MEDIRL: Predicting the Visual Attention of Drivers via Maximum Entropy Deep Inverse Reinforcement Learning (Supplementary Material)

Sonia Baee[1], Erfan Pakdamanian[1], Inki Kim[2], Lu Feng[1], Vicente Ordonez[3], Laura Barnes[1]
[1]University of Virginia, [2]University of Illinois at Urbana Champaign, [3]Rice University
sb5ce@virginia.edu, ep2ca@virginia.edu, inkikim@illinois.edu
Lu.feng@virginia.edu, vicenteor@rice.edu, lb3dp@virginia.edu

We propose a novel inverse reinforcement learning formulation using Maximum Entropy Deep Inverse Reinforcement Learning (MEDIRL) for predicting the visual attention of drivers in accident-prone situations. In addition, we introduce EyeCar, a new driver attention dataset in accident-prone situations.

In this document, we provide more details to the main paper and show extra results on ablation studies. We provide further details about the EyeCar dataset in Section S-1, and more details on the architecture and implementation of MEDIRL in Section S-2. We also provide additional results from experiments and ablation studies (Section S-5). You can find the code and dataset in our Github repository[1].

S-1. EyeCar

EyeCar is a new driver attention allocation in accident-prone situations. We follow BDD-A and DADA established and standardized experimental design protocol for collecting in-lab driver attention and create the EyeCar dataset exclusively for various driving tasks which end in rear-end collisions. EyeCar covers more realistic and diverse driving scenarios in accident-prone situations. Unlike DADA-2000, EyeCar captures collisions from a collision point-of-view (POV) perspective (egocentric) where the ego-vehicle is involved in the accident.

Participants: We recruited 20 participants, 5 of them were women, and the rest were men with at least three years of driving experience. You can find more details of our participants in Table S-1. Participants watched all the selected dash-cam videos to identify hazardous cues in rear-end collisions.

Driving videos: We selected 21 front-view videos from the naturalistic driving dataset [2] that included rear-end collisions with high traffic density. The videos were captured in various driving conditions. These conditions contain: traffic conditions (e.g., crowded and not crowded), weather conditions (e.g., rainy and sunny), landscapes (e.g., town and highway), and times of the day (e.g., morning, evening, night). It also contains typical driving tasks (e.g., lane-keeping, merging-in, and braking) ending to rear-end collisions. Each rear-end collision video lasted for 30 seconds, had a resolution of 1280×720 pixels, and had a frame rate of 30 frames per second. All the conditions were counterbalanced among all the participants.

Gender	Age	Driving Experience	Semi-autonomous vehicle	Accident
5 female, 15 male	22-39	9.71(± 5.8)	25%	1%

Table S-1: Detailed information about individuals who participate in the study.

Apparatus: We conducted this study in an experiment booth with controlled lighting. The experiment was designed to maximize the accuracy of the eye tracker to be used as the ground truth for the evaluation of the estimated driver attention allocation. The driving scenes were displayed on a 20-inch monitor with a pixel resolution of 2560 by 1440. Participants were seated approximately 60 cm away from the screen. The head was stabilized with a chin and forehead rest. A Logitech G29 steering wheel is placed in front of the participants who were asked to view the videos by assuming that they were driving a car. To control the lighting and minimize possible shadows, a Litepanels LED-daylight was used.

Eye movements were recorded using the screen-mounted Tobii X3-120 system with a sampling rate of 120 Hz. The eye-tracker was mounted under the screen of the monitor placed in front of the participants. Due to the sensitivity of the eye tracker, the vertical placement of the screen was adjusted such that the center of the screen was at eye-level for each participant. The system had to be calibrated for

[1]https://github.com/soniabaee/MEDIRL-EyeCar

Dataset	Videos	Accidents	Events	Gaze providers	Duration(hrs)	Number of Frames	Annotation type	Gaze pattern (per frame)	Fixations
EyeCar	21	21	rear-end collisions	20	3.5	315K	spatial and temporal	raw and average	1,823,159

Table S-2: The EyeCar dataset detailed information.

Data Source	Feature	Type	Values	Scale
videos	day light	categorical	0/1	per frame
	speed	categorical	slow, normal, fast	per frame
eye's information	distance	categorical	near-reach, medium-reach, far-reach	per fixation in a frame
	fixation duration	integer	[110ms-447ms]	per fixation in a frame

Table S-3: Detailed information about the videos and the fixations on each frame of each video.

