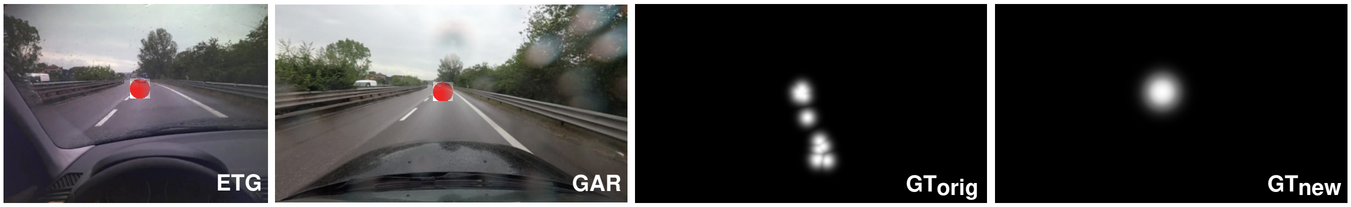

Fig. 2: Original (GT$_{orig}$) and new (GT$_{new}$) ground truth for the same frame (with fixations indicated with red circles in the drivers' (ETG) and rooftop camera (GAR) views). Note the "trail" of points in the original ground truth where the driver made a brief saccade towards the speedometer. The data points corresponding to saccaded are propagated across dozens of frames due to the aggregation procedure. These are absent in the new ground truth shown in the leftmost image.

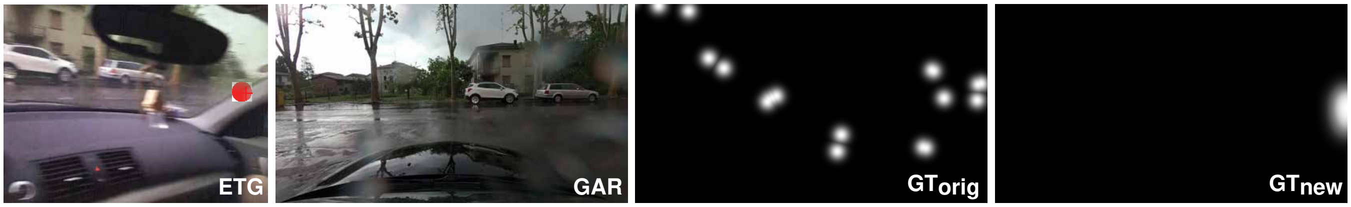

Fig. 3: Original (GT$_{orig}$) and new (GT$_{new}$) ground truth for the same frame (with fixations indicated with the red circle in the drivers' (ETG) view). The driver is looking through the side window, however, in the original ground truth, points representing saccades in the past frames and out-of-bounds fixations are still projected onto the scene in a random pattern. In the new ground truth, to preserve some gaze information, we push out-of-bounds fixations towards the edge of the frame to indicate the approximate direction of where the driver is attending.

The Python implementation of KLD used in the DR(eye)VE evaluation code[2] follows the widely used Matlab implementation [4] for the MIT saliency benchmark[3], but this Python version is numerically unstable for some inputs. We compared the DR(eye)VE version of KLD to two other Python implementations: Fahimi & Bruce [5], [6] and `pysaliency` [7], [8], the official evaluation code for the MIT/Tuebingen saliency benchmark [4].

Table II shows results of testing different implementations with different values of ϵ on three test cases:

1) *Two randomly generated identical images*. KLD for two identical images is expected to be 0, however, DR(eye)VE implementation fails to produce this value with MATLAB and small constant ϵ for two randomly

[2] `https://github.com/ndrplz/dreyeve/blob/master/experiments/metrics/metrics.py`

[3] `saliency.mit.edu`

[4] `saliency.tuebingen.ai`

TABLE II: Results of testing different KLD implementations with several standard values of ϵ on three scenarios. Diverging results of DR(eye)VE implementation are highlighted in orange.

		ϵ		
		MATLAB	Numpy	Small constant
Data	Implementation	1.19E-07	2.22E-16	0.0001
Identical images	DR(eye)VE	-0.029013	0	-2.771208
	Fahimi & Bruce	0	0	0
	pysaliency	0	0	0
Random images	DR(eye)VE	0.087539	0.499305	-5.205599
	Fahimi & Bruce	0.499304	0.499305	0.498398
	pysaliency	0.499304	0.499306	0.498398
Salmaps from DR(eye)VE	DR(eye)VE	2.322985	7.277252	-2.70738
	Fahimi & Bruce	5.135229	9.888739	3.525612
	pysaliency	5.135229	9.888739	3.525612

TABLE III: Evaluation results on the original DR(eye)VE dataset ground truth using DR(eye)VE implementation with MATLAB ϵ and Fahimi & Bruce code with Numpy ϵ.

