

AEGIS: Human Attention-based Explainable Guidance for Intelligent Vehicle Systems

Zhuoli Zhuang
University of Technology Sydney
Sydney, NSW, Australia
zhuoli.zhuang@student.uts.edu.au

Cheng-You Lu
University of Technology Sydney
Sydney, NSW, Australia
cheng-you.lu@student.uts.edu.au

Yu-Cheng Fred Chang
University of Technology Sydney
Sydney, NSW, Australia
fred.chang@uts.edu.au

Yu-Kai Wang
University of Technology Sydney
Sydney, NSW, Australia
yukai.wang@uts.edu.au

Thomas Do
University of Technology Sydney
Sydney, NSW, Australia
thomas.do@uts.edu.au

Chin-Teng Lin
University of Technology Sydney
Sydney, NSW, Australia
chin-teng.lin@uts.edu.au

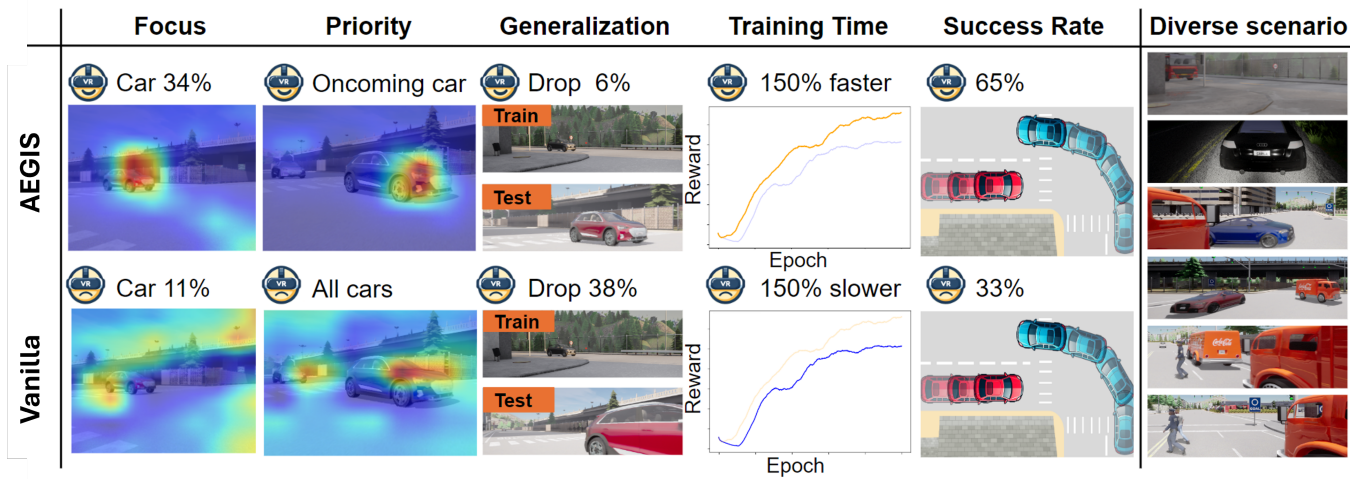


Figure 1: Comparison between AEGIS and RL without human attention guidance (Vanilla). AEGIS enhances focus on the regions of interest, prioritizes important objects, achieves robust performance in unseen environments, accelerates training speed, and achieves better performance for collision avoidance in the left-turn scenario. We analyze and benchmark our approach in six challenging scenarios: car-following, left-turn, and four diverse occlusion scenes.

ABSTRACT

Improving decision-making capabilities in Autonomous Intelligent Vehicles (AIVs) has been a heated topic in recent years. Despite advancements, training machine to capture regions of interest for comprehensive scene understanding, like human perception and reasoning, remains a significant challenge. This study introduces a novel framework, Human Attention-based Explainable Guidance for Intelligent Vehicle Systems (AEGIS¹). AEGIS uses a pre-trained human attention model to guide reinforcement learning (RL) models to identify critical regions of interest for decision-making. By

¹In Greek mythology, the Aegis is a protective shield associated with Zeus and Athena, symbolizing guidance and protection.

collecting 1.2 million frames from 20 participants across six scenarios, AEGIS pre-trains a model to predict human attention patterns. The learned human attention² guides the RL agent's focus on task-relevant objects, prioritizes critical instances, enhances robustness in unseen environments, and leads to faster learning convergence. This approach enhances interpretability by making machine attention more comparable to human attention and thus enhancing the RL agent's performance in diverse driving scenarios. The code is available in <https://github.com/ALEX95GOGO/AEGIS>.

CCS CONCEPTS

• **Human-centered computing** → HCI theory, concepts and models; • **Computing methodologies** → Vision for robotics; Markov decision processes; Sequential decision making.

KEYWORDS

Eye-tracking, Virtual reality, Human-centered computing

²We refer to the prediction of the human attention model as learned human attention.

ACM Reference Format:

Zhuoli Zhuang, Cheng-You Lu, Yu-Cheng Fred Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. 2025. AEGIS: Human Attention-based Explainable Guidance for Intelligent Vehicle Systems. In *CHI Conference on Human Factors in Computing Systems (CHI '25), April 26–May 1, 2025, Yokohama, Japan*. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3706598.3713779>

1 INTRODUCTION

Deep learning methods for autonomous intelligent vehicles (AIVs) have been rapidly developing over the past two decades. However, these models raise concerns about safety of AIVs due to the non-transparency of the decision-making process [38]. Despite recent attempts to improve the transparency of deep learning models for AIVs [65, 66], these deep learning models lack explicit scene understanding and reasoning, the two basic cognitive skills of human drivers. Specifically, the decision-making of human drivers starts with visual perception via eye movements. In this way, task-relevant visual information can be extracted to better understand the scene. Moreover, human visual perception and processing are heavily influenced by top-down cognitive control and prior knowledge [21] that allocates human attention to task-relevant objects [19, 29, 43]. Such selective focus of attention to task-relevant objects hence supports human decision-making [72, 82]. For instance, experienced drivers can easily identify and attend to task-relevant information (e.g., cars or pedestrians in the front) for driving performance and ignore salient but irrelevant information. The reciprocal link among eye movements, visual perception, attention, and decision-making in humans suggests that an AIV needs to be trained with a sophisticated reasoning mechanism similar to that of humans. Motivated by this, prior works [23, 54, 87] have aimed to model drivers' visual attention in driving contexts, suggesting where drivers need to attend. However, the integration of this human attention prediction system into autonomous driving systems to help AIV decision-making and reasoning receives scant attention. Hence, it remains an open question whether using human attention as training guidance for machine decision-making has an effect on reinforcement learning (RL) for AIVs.

Recent decision-making methods for AIVs have not incorporated human attention and have mainly adopted two end-to-end deep learning approaches: imitation learning (IL) [57] and deep reinforcement learning (DRL) [28]. IL aims to learn driving strategies from an expert, such as a human driver, by mimicking their control actions in similar situations [12, 13, 45, 57, 66, 81]. However, IL faces a notable limitation: its vulnerability to the distribution shift problem [10]. During training, an IL model learns from a specific distribution of states and actions from the expert. This means that IL models usually do not explore sufficiently in scenarios where unforeseen failures occur, hindering their ability to respond correctly under adverse conditions [24, 77, 89, 96]. Unlike IL, DRL mitigates the distribution shift problem because it enables agents to learn through trial and error by rewarding chosen actions and allowing them to adapt to new environments [49]. However, DRL has two limitations. One major limitation of DRL is the substantial data and time requirements for convergence. This is due to the sparse reward signals and the RL agent needing extensive exploration to learn effective policies [36]. Another limitation of DRL, which it shares

with IL, is the lack of explainability inherent in the deep neural networks used by both approaches. These networks map perception to actions in an opaque manner, making the decision-making process of machines nontransparent [53, 90].

Current AIV models have been reported to be involved in several accidents [2, 6, 73]. The lack of interpretability in these models has raised public concerns about the need for explainable decision-making systems [3] and even calls for legislation [1]. This lack of transparency makes it difficult to understand the decisions that lead to these worrying accidents. Unlike earlier approaches focusing mainly on performance, our method emphasizes human-centric computing with a focus on interpretability. Interpretability is the ability to explain or provide meaning in terms understandable to humans [5]. The interpretability of our framework has two main aspects. First, aligning machine attention with human attention makes the model's behavior more understandable to humans. Human attention reflects how people focus on important information, and with proper guidance, visual search can be influenced by top-down cognitive control [83]. Second, our network uses a self-attention mechanism to represent machine attention. This enhances interpretability by helping the model learn to focus on the most important features of the input [47]. Additionally, we conducted a survey of 80 respondents (see Fig. 14), which shows that our model's attention and decision-making process are easier to understand than those of existing approaches.

This paper introduces Human Attention-based Explainable Guidance for Intelligent Vehicle Systems (AEGIS) as a solution to the interpretability issue and a response to the open question of the effectiveness of human attention guidance in RL for AIVs. In contrast to previous human-guided RL approaches [84–86] that have primarily provided action guidance, AEGIS leverages human attention to guide the RL agent on the latent code of action. The proposed attention-based guidance enables the RL agent to learn task-relevant objects, thereby improving its generalizability (see Tab. 4 and Tab. 5). To acquire human drivers' attention, we collected large-scale eye-tracking data using a realistic VR driving simulator (see Sec. 3.1). We recorded the active engagement of drivers and propose a unique framework that incorporated these human attention data into the training of DRL for autonomous driving tasks. Our dataset includes 20 participants engaged in two challenging scenarios (see Fig. 2 and Fig. 3) and four diverse occlusion scenarios inspired by [66] (see Fig. 4), yielding a total of six scenarios, 1.2 million frames and 600 minutes of driving data. To the best of our knowledge, this is the largest eye-tracking dataset collected using an immersive method with a VR headset and physical simulator (see Fig. 2 and Tab. 1). Leveraging this dataset, we craft an explainable driving model that utilizes human attention predicted from a model pre-trained on the eye-tracking dataset to guide the model's self-attention layer. The pre-trained human attention model, while simplistic, eliminates the need for eye-tracking data during inference. Compared with the traditional RL without human attention integration, our method enhances focus on crucial objects and increases the DRL training speed (see Fig. 1). Moreover, AEGIS ensures the similarity of machine attention with human attention, thereby increasing agent robustness in new scenes. Although AEGIS prioritizes explainability, which is an important topic for AI safety in both research and industry, performance analysis and attention

visualization confirm that integrating human attention significantly benefits agents. The contributions of this work are four-fold:

- A novel and largest in-lab eye-tracking dataset collected using a realistic VR driving simulator, capturing drivers' active engagement across six diverse scenarios.
- The incorporation of human attention guidance aligns machine attention more closely with learned human attention, improving the RL agent's explainability.
- The proposed AEGIS framework involves a human-attention guidance mechanism to enable the RL agent to learn task-relevant objects.
- Comprehensive analysis shows that AEGIS significantly improves training efficiency, robustness in unseen scenes, and overall performance.