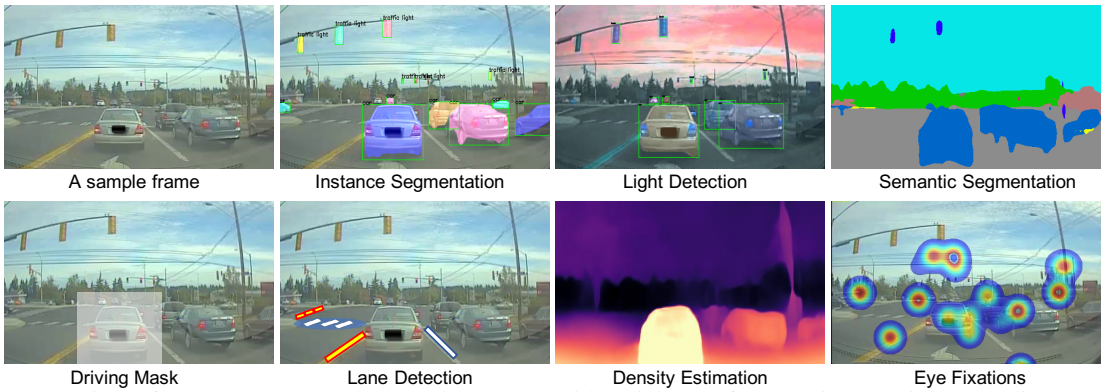

Figure S-1: Overview of our dataset. The dataset also comes with a rich set of annotations: object bounding, lane marking, full-frame semantic, and instance segmentation.

each participant using the Tobii Pro Studio animating nine calibration points. Calibration accuracy was then recorded to be within 0.6 degrees of visual angle for both axes of all participants.

Procedure: Individuals are eligible to participate in this study if they have normal or corrected to normal vision and have at least three years of driving experience. After enrolling in the program, individuals are asked to fill out initial questions consisting of their age, driving experience, gender, whether they have experience with the semi-autonomous vehicle or not, and if they have been involved in any car accidents or not (see Table S-1).

The study had two sessions, and each lasts for 10 ± 2 minutes. To decrease the chance of drivers' fatigue and disengagement, participants watched the first 10 videos in the first session, and then after 5 minutes gap, they watched the other 11 videos in the second session (the whole study takes less than half an hour). The experiment received ethical approval from the University's Institutional Review Board.

During the data collection, we asked participants to 'task-view' the collision videos and were free to fix their eyes on their areas of interest. To incentivize participants to pay attention and play the fall-back ready role in autonomous vehicles, we further modified the experimental design by sitting them in a low-fidelity driving simulator consists of a Logitech G29 steering wheel, accelerator, brake pedal, and eye-tracker.

S-1.1. Data Preprocessing:

Driving videos: EyeCar comes with a rich set of annotations: object bounding, lane marking, full-frame semantic, and instance segmentation (see Figure S-1). You also can see the number of typical instances in each category involved in an accident over all frames of videos in Figure S-2.

Object detection: Understanding the scene is important not only for autonomous driving but the general visual recognition. One of the main elements for a scene is the objects of the scene, therefore locating object is a fundamental task in scene understanding. We provide bounding box and instance mask annotations for each of the frames in EyeCar. The sample of the instances and masks are presented in Figure S-1. In addition, we provide the instance statistics of our object categories in Figure S-2.

Light Detection: Any rear-end collision includes salient stimuli such as brake lights. To detect this type of stimuli, we convert each frame to HSV color space. First, we calculate the average brightness level of each vehicle and traffic light masks. Then, we calculate the brightness anomaly of the selected masks by subtracting their average brightness value from their actual brightness level at each frame. We can determine the pixels corresponding to these anomalies as well as their time of occurrence. Therefore, the location of the target objects and their temporal occurrence interval are annotated in EyeCar.

Depth Estimation: Recognizing the relative distance to the

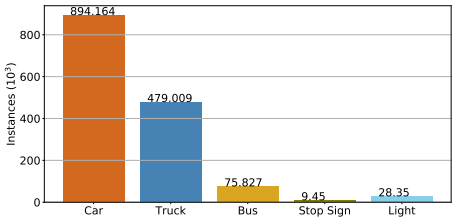

Figure S-2: The distribution of typical instances categories involved in an accident over all frames of the EyeCar videos.

other traffic participants (e.g., the lead vehicle) is crucial for making optimal driving decisions. Therefore, we use a supervised monocular depth estimation model to amplify nearby regions (e.g., distance to a target object) of a driving scene.