Model	KLD (DR(eye)VE implementation)	KLD (Fahimi & Bruce)
BDD-ANet	1.92	3.15
DReyeVENet	1.59	2.63
ADA	1.56	2.33

generated identical images (1000 × 1000 px) .

2) *Two different randomly generated identical images.* Similarly, DR(eye)VE implementation of KLD diverges significantly for MATLAB and small constant ϵ when tested on two *different* randomly generated images. Two other implementations remain stable with small differences in the sixth decimal.

3) *Randomly selected ground truth and predicted saliency maps from DR(eye)VE.* DR(eye)VE implementation diverges from others on a pair of ground truth and predicted saliency maps randomly selected from the dataset. Here, even using the default ϵ generates results different from other implementations.

Table III shows the evaluation results of BDD-ANet [9], DReyeVENet [1], and ADA [10], using DR(eye)VE implementation with MATLAB ϵ and Fahimi & Bruce code with Numpy ϵ. The results obtained with the DR(eye)VE implementation with MATLAB ϵ replicate the KLD values reported in the literature. Fahimi & Bruce implementation produces different absolute values, but preserves the relative ranking of the models. Since finding the cause of the numerical issues is beyond the scope of these experiments, we report the results for all algorithms using the more accurate Fahimi & Bruce implementation with Python ϵ.

B. Additional benchmark results

Table IV shows additional evaluation results on action and context sequences and CC, SIM, and NSS metrics.

IV. Additional SCOUT results

Table V reports additional evaluation results on action and context sequences and CC, SIM, and NSS metrics for variants of the proposed SCOUT model.

References

[1] A. Palazzi, D. Abati, S. Calderara, F. Solera, and R. Cucchiara, "Predicting the Driver's Focus of Attention: the DR (eye) VE Project," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 41, no. 7, pp. 1720–1733, 2018.

[2] J. L. Campbell, C. Carney, and B. H. Kantowitz, "Human factors design guidelines for advanced traveler information systems (ATIS) and commercial vehicle operations (CVO)," Tech. Rep. FHWA-RD-98-057, US department of Transportation, Federal Highway Administration, 1998.

[3] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 41, no. 3, pp. 740–757, 2018.

[4] Z. Bylinski, "MATLAB implementation of saliency metrics." `https://github.com/cvzoya/saliency/blob/master/code_forMetrics`.

[5] R. Fahimi and N. D. Bruce, "On metrics for measuring scanpath similarity," *Behavior Research Methods*, vol. 53, no. 2, pp. 609–628, 2021.

[6] R. Fahimi and N. D. Bruce, "Code for "On metrics for measuring scanpath similarity"." `https://github.com/rAm1n/saliency`.

[7] M. Kümmerer, Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT/Tübingen Saliency Benchmark." `https://saliency.tuebingen.ai/`.

[8] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency Benchmarking Made Easy: Separating Models, Maps and Metrics," in *Proceedings of the European Conference on Computer Vision*, 2018.

[9] Y. Xia, D. Zhang, J. Kim, K. Nakayama, K. Zipser, and D. Whitney, "Predicting driver attention in critical situations," in *ACCV*, 2018.

[10] S. Gan, X. Pei, Y. Ge, Q. Wang, S. Shang, S. E. Li, and B. Nie, "Multisource adaption for driver attention prediction in arbitrary driving scenes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp. 20912–20925, 2022.

TABLE IV: Additional benchmark results on CC, NSS, and SIM metrics.

							Context (CC↑)							
		Actions (CC↑)					Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	Stop	None	Lat	Lon	Lat+Lon	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	0.13	0.54	0.31	0.33	0.18	0.18	0.07	0.66	0.24	0.30	0.32	0.37	0.06
	CDNN†*	0.27	0.75	0.50	0.61	0.38	0.57	0.19	0.77	0.50	0.57	0.40	0.66	0.21
	DeepGaze II	0.19	0.33	0.23	0.28	0.16	0.13	0.08	0.35	0.23	0.23	0.18	0.26	0.06
	UNISAL	0.33	0.44	0.31	0.38	0.22	0.22	0.17	0.45	0.30	0.33	0.30	0.34	0.08
Video	UNISAL	0.32	0.59	0.42	0.52	0.29	0.36	0.22	0.64	0.38	0.50	0.42	0.52	0.14
	ViNet	0.34	0.65	0.46	0.55	0.32	0.37	0.21	0.68	0.48	0.51	0.44	0.57	0.16
	ViNet*	0.32	0.74	0.51	0.61	0.40	0.43	0.25	0.75	0.51	0.58	0.38	0.64	0.30
	DreyeVENet†*	0.34	0.75	0.50	0.61	0.41	0.52	0.18	0.78	0.46	0.62	0.47	0.65	0.23
	BDD-ANet†	0.34	0.65	0.40	0.52	0.26	0.35	0.21	0.70	0.36	0.48	0.33	0.52	0.11
	BDD-ANet†*	0.31	0.72	0.46	0.59	0.35	0.37	0.23	0.75	0.44	0.53	0.42	0.61	0.20
	ADA	0.39	0.75	0.49	0.63	0.35	0.52	0.26	0.77	0.48	0.59	0.45	0.63	0.19