2 RELATED WORK

2.1 Human action-guided RL

RL has shown great success in complex tasks such as playing games that can surpass human players in Atari [49] and Go [70]. However, the training efficiency of RL is hindered by its requirement for extensive interactive sessions and a propensity to converge on suboptimal solutions due to insufficient prior knowledge [86]. Prior works [46, 84–86] have attempted to increase the sample efficiency and performance of RL via human guidance, which provides necessary prior knowledge with human demonstrations to the RL during training. Suboptimal action replacement [46, 84–86], reward shaping [42, 85], and replay buffer prioritization [84–86] have been proposed to accelerate RL training. In human-in-the-loop RL frameworks with suboptimal action replacement [46, 84–86], humans can intervene and replace RL actions with their actions. This is based on the assumption that humans can correct suboptimal RL behaviors when necessary. By doing so, training speed and performance could be improved. Reward shaping is another common method in human-in-the-loop RL, where a negative reward penalizes RL when human actions have deviated from RL actions [42, 85]. Wu et al. [84–86] used prioritized experience replay mechanisms [61] to prioritize human demonstrations based on Q value difference between human and RL actions. However, these studies relied on human actions to replace suboptimal RL agent actions and required substantial human demonstrations, and depended that humans remain available throughout the RL training process. Moreover, these works still leave the decision-making process a black box. Unlike these works, we provide human attention knowledge to the latent space of the RL agent via the attention mechanism and thus enhance the interpretability of our RL agent.

2.2 Machine attention of RL

Extensive research has been conducted to explain the black box behavior of the neural network of the RL agent [27, 31, 50–52, 76, 79, 91, 97]. Joo et al. [32] employed the Gradient-weighted Class Activation (Grad-CAM) [63] method to explain the area of an image that is related to the decision. Similarly, Greydanus et al. [27] showed that the Atari agent could be explained via an occlusion-map method. Nevertheless, these post-hoc explanation methods can be computationally expensive. For example, the occlusion-map approach outlined in [27], requires 256 model inferences to explain

a single 80×80 Atari game image, rendering it impractical for real-time applications. Additionally, as these explanations are generated after model training, they are often approximations and may overlook critical aspects of the model's decision-making process [48].

An alternative approach to achieve interpretability in RL involves the development of inherently transparent models. For instance, Zambaldi et al. [91] developed a relationship module that leveraged the self-attention mechanism [78] to explain the focus area in the Starcraft II environment. Similarly, the self-attention mechanism was applied in [31, 50, 52, 76] to augment agent representation and interoperability. Unlike these works, the self-attention layer of RL policy network in our approach is guided by human attention and explored in AIVs.

2.3 Human attention for RL

Unlike machine attention, human attention can integrate diverse sensory inputs into a coherent understanding and involve sophisticated cognitive processes, such as attention allocation, that contribute to decision-making [39, 44]. While some imitation learning (IL) studies [11, 60, 92–95] have incorporated human attention, to the best of our knowledge, no RL research has investigated the integration of human attention guidance within an RL framework for visuomotor control tasks, such as playing Atari games. Notably, one RL work [29] compared human and machine attention in RL agents for Atari games but did not integrate human attention in the RL framework.

In driving studies, various attempts have been made to predict the drivers' attention in in-car [4, 54] and in-lab datasets [7, 16, 23, 26, 87]. However, the in-car dataset cannot collect repeated scenarios for different drivers, and the in-lab dataset often requires the driver to review a video without active engagement of the task. To solve these issues, we employ a realistic VR simulator for active engagement and realistic data collection. With these datasets, several convolutional neural networks (CNNs) [23, 54, 87] have been proposed to predict drivers' attention. Despite these attempts, a notable gap remains in applying human visual attention as the prior knowledge for training the RL agent in driving tasks. Unlike these works which focused on accurately predicting gaze position, our work focuses on exploring the effect of learned human attention on the RL framework for the control process of AIVs.

3 DATASET

3.1 Eye-tracking data collection

Author Statement: As the authors of this dataset, we take full responsibility for its integrity and any issues related to data rights or ethical standards. We confirm that the collection and use of data comply with relevant regulations, and all participants were compensated at an hourly rate higher than the country's minimum wage. The dataset is shared under an MIT license, allowing use, redistribution, and citation that align with the license term.

Dataset Description: To simulate a realistic driving experience, we integrate VR technology using an HTC VIVE Pro Eye VR headset (resolution: $2 \times 1440 \times 1600$, refreshing rate: 90 Hz, eye-tracking accuracy: 0.5° - 1.1°) and an open-source CARLA simulator [20, 71]. The study includes 20 participants with normal or correct-to-normal

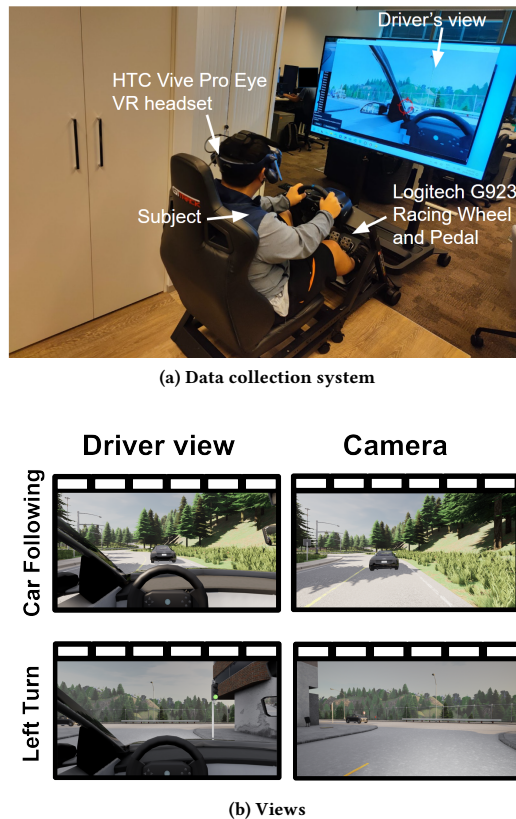


Figure 2: The dataset collection environment. The HTC VIVE Pro Eye VR headset and Logitech G923 Racing Wheel and Pedal give the subject a more realistic driving experience.

vision (age: 24.1 ± 4.8 years) who were provided informed written consent before participation. The participants have an average driver's license possession of 4.9 ± 3.9 years. The participants filled out the driver skill inventory (DSI) form [40] prior to the experiment, with a perceptual-motor skills score of 4.1 ± 0.3 and a safety skills score of 3.7 ± 0.3 (scale: 1-5). The university's human research ethics committee has approved the protocol for involving human participants. The participants signed the consent form prior to the experiment, and were fully aware of the purpose of the data collection. The participants were instructed to complete the driving tasks to the best of their ability, with their eye movements being recorded via the VR headset. Each driving scenario lasts about five minutes, with six diverse scenarios in total, and the total dataset collection time is 600 minutes. The dataset allows training of a human attention model, and no further data collection is required in RL training.

In the experiment, participants could control the vehicle through a racing wheel and pedals, closely mimicking real-life driving (see Fig. 2). This arrangement enables us to collect accurate eye-tracking information while ensuring participants are actively involved. Moreover, it permits the replication of the same scenarios with the same initial settings and exo agent configurations for different individuals

to develop a model that can be applied across individuals. Our data collection pipeline represents a significant advantage by combining scenario-level replication, immersive VR over 2D screens, and active participant control for more realistic and consistent data collection. These represent a distinct advantage over previous eye-tracking driving datasets (see Tab. 1).

To date, the only large on-road dataset is the DR(eye)VE [4] dataset. However, the major limitation of the on-road dataset is that the traffic conditions and exo agents' behaviors are not replicable for different drivers [35]. Several studies [26, 87] have leveraged this dataset and collected eye-tracking data from experienced drivers by asking them to watch the driving videos and imagine as if they are driving. However, the previous approach results in passive observation, which alters gaze distribution compared with actual driving due to the lack of vehicle control and task mindset by participants [87]. Moreover, screen-based displays offer a limited field-of-view (FOV) and a less realistic driving experience. Similarly, 3DDS [8], C42CN [75], DADA-2000 [22], LBW [33], DrFixD(night) [16], and CoCAtt [67] record gaze data in a low-fidelity screen-based driving simulator. To solve the issue of screen-based simulators, we collect our AEGIS dataset, the first large eye-tracking driving dataset recorded from a VR driving simulator. This setting can provide a 360° FOV with much higher fidelity to real driving [34]. Additionally, our dataset includes 1.2M frames collected from 20 participants, with the location of their eye gaze recorded on each frame. These frames are recorded from three cameras, including RGB, semantic, and depth cameras. In addition to visual data, our dataset includes comprehensive vehicle control information, such as throttle and brake inputs, as well as vehicle dynamics data (e.g., speed and acceleration metrics). Similar to DADA-2000 and BDD-A [22, 87], we focus on critical scenarios that require immediate decision-making. We then train a human attention network with the dataset (see Sec. 4.3).

3.2 Scenario design



Figure 3: Car Following: The ego vehicle must avoid collisions with the car ahead by controlling the throttle and brake, ensuring it continues to follow the lead car. Left Turn: The ego vehicle must accurately time its left turn to avoid collisions with vehicles proceeding straight by controlling the throttle and brake.

Our method is validated in six scenarios: car following, left turn (see Fig. 3), and four diverse occlusion scenarios [66] (see Fig. 4). The RL agent is trained and evaluated in different towns (see Fig. 5) to enhance the challenge and diversity, so the agent must develop a robust representation to avoid collisions in unseen scenes. A PID controller controls the lateral movement following [85], while the RL agent concentrates on throttle and brake control, with its action $a_t \in [-1, 1]$.

Car-following scenario: Inspired by [80], our car-following task

Table 1: Comparison between AEGIS dataset and other eye-tracking driving datasets, updated according to [35]. Our eye-tracking dataset is the largest dataset, adopting a realistic VR driving simulator. The following abbreviations are used in the table. Camera: S - scene facing camera, RGB - 3-channel image, sem - semantic segmentation mask. Frame counts marked with * are estimated based on the lengths of the videos and camera frame rate.

Dataset	Active control	Vehicle data	Camera	Hazards	View	#subjects	#frames
AEGIS (Ours)	+	+	<i>srgb,depth,sem</i>	+	VR	20	1.2 M
DrFixD(night) [16]	-	-	<i>Srgb</i>	-	screen	31	67K*
LBW [33]	-	-	<i>Srgb,depth</i>	-	screen	28	123K*
CoCAtt [67]	+	+	<i>Srgb</i>	-	screen	11	17K*
MAAD [26]	-	-	<i>Srgb</i>	-	screen	23	60K
TrafficGaze [17]	-	-	<i>Srgb</i>	-	screen	28	77K*
DADA-2000 [22]	-	-	<i>Srgb</i>	+	screen	20	658K
DR(eye)VE [4]	+	-	<i>Srgb</i>	-	on-road	8	555K
BDD-A [87]	-	+	<i>Srgb</i>	+	screen	45	378K*
C42CN [75]	+	-	<i>Srgb</i>	+	screen	68	-
TETD [19]	-	+	<i>Srgb</i>	-	screen	20	100
3DDS [8]	+	-	<i>Srgb</i>	-	screen	10	192K

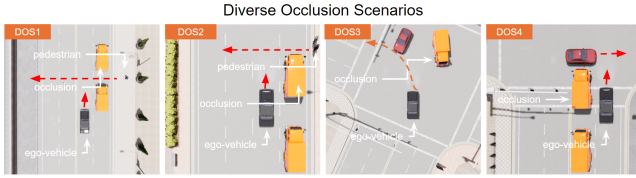


Figure 4: Diverse occlusion scenes. The ego vehicle must control the throttle and brake to prevent collisions with occluded objects, such as pedestrians and cars.