Lane Marking and Lane Changes: The lane marking detection is critical for a task-related visual attention allocation of drivers, as an indicator of the type of maneuver. We recognize the left and right lanes of the ego-vehicle by delineating their boundaries. Our lane markings (Figure S-1) are labeled with five main categories: road curb, double white, double yellow, single white, single yellow. The other categories are ignored during evaluation.

Eye information: We employ iMotion to extract the eyes' features such as; pupil size, gaze location, fixation duration, the sequence of fixation, and the start and end time of the fixation points. Moreover, the visual responses and time delay between the onset of the tasks' stimuli (e.g., brake lights) to perceive it by the participants were captured. The abnormal or missing values of these features can lead us to the wrong conclusion, therefore pre-processing of the raw data is necessary for identifying such values and replacing them with linearly interpolated values, outlier treatment, statistical analysis, and data quality (e.g., calibration, exclusion of trials and participants due to poor recording, track loss).

To clean the data, we first extract the missing values of the eyes' features. We employ linear interpolation if the percentage of the missing values is less than 20%. Then, we calculated the abnormal values of features to detect the outliers. We calculated the mean (μ_{feature}) and the standard deviation (σ_{features}) of each feature (zero values are excluded from our calculation) for each participant. Then, we set the low and high thresholds as follows:

$$\text{low threshold} = \mu_{\text{feature}} - 3 \times \sigma_{\text{feature}}$$
$$\text{high threshold} = \mu_{\text{feature}} + 3 \times \sigma_{\text{feature}}$$

Abnormal values are those that values are less than the low threshold and more than a high threshold. In addition to the exclusion criteria described in the main text, we also excluded the sequences with more than 40% abnormal values for eye fixations (see Figure S-4 for a sample of eye fixation sequence). In this way, we decreased the chance of drivers' fatigue and disengagement. We have about 0.005% of sequences with these conditions.

EyeCar dataset: After implemented all exclusion criteria, we selected 416 variable-length eye fixation sequences. EyeCar includes more than 315,000, rear-end collisions video frames. In addition, each video frame comprises 4.6 vehicles on average, making EyeCar driving scenes more complex than other visual attention datasets. GPS recordings in our dataset show the human driver action given the visual input and the driving trajectories. The proportion of high ($65 \leq v$), normal ($35 \leq v \leq 65$), and low ($35 \geq v$)-speed categories are 38%, 39%, and 23%, respectively see Table S-3.

A total of 1,823,159 fixations were extracted from the eye position data, over the 20 subjects. The EyeCar dataset contains 3.5 hours of gaze behavior from the 20 participants. The fixation maps highlight the direction of human drivers' gaze to a salient object when making driving decisions in rear-end collisions. We also provide a raw fixation map of multiple observers as well as an average fixation map of them. We aggregate and smooth the gaze patterns of these independent observers to make an attention map for each frame of the video [1] and simulate the peripheral vision of human [4].

S-2. Implementation

Depth Estimation: The predicted dense depth map D_t at each time step t is combined with the visual feature F_t by the following formula:

$$F_t \oplus D_t = F_t \odot \lambda * D_t + F_t,$$

where $\lambda = 1.2$. This value of $lambda$ parameter helped us to focus on the lead vehicle more than other surrounding vehicles during rear-end collisions. Note that the above equation is equivalent to the main paper's equation which is written in a recurrent form.

State Representation: In our proposed state representation, we try to formulate the visual system mechanism by considering the high-resolution visual information at the eye fixation location (a selected patch in a grid space) and low-resolution visual information outside of the eye-fixation location.

To model the altering of the state representation followed by each fixation, we propose a dynamic state model. To begin with, the state is a low-resolution frame corresponding to peripheral visual input. After each fixation made