							Context (NSS↑)							
		Actions (NSS↑)					Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	Stop	None	Lat	Lon	Lat+Lon	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	0.55	1.77	1.10	1.24	0.71	0.75	0.26	2.16	0.88	1.16	1.21	1.28	0.20
	CDNN†*	1.14	2.20	1.56	1.95	1.23	1.74	0.67	2.27	1.43	1.77	1.36	1.97	0.76
	DeepGaze II	0.78	1.23	0.91	1.12	0.69	0.58	0.42	1.24	0.87	0.82	0.82	1.03	0.25
	UNISAL	1.43	1.87	1.31	1.64	0.94	1.01	0.71	1.92	1.22	1.36	1.33	1.51	0.31
Video	UNISAL	1.34	2.23	1.51	1.94	1.07	1.50	0.79	2.38	1.48	1.80	1.54	1.93	0.50
	ViNet	1.43	2.40	1.67	2.09	1.21	1.48	0.77	2.54	1.66	1.93	1.66	2.10	0.55
	ViNet*	1.27	2.13	1.56	1.93	1.31	1.50	0.86	2.17	1.45	1.83	1.34	1.94	1.03
	DreyeVENet†*	1.38	2.15	1.54	1.90	1.33	1.71	0.64	2.21	1.37	1.86	1.53	1.96	0.85
	BDD-ANet†	1.36	1.86	1.26	1.67	0.86	1.25	0.71	1.94	1.06	1.53	1.14	1.61	0.34
	BDD-ANet†*	1.32	2.22	1.49	1.97	1.20	1.39	0.78	2.31	1.38	1.74	1.46	1.98	0.68
	ADA	1.57	2.11	1.50	1.95	1.14	1.83	0.89	2.16	1.35	1.81	1.59	1.86	0.62

							Context (SIM↑)							
		Actions (SIM↑)					Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	Stop	None	Lat	Lon	Lat+Lon	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
Image	CDNN†	0.12	0.38	0.25	0.25	0.16	0.16	0.10	0.46	0.22	0.22	0.23	0.27	0.08
	CDNN†*	0.23	0.61	0.41	0.48	0.31	0.42	0.19	0.63	0.41	0.42	0.31	0.51	0.18
	DeepGaze II	0.13	0.17	0.15	0.16	0.12	0.11	0.11	0.17	0.13	0.13	0.12	0.14	0.08
	UNISAL	0.19	0.22	0.18	0.20	0.15	0.16	0.14	0.22	0.16	0.18	0.17	0.18	0.09
Video	UNISAL	0.26	0.42	0.32	0.38	0.24	0.28	0.19	0.45	0.29	0.36	0.30	0.37	0.14
	ViNet	0.25	0.43	0.32	0.37	0.24	0.26	0.19	0.45	0.34	0.34	0.29	0.37	0.14
	ViNet*	0.24	0.53	0.37	0.44	0.29	0.31	0.20	0.53	0.37	0.41	0.27	0.45	0.20
	DreyeVENet†*	0.27	0.62	0.42	0.50	0.34	0.43	0.17	0.65	0.38	0.50	0.38	0.54	0.19
	BDD-ANet†	0.24	0.46	0.30	0.38	0.22	0.26	0.18	0.51	0.29	0.32	0.24	0.35	0.12
	BDD-ANet†*	0.23	0.54	0.36	0.43	0.27	0.27	0.20	0.57	0.36	0.37	0.30	0.44	0.16
	ADA	0.28	0.57	0.38	0.47	0.28	0.30	0.21	0.57	0.39	0.41	0.32	0.46	0.16

* — model is trained on the data; †— model for drivers' gaze prediction. Notation: Stop — vehicle is stopped, Lat — lateral actions only, Lon — longitudinal only, None — no action, RoW — ego-vehicle has right-of-way. ↑ indicates that larger values are better. Green and red colors indicate the best and worst values.

TABLE V: Additional results for variants of SCOUT model on CC, NSS, and SIM metrics.