Figure 5: Training and Testing scene. The car-following model is trained in Town 7, characterized by its rural setting and narrow roads, and then tested in Town 4, a mountainous area featuring highways. The left-turn model is trained in Town 1, a small town, and then tested in Town 5, a town with bridge and cross junctions.

resembles the adaptive cruise control (ACC) for the traffic congestion situation. The ego vehicle must follow a lead vehicle within the same lane, which moves at speeds up to 8 m/s and may brake abruptly. The ego vehicle should brake swiftly to avoid collisions while maintaining a close enough distance for effective following. This scenario design is challenging, and many participants found it harder than the left-turn scenario and some participants stated that the car-following scenario demanded sustained attention over a longer duration to monitor the lead vehicle.

Left-turn scenario: Adopting from [85], our left-turn scenario requires the ego vehicle to perform a left turn at an intersection,

ensuring no collisions with oncoming vehicles. These vehicles, moving at speeds between 3m/s and 5m/s, act aggressively and do not give way to the ego vehicle. The ego vehicle should blend into the traffic at the right moment, aiming to reach a goal point swiftly.

Occlusion scenarios: The four occlusion scenarios follow the public Drive in Occlusion Simulation (DOS) benchmark [66]. The benchmark consists of 100 diverse cases with oncoming vehicles or pedestrians occluded by other vehicles. DOS1 and DOS2 involve scenarios where a pedestrian, occluded by cars, is walking across the road. In DOS1, the ego vehicle can avoid a collision by detecting the pedestrian early, before the occlusion occurs. In DOS2, the ego vehicle must adapt by reducing its speed when approaching the intersection, as the cars completely obscure the pedestrian from view beforehand. In DOS3, the ego vehicle should slow down when passing the intersection to ensure safety. In DOS4, the ego vehicle can identify the oncoming traffic through the gaps between the obstructing vehicles.

4 METHODS

4.1 RL problem definition

Our task is a goal-oriented collision-avoidance task that involves controlling the ego vehicle’s brake and throttle strength. The Markov Decision Process (MDP) for this task can be represented as $\{S, A, P, R\}$. At a time step t , the agent (e.g., ego vehicle) observes the state s_t from all possible states S and outputs an action a_t from the action space A . Given this action, the environment transitions to a new state $s_{t+1} \in S$ according to the transition probability matrix P and provides a reward r_t . The reward is determined by the reward function $R(\cdot|s, a) : S \times A \rightarrow r$. The goal of reinforcement learning (RL) is to find an optimal policy π that maximizes the expected return $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where $\gamma \in (0, 1)$ is the discount factor.

State space: The state space $S = \{s_t \mid s_t = \{I_{t-2}, I_{t-1}, I_t\}, t \in \mathbb{N}\}$ is defined as the set of all sequences of three consecutive segmentation images, where $I \in \mathbb{R}^{h \times w \times 1}$ represents a single segmentation image captured by the camera sensor in the CARLA simulator [20].

In the car-following scenario, the camera faces forward. For the left-turn scenario, the looking vector is adjusted 30° to the left to mimic a driver's perspective. In the occlusion scenarios, we employ three cameras with a front-facing camera and two side cameras facing 60° to the left and right due to pedestrians coming from the sidewalk.

Action space: The action space $A = \{a_t \mid a_t \in [-1, 1], t \in \mathbb{N}\}$ represents the set of possible longitudinal control commands for the ego vehicle. A value of -1 indicates the maximum braking force, whereas 1 denotes the maximum throttle force.

Reward: The primary objective of the RL agent is to avoid collision while completing tasks efficiently. To encourage such behavior, we design a reward function as follows:

$$\begin{aligned} r_t &= R(\cdot | s_t, a_t) \\ &= r_{goal}(s_t \in C_{goal}) + r_{collide}(s_t \in C_{collide}) \\ &\quad + \omega * r_{idle}(s_t \in C_{idle}) + \delta * r_{gap}(s_t) \end{aligned} \quad (1)$$

where C_{goal} , $C_{collide}$, C_{idle} and r_{gap} represent completion, collision, idle status, and time gap, respectively. The agent is awarded a large positive reward $r_{goal} = 100$ when it reaches the goal and receives a large negative reward $r_{collide} = -100$ when a collision happens, a negative reward $r_{idle} = -1$ when the ego vehicle stops moving. The time gap is the time it takes for the ego vehicle to reach the current position of the lead vehicle. Similar to [15], we define a reward term r_{gap} for the car-following scenario is as follows:

$$r_{gap} = \begin{cases} T_{gap} & \text{if } T_{gap} \in [1, 2] \\ \max(-1/T_{gap}, -10) & \text{if } T_{gap} < 1 \\ \max(-T_{gap}, -10) & \text{if } T_{gap} > 2 \end{cases} \quad (2)$$

where $T_{gap} = Dis/V_{ego}$, where Dis is the distance between the ego vehicle and the lead vehicle, and V_{ego} is the speed of the ego vehicle. This reward design encourages the ego vehicle to maintain a safe and optimal distance from the lead vehicle.

4.2 Policy network

We design an interpretable policy network leveraging the self-attention mechanism (see Fig. 6). The model processes three consecutive semantic segmentation images as the input state $s_t \in \mathbb{R}^{h \times w \times 3}$, where h and w denote the height and width of the input images, respectively. The policy network outputs machine attention $M \in \mathbb{R}^{h/16 \times w/16}$, an action $a_t \in [-1, 1]$, which represents throttle and brake control, and time-to-collision (TTC) within the range $[0, 5]$, in a unified framework. A shallow CNN encodes the semantic segmentation images into a feature map $F \in \mathbb{R}^{h/16 \times w/16 \times f}$. This feature map is subsequently flattened into $N \in \mathbb{R}^{n \times f}$, where $n = \frac{h}{16} \times \frac{w}{16}$. The flattened representation is then projected into the query Q , key K , and value V matrices via fully-connected layers f_Q , f_K , and f_V , respectively:

$$Q = f_Q(N), \quad K = f_K(N), \quad V = f_V(N) \quad (3)$$

Here, $Q, K, V \in \mathbb{R}^{n \times d}$, where d represents the dimensionality of the latent space for each token. These matrices are utilized to compute

self-attention as follows:

$$\begin{aligned} \text{SelfAttention} &= \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \\ &= \text{MachineAttention}(Q, K)V \\ &= MV \end{aligned} \quad (4)$$

In this formulation, $QK^T \in \mathbb{R}^{n \times n}$ encodes the pairwise attention scores between all the elements, which are normalized by \sqrt{d} to improve numerical stability. The attention scores are normalized using the softmax function, converting them into a probability distribution where the weights sum to 1 for each element. The resulting matrix $M \in \mathbb{R}^{n \times n}$, referred to as machine attention, determines the relative importance of each element. These attention weights are applied to V , producing the final output of the self-attention mechanism.

The output from the self-attention layer is flattened and inputs into multilayer perceptrons (MLPs) to predict action $a_t \in [-1, 1]$ where -1 represents maximum braking, and 1 indicates maximum throttle, and TTC:

$$TTC = \text{clip}(Dis/(V_{ego} - V_{front}), 0, 5) \quad (5)$$

where Dis is the distance to the closest vehicle, V_{ego} is the speed of the ego vehicle, and V_{front} is the speed of the closest vehicle. We clip TTC to the range $[0, 5]$ s to encourage the agent to concentrate on critical situations.

To regularize learning, we employ Kullback-Leibler Divergence (KL) \mathcal{L}_{kl} to align machine attention with learned human attention from the pre-trained model in Sec. 4.3, and the mean square error \mathcal{L}_{mse} to align the predicted TTC with the ground truth. Overall, the new loss is:

$$\mathcal{L}_{total} = \mathcal{L}_\pi + \alpha * \mathcal{L}_{kl} + \beta * \mathcal{L}_{mse} \quad (6)$$

where \mathcal{L}_π is the loss of the original RL policy network, which depends on the RL method used, and \mathcal{L}_{kl} and \mathcal{L}_{mse} are auxiliary losses used for regularization. Notably, we only supervise the machine attention with learned human attention in the first 500 steps of RL training, allowing the agent to further refine its attention after that.

4.3 Human attention network

In this study, we utilize a CNN model developed in [18], which is based on a lightweight U-Net [59] structure. To mitigate the variability and potential distraction arising from individual differences and ensure consistency in our results (see Fig. 16), we employ a human attention network that can predict the general pattern of humans' focus of attention. This choice is motivated by the model's balance of computational efficiency and speed, making it suitable for real-time applications. The inference time of one image is 0.005s. The ground truth of the model is the human attention obtained from the gaze position of the previous ten consecutive frames, similar to the method in [54]. The discrete gaze positions are converted into continuous distribution via a 2D Gaussian filter with σ that is equivalent to one visual degree [41], which is the visual field of the foveola, the high-acuity region of the retina at the center of gaze [56]. The human attention model is trained using binary cross entropy (BCE) loss until converged.

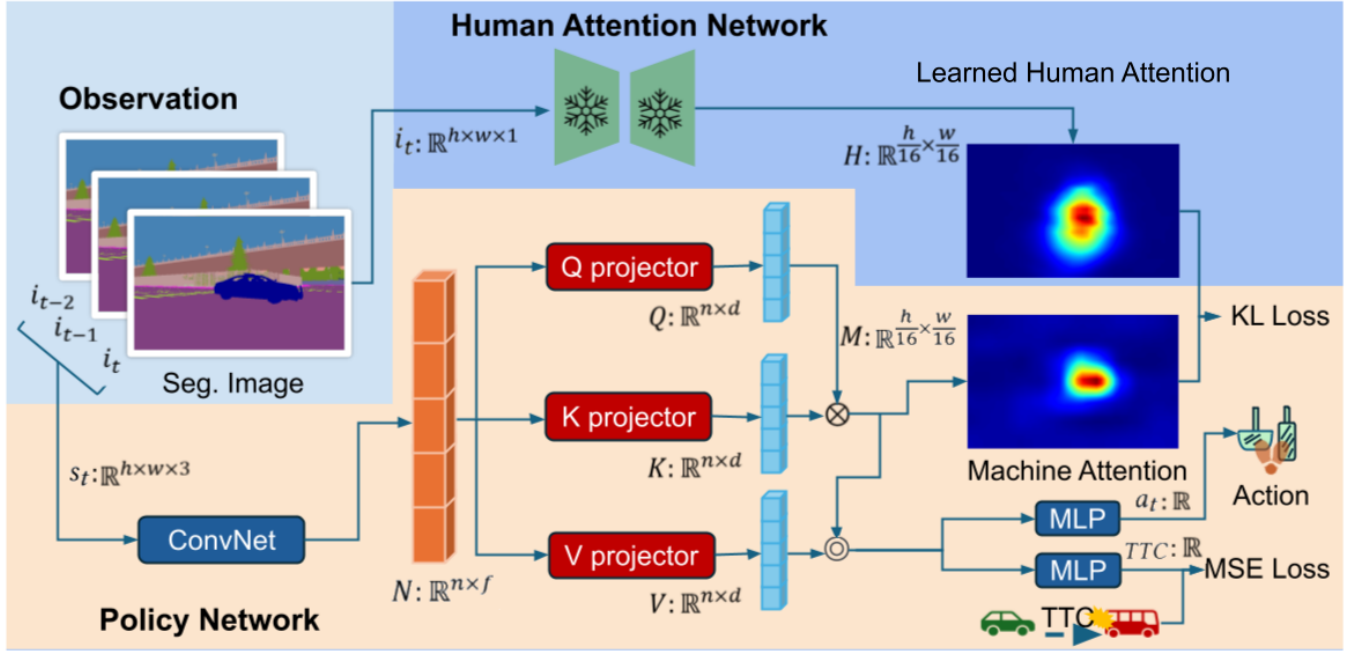


Figure 6: Structure of AEGIS. **Human Attention Network:** This pre-trained network predicts human attention from a segmentation image. **Policy Network:** This network determines the vehicle’s policy from a sequence of three segmentation images, starting with a CNN to extract features. These features are then flattened and processed through a self-attention layer, producing machine attention. This machine attention regulates RL training using the KL divergence loss relative to human attention. The policy network includes two MLP prediction heads: one for estimating the throttle and brake strength and another for predicting TTC, which aids in training regularization through the MSE loss. \otimes represents dot product, and \odot represents scaled dot product (after normalization and Softmax).