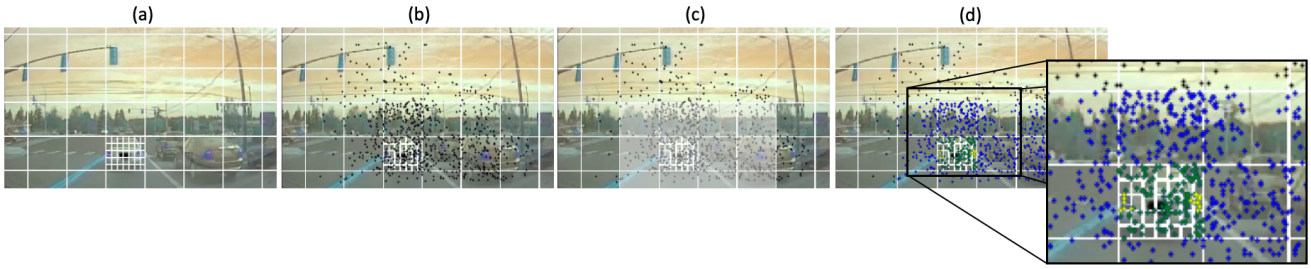

Figure S-3: Illustration of a discretize frame along with gaze location points of all drivers in the EyeCar dataset. Drivers allocate their attention to the driving task-related salient regions of the driving scene. The points show the gaze location of drivers. The black points are out of the task-related regions. The blue points are in the driving mask (the gray area in the frame). The green points are in the lead vehicle bounding box, and the yellow points are in the area of the target object (i.e., braking lights).

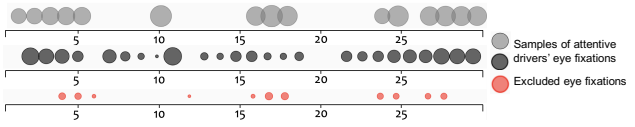

Figure S-4: A sample of attentive drivers' average eye fixations sequence for a given front-view video as well as the excluded sequence. Note that the sizes of the circles are corresponding to the duration of the average eye fixation over the last 30 frames before a collision.

Figure S-5: Examples of MEDIRL generated visual attention allocations on the EyeCar dataset. An attentive driver eye fixation sequences are colored in green, and the model generated is in blue. You can see that MEDIRL mainly focused on driving tasks related to rear-end collisions.

by a driver, we update the state by replacing the portion of the low-resolution features with the corresponding high-resolution portion obtained at each new fixation location. At a given time step t, feature maps H for the original frame (high-resolution) and feature maps L for the blurred frame (low-resolution) are combined as follows:

$$O_{0,1} = L_{0,1}, O_{k+1,t} = E_{k,t} \odot H_t + (1 - E_{k,t}) \odot O_{k,t},$$

where $\odot$ is an element-wise product. $O_{k,t}$ is a context of spatial cues after k fixations. $E_{k,t}$ is a binary map with 1 at current fixation location and 0 elsewhere in a discretize frame. The size of each patch is equal to the smallest size(furthest) of the lead vehicle in the scene 12×17 (about 1° visual angle). To jointly aggregate all the temporal information, we update the next frame by considering all context of spatial cues in the previous frame as follows:

$$O_{k,t+1} = E_{k,t+1} \odot H_{t+1} + (1 - E_{k,t+1}) \odot O_{\mathcal{K},t},$$

where $O_{\mathcal{K},t}$ is visual information after all fixations $\mathcal{K}$ of time step t(previous frame).

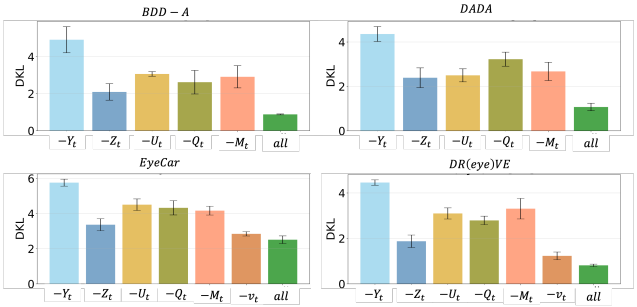

Figure S-6: Ablation study on the proposed state representation. We remove one part by masking out or simply removing from the state representation at each time.

Action Space: We aim to predict the next eye fixation of drivers. It means, we need to predict the pixel location where the driver is looking in the driving scene during accident-prone situations. We discretize each frame into a $n \times m$ grid where each patch matches the smallest size (furthest) of the lead vehicle bounding box (see Figure S-3. The maximum approximation error due to this discretization procedure is 1.27 degrees, visual angel. Action $a_{k,t}$ represents where the focus of attention can move at fixation k of time step t. The policy selects one out of $n * m$ patches in a given discretize frame. The center of the selected patch in the frame is a new fixation. Finally, the changes (Δ_x, Δ_y) of the current fixation and the selected fixation define the action space A_t: {left, right, up, down, focus-inward, focus-outward, stay}, as shown in Figure **??** which has three degrees of freedom (vertical, horizontal, diagonal). We also excluded the patches that have no visits or less than five visits for computational efficiency. It should be noted that we did not pre-defined the radius of the direction for the agent. Therefore, the agent has the freedom to pick any patch among the created ones.