							Context (CC↑)							
		Actions (CC↑)					Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	Stop	None	Lat	Lon	Lat+Lon	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA	0.39	**0.75**	0.49	**0.63**	0.35	0.52	0.26	**0.77**	0.48	0.59	0.45	0.63	0.19
	ViNet	0.32	0.74	0.51	0.61	0.40	0.43	0.25	0.75	**0.51**	0.58	0.38	0.64	0.30
SCOUT	w/o task	**0.27**	0.70	0.50	0.57	0.38	**0.53**	0.16	0.72	0.48	0.59	0.25	0.61	0.17
	w/o task + weighted loss	0.35	0.72	0.50	0.61	0.42	0.47	0.22	0.73	0.45	0.62	0.40	0.64	0.24
	w/o task + weighted sampling	0.30	0.72	0.50	0.60	0.42	0.45	0.21	0.75	0.46	0.60	0.41	0.64	0.25
	w/ task CA[3]	0.34	0.73	0.50	0.62	0.43	0.41	0.32	0.75	0.46	0.61	0.44	0.63	0.33
	w/ task GC+LC+CA [2, 3, 4]	0.37	0.73	**0.53**	0.62	**0.46**	0.38	0.30	0.75	0.46	0.63	0.25	0.66	0.33
	w/ task GC+LC+CA [3]	0.31	0.73	0.51	0.61	0.43	0.46	0.26	0.75	0.46	0.62	0.39	0.63	0.28
	w/ task LC [2]	0.34	0.74	**0.53**	0.62	**0.50**	0.42	**0.34**	0.75	0.49	**0.64**	**0.46**	**0.67**	**0.34**

							Context (NSS↑)							
		Actions (NSS↑)					Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	Stop	None	Lat	Lon	Lat+Lon	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA	1.57	2.11	1.50	1.95	1.14	1.83	0.89	2.16	1.35	1.81	1.59	1.86	0.62
	ViNet	1.27	2.13	1.56	1.93	1.31	1.50	0.86	2.17	1.45	1.83	1.34	1.94	1.03
SCOUT	w/o task	1.63	4.40	2.97	3.51	2.20	3.07	0.92	4.53	2.97	3.55	1.54	3.76	0.98
	w/o task + weighted loss	**2.07**	4.51	3.03	3.70	2.48	**2.70**	1.17	4.58	2.77	3.72	2.47	3.91	1.46
	w/o task + weighted sampling	1.80	4.52	3.00	3.68	2.50	2.59	1.17	**4.73**	2.77	3.57	2.57	3.91	1.65
	w/ task CA[3]	**2.07**	4.57	3.05	3.77	2.59	2.39	1.99	4.70	2.81	3.68	2.76	3.90	2.34
	w/ task GC+LC+CA [2, 3, 4]	2.23	4.56	3.21	3.79	2.77	2.18	1.88	4.70	2.78	3.77	1.59	4.02	2.34
	w/ task GC+LC+CA [3]	1.87	4.53	3.04	3.72	2.52	2.65	1.49	4.72	2.80	3.68	2.40	3.86	1.81
	w/ task LC [2]	2.03	**4.64**	**3.24**	**3.82**	**3.09**	2.40	**2.09**	4.71	**3.00**	**3.80**	**2.91**	**4.12**	**2.43**

							Context (SIM↑)							
		Actions (SIM↑)					Roundabout		Highway ramp		Signalized		Unsignalized	
	Model	Stop	None	Lat	Lon	Lat+Lon	RoW	Yield	RoW	Yield	RoW	Yield	RoW	Yield
	ADA	0.28	0.57	0.38	0.47	0.28	0.30	0.21	0.57	0.39	0.41	0.32	0.46	0.16
	ViNet	0.24	0.53	0.37	0.44	0.29	0.31	0.20	0.53	0.37	0.41	0.27	0.45	0.20
SCOUT	w/o task	0.23	0.58	0.41	0.47	0.31	**0.41**	0.16	0.59	**0.41**	0.47	0.23	0.50	0.16
	w/o task + weighted loss	0.26	0.57	0.40	0.48	0.33	0.36	0.20	0.57	0.36	0.48	0.29	0.49	0.18
	w/o task + weighted sampling	0.24	0.58	0.41	0.48	0.34	0.36	0.19	0.60	0.37	0.47	0.32	0.51	0.20
	w/ task CA[3]	0.27	0.60	0.41	0.49	0.34	0.33	0.25	**0.61**	0.38	0.49	**0.33**	0.51	0.22
	w/ task GC+LC+CA [2, 3, 4]	**0.30**	0.60	0.42	0.49	0.35	0.30	0.25	**0.61**	0.38	0.48	0.21	0.51	**0.23**
	w/ task GC+LC+CA [3]	0.23	0.57	0.40	0.48	0.33	0.34	0.22	**0.61**	0.37	0.47	0.26	0.49	0.19
	w/ task LC [2]	0.26	**0.61**	**0.43**	**0.50**	**0.38**	0.33	**0.26**	**0.61**	0.39	**0.51**	0.32	**0.54**	0.22

Notation: Stop — vehicle is stopped, Lat — lateral actions only, Lon — longitudinal only, None — no action, RoW — ego-vehicle has right-of-way. ↑ indicates that larger values are better. Green and red colors indicate the best and worst values.