5 RESULTS

5.1 Experimental Settings

In this section, we discuss the training and testing scene settings and their differences for all scenarios in *Scene Settings*, present the metrics for benchmarks and attention maps in *Evaluation Metrics*, and provide details of the baselines and the hyperparameters of AEGIS in *Baseline Methods*.

Scene settings. We use CARLA [20] to construct the learning and testing environment for AEGIS. CARLA is an open-source driving simulator that offers maps of 14 towns for urban driving simulation. In our setup, the car-following scenario uses Town 7 as the training scene and Town 4 as the testing scene (see Fig. 5). The left-turn scenario uses Town 1 as the training scene and Town 5 as the testing scene. Although we use semantic segmentation masks as inputs, significant differences still exist between different towns. For example, bridges and cross junctions in Town 5 represent new categories that are not present in the training scenes from Town 1. Additionally, variations in traffic conditions and road structure, such as narrower or broader roads, also influence performance and contribute to the differences between towns. For four diverse occlusion scenarios, we follow the public Drive in Occlusion Sim (DOS) benchmark proposed in ReasonNet [66]. We follow the training/evaluation setting in ReasonNet and use 5 cases for training and other 20 cases for evaluation for each scenario. By testing in

the unseen scenes, we can evaluate the models’ generalizability in new environments. We train five models per method with different random seeds and report the average performance for a fair comparison. In the **free action setting**, each method can execute its own actions, which allows us to verify their performance. However, this setting leads to different observations from the environment within the same town. To compare visualization results across methods fairly, we also employ a **fixed action setting**, where the throttle value remains at 0.6 during inference.

Evaluation metrics. We evaluate the performance of the methods using the following metrics: a) Success rate, defined as the percentage of trials the agent reaches its destination without collision; b) Survival distance, which measures the distance traveled without any incidents; c) TTC as derived from Eq. 5; and d) Reward, calculated based on Eq. 1. We evaluate the similarity between machine attention and human attention through the following distribution-based metrics: a) Pearson’s Correlation Coefficient (CC); b) KL divergence; c) Similarity (SIM) [9, 74]; and d) location-based metric Normalized Scanpath Similarity (NSS) [55]. Our analysis goes beyond these metrics (see Sec. 5.3).

Baseline methods. *Benchmark for car-following and left-turn scenarios:* In these two scenarios, we adopt the TD3 algorithm [25] as the RL method for both AEGIS and all baseline methods. Specifically, we introduce a baseline referred to as Vanilla, which uses the same policy network architecture shown in Fig. 6 but excludes the

\mathcal{L}_{kl} component. In addition to Vanilla, we include behavior cloning (BC) [30, 57], a widely used approach in imitation learning (IL), as another baseline method. For the BC baseline, we construct a model to mimic human policy using the same network structure as the policy network in AEGIS. The model is trained with MSE loss on collected human demonstration data and then fine-tuned using RL. Unlike Vanilla, which starts training from scratch without prior knowledge, the BC baseline begins with a pretrained human policy and further refines it through RL. This approach allows the BC model to incorporate human guidance directly in the action space. By comparing AEGIS with both Vanilla and BC, we aim to emphasize the interpretability of human attention guidance.

As for training details of RL, the hyperparameters of TD3 include a replay buffer capacity of 38,400 and a minibatch size of 16, ensuring efficient data utilization during training. Learning rates for the actor and critic networks are 5×10^{-4} and 2×10^{-4} , respectively, with a learning rate decay of 0.995 per episode to stabilize adjustments over time. Exploration rates decrease from 0.5 to 0.05 to balance exploration with exploitation, and the discount factor γ is 0.95. The hyperparameter settings for the loss in Eq. 6 are $\alpha = 0.05$ and $\beta = 0.1$. α is a regularization term that allows the RL to maximize the reward while minimizing the KL divergence between human and machine attention. The hyperparameter settings for the reward in Eq. 1 are $\omega = -1$ and $\delta = 1$. All baseline methods use segmentation masks as input.

Benchmark for occlusion scenarios: In these four scenarios [66], we adopt the PPO algorithm [62] as the RL method for both AEGIS and all RL-based baseline methods to demonstrate the compatibility of the proposed framework with different RL methods. We benchmark AEGIS with different attention guidance, including Class Activation Maps [97] (CAM-guided) and random masks (Random-guided) by replacing human attention. By comparing AEGIS with Random-guided, we aim to demonstrate the advantages of focusing on task-relevant objects. Similarly, comparing AEGIS with CAM-guided allows us to assess the benefits of incorporating human guidance. We also compare the result with ReasonNet [66], a recent imitation learning method with cameras and LIDAR sensor fusion for a more comprehensive benchmark. We choose ReasonNet since it **ranked 1st** in the CARLA Leaderboard 1.0.

For the training details of RL, the learning rate is 3×10^{-4} . The value function’s importance is weighted by a coefficient of 0.5, and a clipping range of 0.2 helps maintain policy stability. The hyperparameter settings for the loss in Eq. 6 are $\alpha = 0.05$ and $\beta = 0$ for occlusion scenarios. The hyperparameter settings for the reward in Eq. 1 are $\omega = -0.2$ and $\delta = 0$. Other hyperparameters of PPO are implemented using default settings in stable baselines 3 [58]. All the baseline methods use segmentation masks as input, except for ReasonNet, which is a closed-source method that uses additional input such as LIDAR.

5.2 Benchmarks

In this section, we show the faster convergence speed and training performance of AEGIS in the *Evaluation of Training Phase*, and analyze the better testing performance, generalization ability, decision-making and attention mechanisms of AEGIS in the *Evaluation of Testing Phase*.

Evaluation of the training phase. We first investigate whether learned human attention can improve RL learning by comparing it against Vanilla and BC in our car-following and left-turn scenarios. The average reward and survival distance serve as metrics to demonstrate performance across episodes. As depicted in Fig. 7, AEGIS outperforms Vanilla and BC by achieving the highest average reward at the end of training for both scenarios. Notably, AEGIS reaches the highest reward achieved by Vanilla in fewer episodes, 270% faster for the car-following scenario and 150% faster for the left-turn scenario. Note that all methods require similar GPU training time, as the inference time for the pretrained human attention network is merely 0.005 seconds. Additionally, AEGIS shows lower variance than Vanilla, indicating more robust performance. While BC exhibits a faster convergence speed initially, it does not perform well in the testing scene and thus suffers from severe overfitting issues (see Tab. 4 and Tab. 5). The results of survival distance further verify the statement.

Evaluation of testing phase. We demonstrate that AEGIS outperforms Vanilla and BC in performance with the free action setting on an unseen scene. Tab. 2 shows that AEGIS consistently outperforms the other baselines in two scenarios, achieving an average success rate of 62% in the car-following scenario and 65% in the left-turn scenario. Moreover, AEGIS achieves the highest survival distance and TTC, showing its superior ability to maintain a safe driving distance and avoid collisions compared to other methods. Notably, the success rates of Vanilla and BC decrease significantly by 17% and 23% for the car-following scenario and 38% and 50% for the left-turn scenario in the new scene, highlighting their poor generalization capabilities, while AEGIS has a minor decrease of 2% and 6% for car-following and left-turn scenarios respectively (see Tab. 4 and Tab. 5). Although segmentation masks are used as input to mitigate the domain gap, BC and Vanilla still suffer from significant overfitting, whereas AEGIS does not, highlighting the importance of using learned human attention guidance to identify critical objects.

Additionally, we visualize the attention learned by AEGIS, Vanilla, and BC within the car-following and left-turn scenarios, as shown in Fig. 8a and Fig. 9a, respectively. In the car-following scenario, AEGIS focuses on the most critical object, the lead vehicles, at all four time steps. This aligns with learned human attention (refer to the bottom row of Fig. 8a), indicating that the AEGIS is effectively guided by learned human attention through our framework. Remarkably, AEGIS is able to prioritize the important instances (see t2 in Fig. 8a) when the vehicles are far apart and shifts its focus to the surrounding vehicles as they get closer (see t3 in Fig. 8a). In contrast, the machine attention of Vanilla appears more random and is often focused on the ground at all time steps. Although the machine attention of BC focuses on the lead vehicle, it is more scattered compared to AEGIS, with unnecessary attention to the background, which is less relevant to the task (see Sec. 5.3).

We further analyze the actions of AEGIS alongside baseline methods at four time steps as presented in Fig. 8a and Fig. 8b. In the car-following scenario, the lead vehicle is closer to the ego vehicle around t1, moves away around t2, starts braking around t3, and is hit around t4. AEGIS performs well by choosing to initially brake around t1, throttle around t2, brake timely around t3, and ultimately performing a full brake around t4. Unlike AEGIS, Vanilla

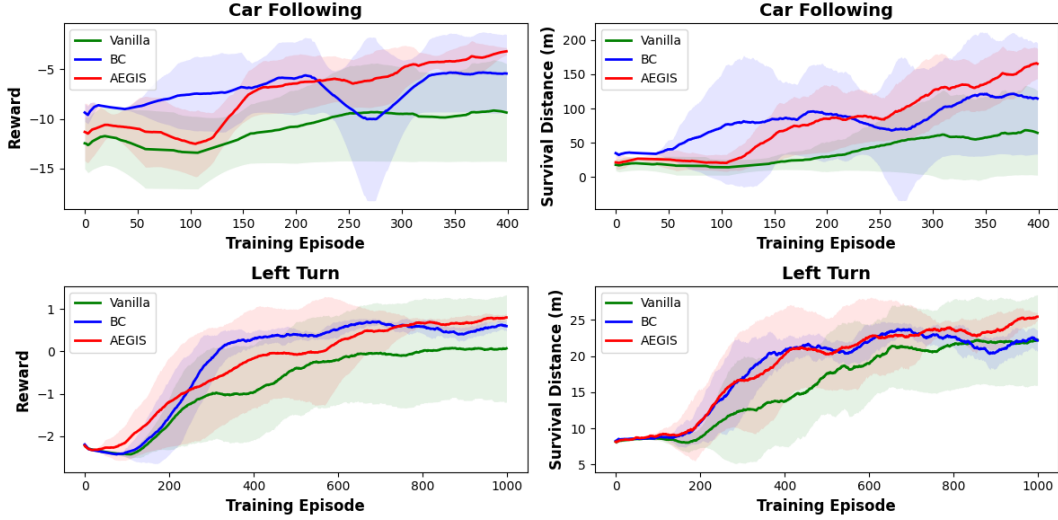


Figure 7: The training curves of AEGIS, Vanilla, and BC evaluated with reward and survival distances. With the help of learned human attention, AEGIS achieves the highest performance of Vanilla with fewer episodes.