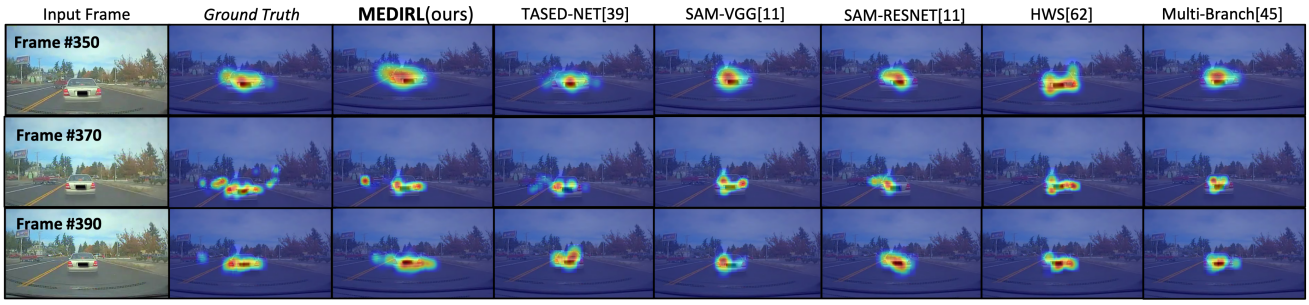

Figure S-7: Predicted driver attention in a braking task for each compared model and MEDIRL (video #17). They all trained on BDD-A. MEDIRL can learn to detect most task-related salient stimuli (e.g., traffic light, brake light). The redder color indicates the expectation of a higher reward for fixation location.

Driving task and ego-vehicle speed: We embed the task in our model by one-hot encoding maps which spatially repeat the one-hot vector. Therefore, we concatenate the task embedding with other features in our proposed state representation to have a task-dependent bias term for every convolutional layer. We then add another fully-connected layer to encode the current speed of the ego-vehicle and concatenate the state with the speed vector.

Visual attention allocations: The eye fixation location is generated from the probability map that MEDIRL has produced. We also applied Inhibition-of-Return to decrease the likelihood that a previously inspected (possibly salient) region in the scene will be re-inspected, thereby encouraging visual attention toward the next salient region in a driving scene. Therefore, MEDIRL generates a new spatial probability map at every step.

Figure S-5 demonstrates the generated fixation location in a single frame of different driving videos by MEDIRL. MEDIRL mainly focused on driving tasks related to rear-end collisions.

Maximum Entropy: To learn the policies, we maximize the joint posterior distribution of visual attention allocation demonstrations $\Xi = \{\xi_1, \xi_2, ..., \xi_q\}$, under a given reward structure and of the model parameter, θ, across I. For a single frame and given visual attention allocation sequence $\xi_q = \{(s_1, a_1), ..., (s_\tau, a_\tau)\}$ with a length of $|\tau|$, the likelihood is:

$$\mathcal{L}_\theta = (1/\Xi) \sum_{\xi^i \in \Xi} logP(\xi^i, \theta),$$

, where $P(\xi^i, \theta)$ is the probability of the trajectory ξ^i in demonstration Ξ. In each iteration j of maximum entropy deep inverse reinforcement learning algorithm, we first evaluate the reward value based on the state features and the current reward network parameters θ_j. Then, we determine the current policy, π_j, based on the current approximation of reward, R_j and transition matrix (i.e., the outcome state-space of a taken action), $\mathcal{T}$. Therefore, we can benefit from the maximum entropy paradigm, which enables the model to handle sub-optimal behavior as well as stochastic behavior of experts, by operating on the distribution over possible trajectories [6, 5].

Principle of Maximum Entropy [3] demonstrates that the best distribution overcurrent information is one with the largest entropy. Maximum Entropy also prevents issues with label bias which means portions of state space with many branches will each be biased to be less likely, and while areas with fewer branches will have higher probabilities (locally greedy). Maximum Entropy gives all paths equal probability due to equal reward and uses a probabilistic approach that maximizes the entropy of the actions, allowing a principled way to handle noise, and it prevents label bias. It also provides an efficient algorithm to compute empirical feature count, leading to a state-of-the-art performance at the time. This process maximized total reward, even over the short period of time ($0.6\pm$ 0.2 seconds) that our attentive drivers detect the target objects (brake light) in rear-end collisions.