Table 2: The evaluation results of car following / left turn in the unseen town with the free action setting. AEGIS outperforms Vanilla and BC.

Model	Success rate \uparrow	Survival distance \uparrow	TTC \uparrow
AEGIS (Ours)	$0.62 \pm 0.48 / 0.65 \pm 0.27$	$134 \pm 72 / 18 \pm 11$	$2.4 \pm 0.2 / 2.0 \pm 0.1$
BC	$0.46 \pm 0.40 / 0.23 \pm 0.22$	$97 \pm 76 / 13 \pm 5$	$2.3 \pm 0.1 / 1.9 \pm 0.1$
Vanilla	$0.18 \pm 0.36 / 0.33 \pm 0.36$	$35 \pm 41 / 15 \pm 6$	$2.1 \pm 0.17 / 1.8 \pm 0.1$

tends to maintain the throttle status most of the time, and BC tends to keep the highest throttle and act quickly. In the left-turn scenario, a similar pattern occurs, whereas a difference is that Vanilla acts somewhat akin to AEGIS. However, compared with AEGIS, Vanilla RL tends to increase throttle around t_2 , resulting in less robust and inconsistent decision-making.

For benchmark purposes, we further test AEGIS in four diverse occlusion scenarios following [66]. AEGIS has the best overall performance in the DOS benchmark. Remarkably, AEGIS is competitive with ReasonNet [66] in all four occlusion scenarios while maintaining explainability (see Tab. 3). Although the success rate of AEGIS is slightly higher than that of CAM-guided in Tab. 3, the

Table 3: The success rate of AEGIS and baselines over four occlusion scenarios in the unseen town with a free action setting.

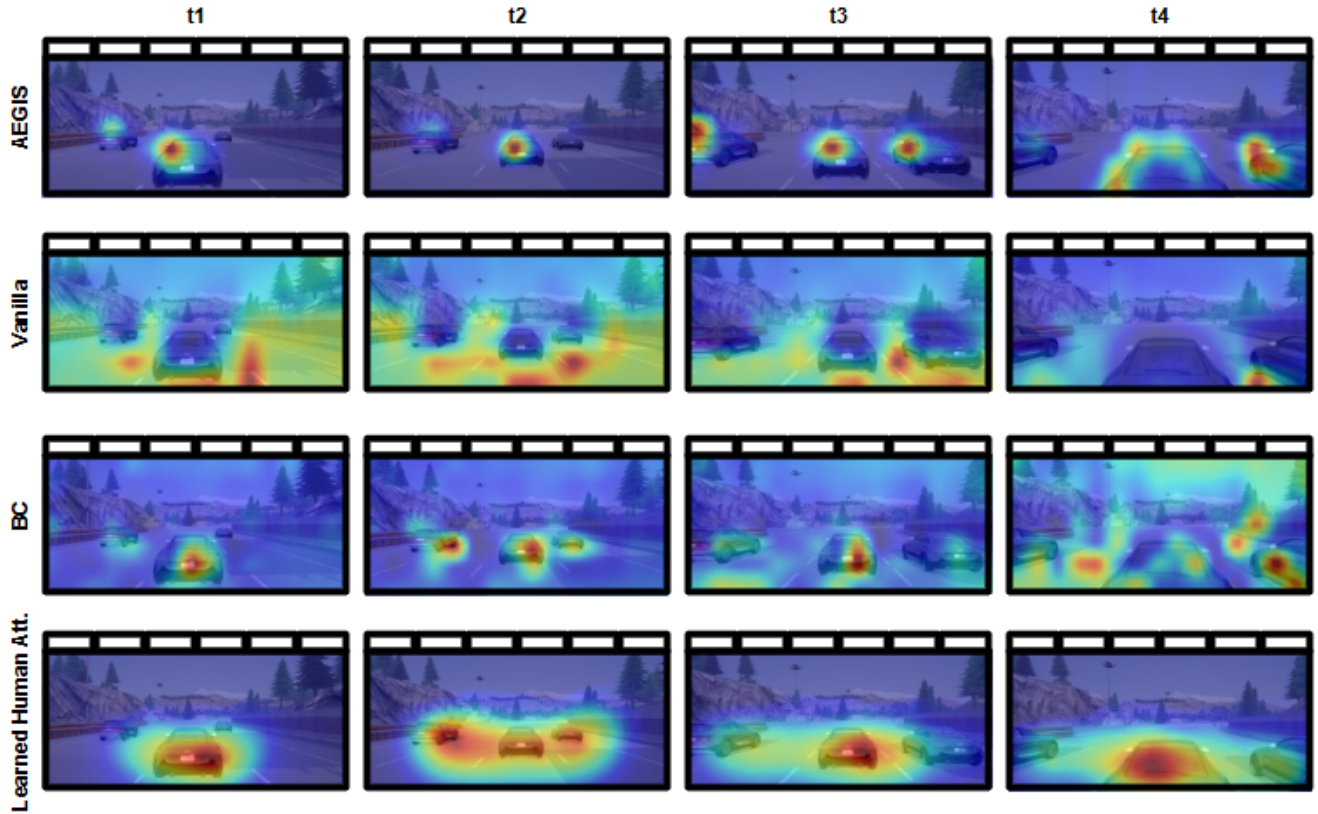
Model	DOS1	DOS2	DOS3	DOS4
AEGIS (Ours)	0.66 ± 0.13	0.72 ± 0.14	0.84 ± 0.07	0.79 ± 0.11
CAM-guided	0.63 ± 0.04	0.63 ± 0.15	0.83 ± 0.16	0.77 ± 0.10
Random-guided	0.23 ± 0.18	0.43 ± 0.40	0.72 ± 0.20	0.65 ± 0.27
Vanilla	0.21 ± 0.05	0.59 ± 0.26	0.75 ± 0.06	0.70 ± 0.06
ReasonNet [66]	0.63 ± 0.04	0.73 ± 0.03	0.80 ± 0.04	0.70 ± 0.06

attention is more reasonable and explainable in Fig. 10. We also observe that integrating CAM guidance into the framework, even using the basic version [97], approximately doubles the training time. The visualization results of machine attention in Fig. 10 show that AEGIS effectively identifies pedestrians in scenarios where they suddenly cross the road, allowing the ego vehicle to respond appropriately.

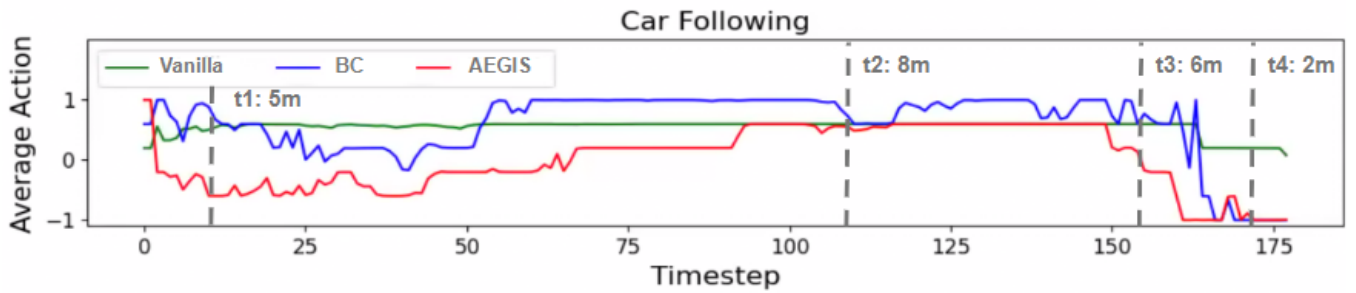
5.3 Human-like attention and its benefits

In this section, we show that the attention of AEGIS is closer to learned human attention in *Similarity between Machine and Learned Human Attention*, focusing more frequently on critical categories (e.g., vehicles or pedestrians) than humans due to the regularization design in *Ratio of Focus Categories*, and exhibiting less scatter in *Spatial Entropy of Machine Attention*, regardless of the focus categories. In addition, we demonstrate that our human attention guidance design is reasonable in *Correlation Between Human Attention and Rewards*, along with quantitative results showcasing improved interpretability in *Interpretability*.

Similarity between Machine and Learned Human Attention. The machine attention of AEGIS closely aligns with human attention, especially in two *distribution-based* metrics: CC and KL (see Tab. 6 and Tab. 7). Although the SIM and NSS of AEGIS and BC are similar in the left-turn scenario, the CC and KL of AEGIS in the



(a) The learned human attention and the attention of AEGIS, Vanilla, BC. AEGIS focuses on cars and prioritizes instances (t_2).

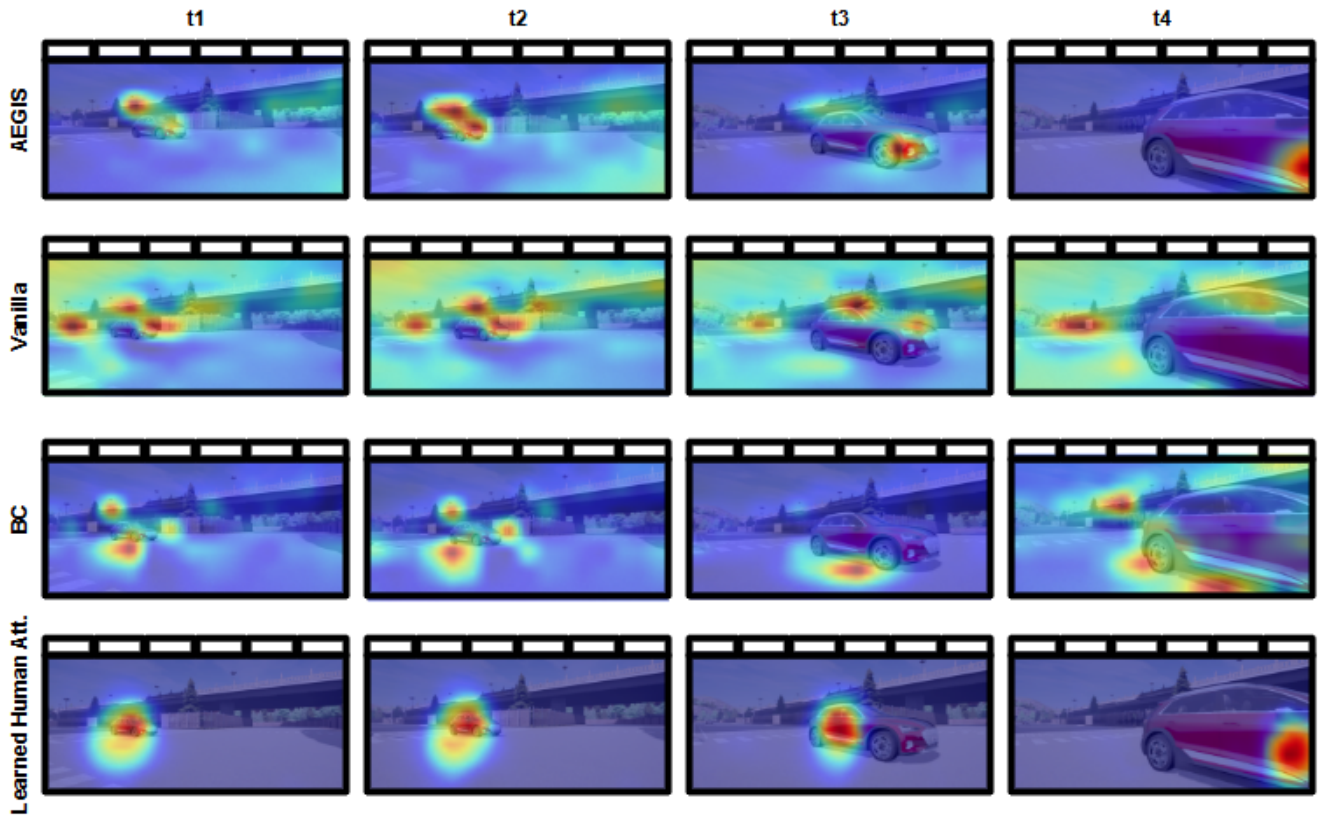


(b) The action of the methods across the time step. The dotted lines show the four time steps of the top figure with the distance to the lead vehicle.

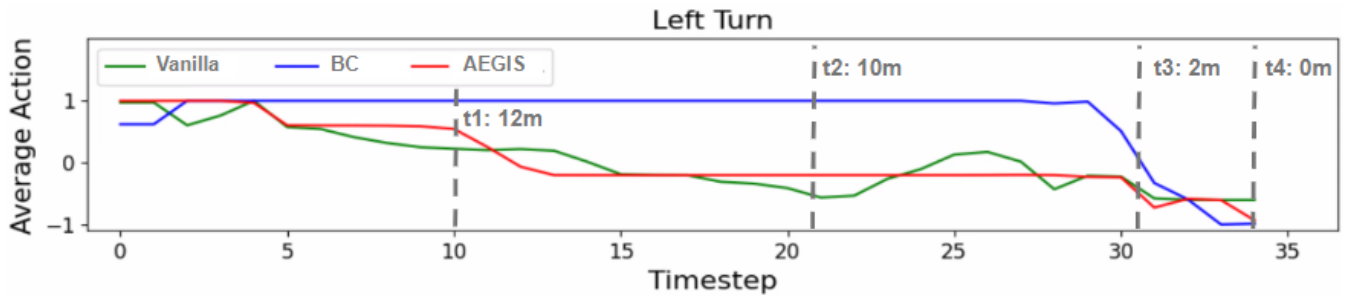
Figure 8: The visualization results of attention of the car-following scenario with the fixed action setting, and the average action across time steps for each method.

Table 4: Performance gap between training to evaluation environment in the car-following scenario. AEGIS drops less than Vanilla and BC.

Model	Training success rate \uparrow	Testing success rate \uparrow	Drop \downarrow
AEGIS (Ours)	0.64 ± 0.33	0.62 ± 0.48	0.02
BC	0.69 ± 0.36	0.46 ± 0.40	0.23
Vanilla	0.35 ± 0.43	0.18 ± 0.36	0.17



(a) The learned human attention and the attention of AEGIS, Vanilla, BC. AEGIS can focus on the critical object.



(b) The action of the methods across the time step. The dotted lines show the four time steps of the top figure with the distance to the closest vehicle.

Figure 9: The visualization results of attention of the left-turn scenario with the fixed action setting, and the average action across time steps for each method.

Table 5: Performance gap between training to evaluation environment in the left-turn scenario. AEGIS drops less than Vanilla and BC.

Model	Training success rate \uparrow	Testing success rate \uparrow	Drop \downarrow
AEGIS (Ours)	0.71 ± 0.21	0.65 ± 0.27	0.06
BC	0.73 ± 0.16	0.23 ± 0.22	0.50
Vanilla	0.71 ± 0.21	0.33 ± 0.36	0.38

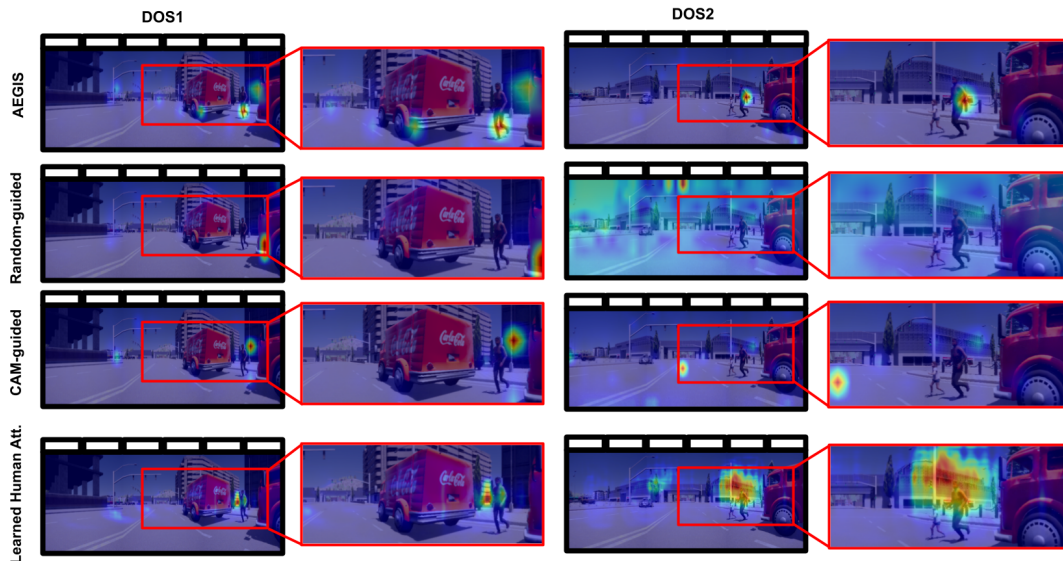


Figure 10: Visualization results of the occlusion scenes with pedestrians (DOS1 and DOS2) with the fixed-action setting. AEGIS successfully identifies the pedestrian near the red vehicle, while other baselines fail to recognize the pedestrian.

Table 6: The similarity between human and machine attention in the car-following scenario in an unseen town with a fixed action setting. The attention of AEGIS is more similar to human attention.

Model	CC \uparrow	KL \downarrow	SIM \uparrow	NSS \uparrow
AEGIS (Ours)	0.43 ± 0.03	2.15 ± 0.06	0.62 ± 0.07	0.13 ± 0.55
BC	0.32 ± 0.06	2.34 ± 0.12	0.59 ± 0.04	-0.23 ± 1.00
Vanilla	0.24 ± 0.10	2.47 ± 0.20	0.52 ± 0.05	-1.17 ± 0.89

Table 7: The similarity between human and machine attention in left-turn scenario in an unseen town with a fixed action setting. The attention of AEGIS is more similar to human attention.

Model	CC \uparrow	KL \downarrow	SIM \uparrow	NSS \uparrow
AEGIS (Ours)	0.25 ± 0.08	4.26 ± 0.29	0.34 ± 0.05	0.20 ± 0.18
BC	0.21 ± 0.01	4.42 ± 0.02	0.34 ± 0.01	0.20 ± 0.36
Vanilla	0.18 ± 0.07	4.50 ± 0.14	0.29 ± 0.07	-0.38 ± 0.97

same scenario demonstrate AEGIS is closer to learned human attention. Overall, AEGIS demonstrates greater similarity to learned human attention when considering both scenarios together. Moreover, in the DOS scenario, AEGIS demonstrates greater similarity to learned human attention compared with CAM-guided attention (see Tab. 8). The visualization results in Fig. 8a, Fig. 9a and Fig. 10 further verify this statement.

Ratio of Focus Categories. We further analyze the focus categories of machine attention from the methods and compare them

Table 8: KL divergence between machine attention and learned human attention for CAM-guided and AEGIS models across occlusion scenarios (DOS1-DOS4).

Model	DOS1	DOS2	DOS3	DOS4
AEGIS (Ours)	1.86	1.13	3.24	3.67
CAM-guided	2.11	1.46	3.74	4.92

with human attention (see Fig. 11). To precisely quantify the focus categories, we first filter out the less relevant regions. This process converts machine attention maps to binary masks using a 0.1 threshold and removes regions with attention levels lower than the threshold. Subsequently, we intersect these masks with semantic segmentation data to calculate category-specific ratios. Finally, we average these ratios across all the images and models to identify the primary focus areas. We find that AEGIS can focus on the most critical objects, the cars, more often than other baseline methods can, with a ratio of 33.5 % in the left-turn scenario and 24.6% in the car-following scenario. Although the focus ratio of the road is similar between BC and learned human attention in the left-turn scenario, BC allocates only 14.6% of its attention to the most critical objects, the cars, whereas learned human attention allocates 22.0%. This confirms our design of human attention guidance, which serves as a regularization term and allows the machine to refine the attention on its own further instead of replicating learned human attention identically. For overall similarity between machine attention and learned human attention, please refer to *Similarity between Machine and Learned Human Attention*.

In the DOS3 and DOS4 scenarios (see Fig. 12), AEGIS allocates the highest percentage of attention to vehicles compared with other

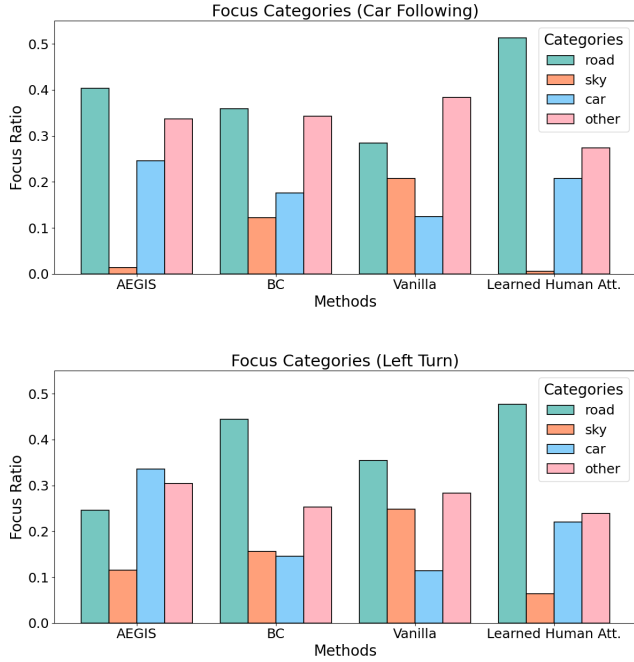


Figure 11: Ratio of focus categories from different methods and learned human attention in the car-following and left-turn scenarios. AEGIS can concentrate on the most crucial object, the car.

baseline methods, potentially reflecting its prioritization of task-critical elements. In the DOS1 and DOS2 scenarios, which emphasize pedestrian avoidance, AEGIS demonstrates the highest focus on pedestrians.

Table 9: Spatial entropy of the attention maps from car-following and left-turn scenario. The spatial entropy indicates the degree of sparsity in attention maps, regardless of whether the focus is on the critical object or not. The spatial entropy of AEGIS is smaller than Vanilla and BC, indicating more concentrated attention.