S-3. More Evaluations

S-3.1. Training and Testing on EyeCar

We train and test MEDIRL on EyeCar. To be able to do it, we used leave-one-out cross-validation (one video as test) and obtained the following results: (CC: 0.85, s-AUC: 0.8, KLD: 0.92), (CC: 0.83, s-AUC: 0.74, KLD: 1.58), (CC: 0.79, s-AUC: 0.77, KLD: 1.29), on lane-keeping, merging-in, and braking driving tasks, respectively.

S-3.2. Training on EyeCar and Testing on Benchmarks

We report the results of training on EyeCar and testing on each benchmark for each driving task.

Dataset \ Task	Merging-in			Lane-keeping			Braking		
	CC↑	s-AUC↑	KLD↓	CC↑	s-AUC↑	KLD↓	CC↑	s-AUC↑	KLD↓
DR(eye)VE	0.88	-	0.89	0.91	-	0.70	0.88	-	0.81
BDD-A	0.92	0.89	0.87	0.94	0.94	0.82	0.96	0.90	0.86
DADA-2000	0.77	0.71	1.06	0.93	0.72	0.92	0.85	0.88	0.99

Table S-4: The results of training on EyeCar and testing on each benchmark for each driving task.

S-4. Qualitative Comparison

We provide a qualitative comparison of MEDIRL against other models in Figure S-7. It shows that MEDIRL can reliably manage to capture the important visual cues in a braking task in the case of a complex frame. In contrast, nearly all other models partially capture the spatial cues and predict attention mainly towards the center of the frame, thereby ignoring the target and non-target objects (i.e., spatial cues).

S-4.1. Challenging Environment

We also evaluate MEDIRL performance under extreme weather conditions such as foggy weather. The BDD-A dataset includes severe weather (e.g., foggy and snowy). We report the results of MEDIRL trained and tested on BDD-A in Table 2 of the paper, showing MEDIRL surpasses almost all the models. We further compared MEDIRL with TASED-NET exclusively on the **foggy** videos extracted from BDD-A.

Weather	Methods	CC↑	s-AUC↑	KLD↓
Foggy	TASED-NET	0.64	0.53	2.12
	MEDIRL	0.72	0.62	1.45

Table S-5: Evaluating MEDIRL in a challenging driving environment (i.e., foggy).

Despite the foggy weather conditions, the results highlight that MEDIRL still performs better under all evaluation metrics. The results will be added to supplementary material. Generally, MEDIRL is more sensitive to false-negative prediction, leading to significant improvement in KLD.

S-5. Ablation Studies

Figure S-6 shows the ablation study of the full state representation on different test datasets. We can see that the most important feature categories were semantic/instance (Y_t), followed by target object (U_t), and types of driving tasks (Q_t). The depth map features (Z_t) are also beneficial for the model performance whereas ego-vehicle speed (v_t) weakly impacted model performances. The results confirm the incorporation of low and mid-level visual cues, and driving-specific visual features.

We study the benefits of each component of MEDIRL by running ablation experiments (see Table S-5) with the trained model on the BDD-A dataset and tested on EyeCar and BDD-A test dataset. We employed our general scene features and driving-related features to have a rich state representation for our proposed MEDIRL model. To understand the contribution of each component, we removed the maps of each group one at a time and compared the corresponding performance of the model. MEDIRL is not restricted to these backbones and could potentially incorporate new and more robust networks as submodules.

Ablated versions \ Dataset	EyeCar			BDD-A		
	CC ↑	KLD ↓	F_β ↑	CC ↑	KLD ↓	F_β ↑
-general features	0.36	3.55	0.21	0.41	3.51	0.27
-driving-related features	0.69	2.21	0.30	0.60	2.07	0.39
MEDIRL	**0.84**	**0.81**	**0.61**	**0.89**	**0.88**	**0.78**

Table S-6: Ablative study of MEDIRL using different combination of modules. The model used here is trained on BDD-A dataset and tested on EyeCar and BDD-A test dataset.

References

[1] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. How do drivers allocate their potential attention? driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):2146–2154, 2019.

[2] Thomas A Dingus, Jonathan M Hankey, Jonathan F Antin, Suzanne E Lee, Lisa Eichelberger, Kelly E Stulce, Doug McGraw, Miguel Perez, and Loren Stowe. *Naturalistic driving study: Technical coordination and quality control*. 2015.

[3] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[4] Anwesan Pal, Sayan Mondal, and Henrik I Christensen. "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11883–11892, 2020.

[5] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.

[6] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.