Model	Left Turn	Car Following
AEGIS (Ours)	0.896	0.869
BC	0.945	0.928
Vanilla	0.965	0.954
Learned human attention	0.872	0.923

Spatial Entropy of Machine Attention. We analyze the spatial entropy of the machine attention to quantify the overall uncertainty of the focus of attention, following the methodology in [68]. We partition the image to a 4×4 grid and calculate the spatial entropy [69] via Shannon’s entropy equation [64]. Notably, spatial entropy illustrates the scattered degree of attention, independent of whether

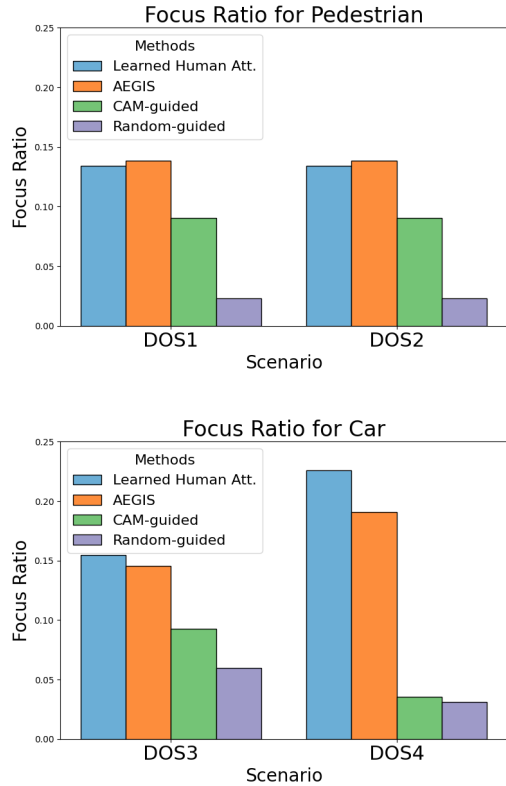


Figure 12: Ratio of focus information for pedestrian-avoidance scenarios (DOS1 and DOS2) and car-avoidance scenarios (DOS3 and DOS4). AEGIS demonstrates a higher focus on dynamic critical objects such as pedestrians in DOS1 and DOS2 and cars in DOS3 and DOS4.

it is directed at critical objects. While BC’s spatial entropy is closer to human levels for car following scenario of Tab. 9, it may not align with human object focus. For instance, in Fig. 11, BC focuses less on the car and more on the sky than learned human attention and AEGIS in the car-following scenario. We prefer smaller spatial entropy than the human attention network, as it indicates that RL agents not only learn human attention patterns but also refine learned human attention by filtering out less relevant objects. This can be observed in Fig. 8a, where AEGIS in t2 focuses solely on the critical car, while BC in t4 is overly scattered.

AEGIS has lower spatial entropy than BC and Vanilla in both car-following and left-turn scenarios (see Tab. 9). This illustrates that the machine attention of AEGIS is more concentrated and less random across spatial locations. For overall similarity between machine attention and learned human attention, please refer to *Similarity between Machine and Learned Human Attention*.

Correlation between Human Attention and Rewards. Like [29], we further investigate the relationship between the rewards and the similarity with human attention. To this end, a linear regression analysis is conducted to examine the correlation between RL rewards and Pearson’s correlation coefficient (CC) between human

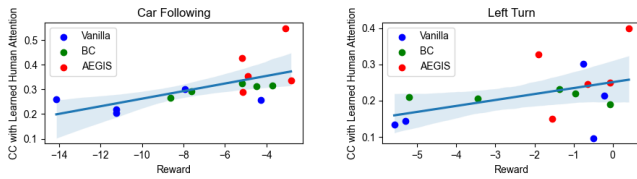


Figure 13: Linear regression of the Pearson's CC between human and machine attention regarding reward. The X-axis represents the reward, and the Y-axis represents the CC (similarity) between machine and learned human attention. The CC is positively correlated with the reward in two scenarios.

and machine attention (see Fig. 13). The points representing CC and rewards in the linear regression are obtained from the five models trained with different random seeds for each method, as reported in Tab. 6 and Tab. 7. The X-axis represents the reward, and the Y-axis shows the CC (similarity) between machine and learned human attention. Notably, AEGIS has the highest average CC (see red dots in Fig. 13, Tab. 6 and Tab. 7). In both scenarios, rewards (performance) are positively correlated with CC between machine and learned human attention. In the left-turn scenario, the linear regression model yields an R-squared value of 0.377 ($p = 0.0148$). These results indicate that the positive correlation between CC and RL rewards is statistically significant at conventional significance levels. For the car-following scenario, while a positive correlation is observed with an R-squared value of 0.128, the associated p value of 0.19 indicates that this correlation does not achieve statistical significance.

Interpretability. In this work, interpretability is defined as the degree to which a model's decision-making process can be understood and explained by humans [3]. Naturally, if a model's machine attention map is closer to human attention, the model becomes more interpretable and human-understandable. As shown in Tab. 6 and Tab. 7, AEGIS achieves closer alignment with human attention. We also conducted a survey with 80 human participants, asking three questions related to the interpretability and safety of the model. The participants were compensated at a rate higher than the country's law. The survey involved showing videos of AEGIS, Vanilla, and BC in the video figure.

- **QA:** Rank the three videos from easiest to most difficult to understand, based on the relationship between the visualization and the action.
- **QB:** Rank the three videos based on how well their visualizations and actions meet your expectations.
- **QC:** Rank the three videos based on how confident you feel about their safety and reliability in performing autonomous driving tasks.

QA and QB focus on the interpretability of the models, aiming to identify which model is more human-understandable, while QC evaluates the safety and reliability of the models. Fig. 14 demonstrates that AEGIS achieves better average rankings, indicating that human attention guidance can enhance both interpretability and safety.

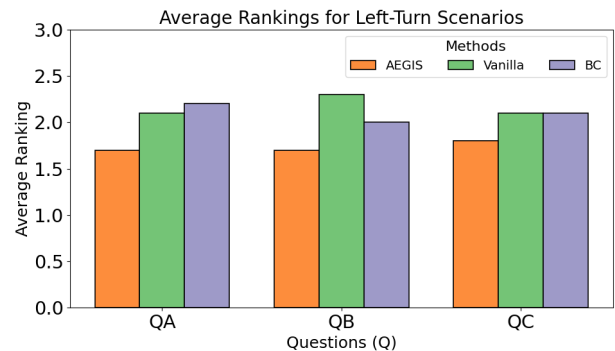
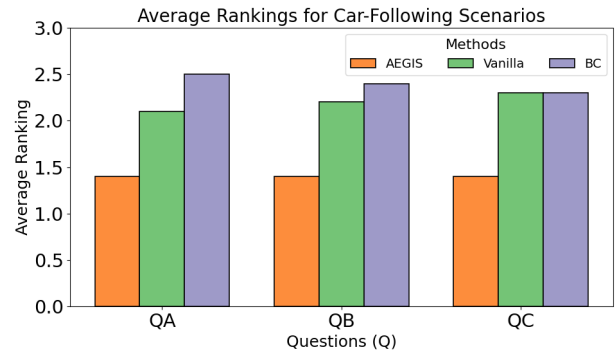


Figure 14: The average rankings for interpretability (QA and QB) and safety (QC), derived from a survey conducted with 80 participants, indicate that AEGIS improves both interpretability and safety.

6 DISCUSSION

In this work, we present AEGIS, a framework to increase the interpretability and performance of intelligent vehicles via human attention guidance. By aligning machine attention with learned human attention closely in the early training phase of RL, AEGIS can learn to focus on task-relevant objects. This is essential for preventing overfitting in RL training, as demonstrated in Fig. 15. Both the Vanilla RL and AEGIS can successfully execute a left turn. However, AEGIS performs the maneuver based on the correct cue, the oncoming vehicle, whereas the Vanilla RL incorrectly bases its action on the sidewalk. This is supported by the dramatic drop of 38% in the performance of Vanilla from training to testing (see Tab. 4 and Tab. 5). Our framework can improve the interpretability of current RL, evidenced by higher similarity with learned human attention and higher ranking in the interpretability survey. In addition, we propose the largest human eye-tracking dataset that is designed for the task. Compared with existing in-car datasets, Our dataset faces fewer ethical concerns, has lower costs, and can provide repeatable scenarios for different drivers. Compared with other in-lab datasets, our dataset offers a more immersive experience using a VR headset.

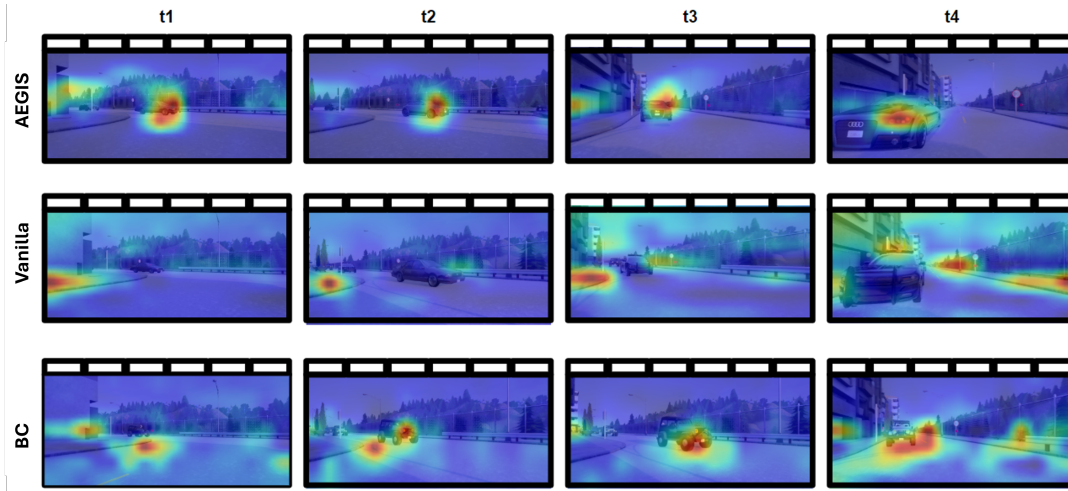


Figure 15: Visualization results of the left-turn scene with the free-action setting during training. Vanilla and BC overfit on the non-vehicle objects, leading to the performance drop during testing.

6.1 Human attention network

The pre-trained human attention network accurately predicts where most people are likely to focus their attention in a scene. For example, the human attention network can identify the lead vehicle in a car-following scenario (see Fig. 16). This network ensures consistency in predicting learned human attention across participants, who may exhibit diverse eye-tracking patterns. For example, in Fig. 16, S07’s attention is primarily directed toward the road fence, whereas S09 focuses more on the lead vehicle, however, the learned human attention can predict where most people look. Additionally, the human attention network enables the prediction of human attention across different observations, allowing agents to take their actions. The use of a pre-trained network also reduces the burden on human operators in practical applications, as the pre-trained human attention network eliminates the need for human involvement during RL training unlike human-in-the-loop RL in which humans need to stand by [84–86], and it allows human-free inference which can have broader application when human gaze data are unavailable.

To evaluate the central bias of the human attention model, we adopt the approach from [18, 54] and employ information gain (IG) [37] as a metric. IG measures the quality of the learned human attention model by comparing learned human attention to ground truth human attention and a baseline map. For this analysis, we use a centered Gaussian baseline [18] as the baseline map. An IG score greater than zero indicates that the learned human attention surpasses the centered Gaussian baseline in predicting human attention, thereby demonstrating less central bias. The learned human attention achieves a better KL, NSS, SIM, CC than the centered Gaussian baseline. An IG score much greater than zero not only demonstrates reduced central bias but also suggests the network’s capability to predict task-driven changes in gaze direction.

Table 10: KL, NSS, SIM, CC, and Information Gain (IG) are used to evaluate the learned human attention compared to a centered Gaussian. Lower KL value and higher NSS, SIM, CC, and IG indicate that the learned human attention is not simply predicting attention at the center of the image, demonstrating its effectiveness beyond a mean predictor.

Attention	KL↓	NSS↑	SIM↑	CC↑	IG↑
learned human attention	2.11	0.54	0.37	0.46	4.95
centered Gaussian	6.27	0.52	0.13	0.16	0.00

6.2 Analysis of tolerance to noisy input

In this study, the RL model is fed with perfect segmentation images from the simulator. To assess the model’s robustness and reliability, we investigate the resilience of our model to noise by using RGB images as the input and employing a pre-trained segmentation model to obtain segmentation images. We employ Segformer [88] pre-trained on the Cityscapes [14] dataset, resulting in noisier segmentation images than the segmentation directly from the simulator (see Fig. 17). Note that the segmentation images from the CARLA simulator contain customized classes like bridges and road lines that are not presented in Cityscapes. We then evaluate the trained RL agents in the noisier segmentation images. AEGIS outperforms BC and Vanilla with a success rate of 55% and 57% in the left-turn and car-following scenarios, respectively (see Tab. 11). This analysis reveals the potential of our model to maintain high performance even in less-than-ideal conditions.

6.3 Impact of input frame number

For all the experiments, we use three frames of images as the input of RL. Tab. 12, we investigate the impact of temporal information by varying the number of input frames. Our finding suggests that using 3 frames can achieve the best results in both scenarios. The 1-frame setting does not include sufficient temporal information

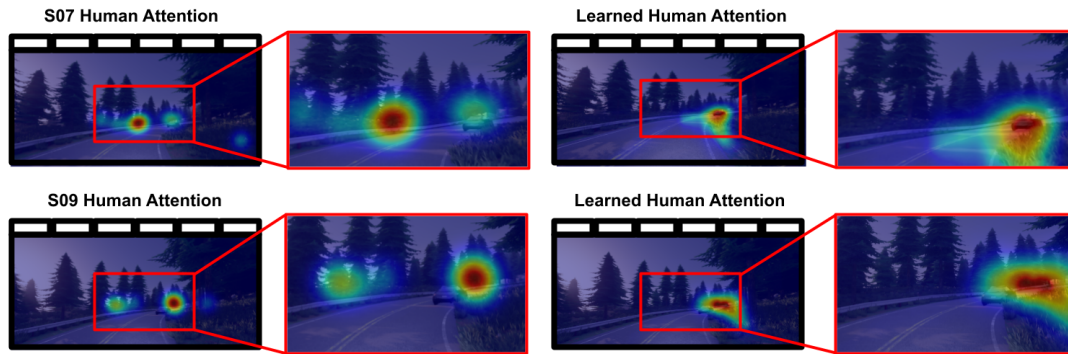


Figure 16: Human attention from subjects S07 and S09 reveals different focus patterns in similar scenes. S07’s attention is primarily directed toward the road fence, while S09 focuses more on the lead vehicle. Incorporating a human attention model can help mitigate distraction issues.

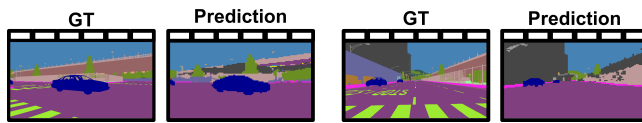


Figure 17: Comparison between segmentation images directly from the simulator (GT) and segmentation images from pre-trained Segformer (Prediction).

Table 11: Evaluation results of the success rate using segmentation images from Segformer. AEGIS outperforms BC and Vanilla.

Model	Left turn	Car following
AEGIS (Ours)	0.55 ± 0.14	0.57 ± 0.31
BC	0.26 ± 0.14	0.17 ± 0.20
Vanilla	0.14 ± 0.26	0.17 ± 0.34

and, thereby, cannot accurately predict the vehicle dynamics. On the other hand, the 5-frame setting also degrades performance, likely due to the model’s limited capacity to process the increased information effectively. Therefore, we adopt the 3-frame setting for all the analyses to achieve a balance between incorporating sufficient temporal information and maintaining manageable model capacity.

Table 12: Performance of AEGIS when training with 1, 3, 5 frames. 3-frames has the best performance due to a balance of model capacity and temporal information.

Scenario	1 frame	3 frame	5 frame
left turn	0.32 ± 0.10	0.65 ± 0.27	0.27 ± 0.12
car following	0.14 ± 0.28	0.62 ± 0.48	0.43 ± 0.47

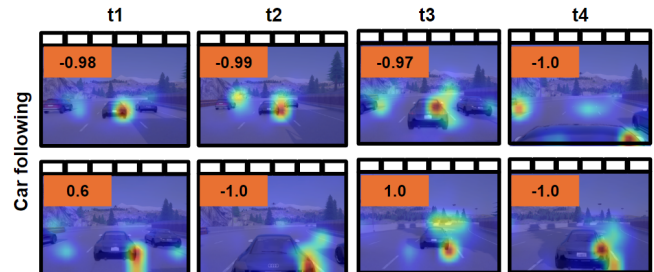


Figure 18: Failure cases of AEGIS with action value at the top-left side in the car-following scene. In the top row, the vehicle collides with the lead vehicle due to failure to maintain a safe following distance. In the bottom row, the vehicle collides with the front car in the road curve (t3-t4).

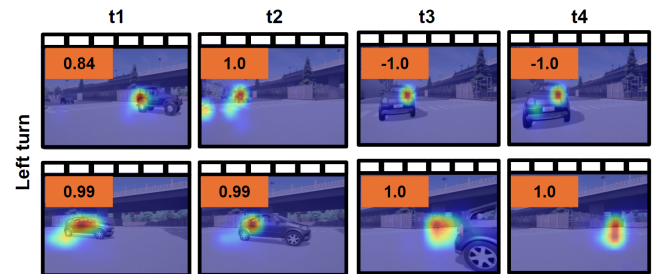


Figure 19: Failure cases of AEGIS with action value at the top-left side in the left-turn scene. In the top row, the vehicle chooses to stop as it cannot safely cross the junction. In the bottom row, the rear of the vehicle collides with another vehicle that has just passed through.

6.4 Analysis of failure cases

We collect a few interesting failure cases from AEGIS. In the top row of Fig. 18, the vehicle is unable to maintain a safe distance and collides with the car ahead despite executing an immediate braking decision. In the bottom row of Fig. 18, the vehicle successfully

avoids the collision on the straight road (t1 and t2), while hitting the lead vehicle on the curved road (t3 and t4). The visualization results suggest that the vehicle’s attention remains fixated on the curved corner, which may mean it is anticipating incoming traffic. However, this may hinder its ability to clearly see the lead vehicle and end up in a collision. In the top row of Fig. 19, the vehicle detects an approaching vehicle moving quickly towards it, leading it to choose to stop. This reaction closely mirrors human behavior, as humans tend to stop the vehicle in emergency situations. In the bottom row of Fig. 19, as the ego vehicle accelerates during a left turn, it collides with another vehicle almost completely past the intersection. This collision happens because the ego vehicle incorrectly estimates that there is sufficient space to complete its turn, failing to account for the other vehicle’s proximity and speed.

6.5 Limitation and future work

Although AEGIS shows improved success rates, quicker training speeds, and robust performance across different scenes, our current focus is limited to collision avoidance, with an emphasis on the vehicle’s speed. Furthermore, the RL agent in the current work still relies on hand-crafted reward functions, which requires hyperparameter tuning. Our goal is to develop a framework that leverages human attention to guide the RL agent and demonstrate that the RL agent can benefit from human attention rather than presenting a state-of-the-art method and claiming that attention is all you need. In future research, we aim to gather trajectory data from human drivers in more scenarios. However, collecting long-term and diverse driving data has been challenging due to motion sickness experienced by participants using VR headsets. Consequently, we intend to acquire more extensive and varied data from participants without motion sickness in future studies. Our works shed light on developing explainable guidance for RL autonomous driving tasks using human data. Although we do not claim that “attention is all you need” and that AEGIS can generalize well to any unseen scene, one interesting future step is to continue maintaining and augmenting the dataset with more scenarios. Another future step is to collect more human action data for steering control to create generalizable and scalable RL models.

Currently, this work is limited to simulations, as the simulator provides a cost-effective environment for closed-loop training compared to real-world data collection. To explore the potential real-world applicability of our model, we conducted a case study visualizing the learned machine attention on a single real-world video, as shown in the video figure. While the results demonstrate promise, further investigation and testing across a broader range of real-world scenarios are necessary to assess and improve the model’s generalizability in future work.

7 CONCLUSION

This paper presents a novel human attention-based explainable guidance for intelligent vehicle systems (AEGIS) framework as a solution to enhance the RL agent’s learning efficiency, generalization, and interpretability in complex driving scenarios. We collect eye-movement data from a VR simulator with 20 participants through our in-lab realistic data collection system and design a policy network with the self-attention layer guided by learned

human attention. With human attention guidance, AEGIS achieves the highest reward in a shorter timeframe in the car-following and left-turn scenarios. Moreover, AEGIS maintains the highest success rate in unseen maps of both scenarios, highlighting its robustness and generalization capability. We also conduct further analysis on the focus information of machine attention learned by AEGIS. It demonstrates that AEGIS can learn to prioritize task-relevant objects more effectively by aligning machine attention more closely with human attention. Moreover, a survey with 80 participants demonstrates the attention and action of AEGIS is more interpretable compared with methods without human attention guidance. The study implies the potential of incorporating human attention in the development of AIVs and emphasizes the benefits of integrating human cognitive processes with machine learning algorithms in autonomous driving.

8 ACKNOWLEDGMENTS

This work was partly supported by the Australian Research Council (ARC) under discovery grants DP210101093 and DP220100803 and the UTS Human-Centric AI Centre funding sponsored by GrapheneX (2023-2031). Research was partially sponsored by the Australia Defence Innovation Hub under Contract No. P18650825, Australian Cooperative Research Centres Projects (CRC-P) Round11 CRCPXI000007, USOfficeofNavalResearchGlobal under Cooperative Agreement Number ONRG- NICOP- N62909-19-12058, and AFOSR– DST Australian Autonomy Initiative agreement ID10134. We also thank the NSW Defence Innovation Network and the NSW State Government of Australia for financial support in part of this research through grant DINPP2019 S1-03/09 and PP2122.03.02. We also thank Yi-Shan Hung for contribution in the figures. Special thanks to Mrs Haiting Lan for proofreading.

REFERENCES

- [1] 2024. Artificial Intelligence (AI) Act: Council Gives Final Green Light to the First Worldwide Rules on AI. <https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/>. Accessed: 2024-11-27.
- [2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access* 6 (2018), 52138–52160.
- [3] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Diaz-Rodriguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information fusion* 99 (2023), 101805.
- [4] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. 2016. Dr (eye) ve: a dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 54–60.
- [5] Alejandro Barredo Arrieta, Natalia Diaz-Rodríguez, Javier Del Ser, Adrien Ben-netot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.
- [6] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. 2024. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *IEEE Access* (2024).
- [7] Sonia Bae, Erfan Pakdamanian, Inki Kim, Lu Feng, Vicente Ordóñez, and Laura Barnes. 2021. MEDIRL: Predicting the Visual Attention of Drivers via Maximum Entropy Deep Inverse Reinforcement Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13178–13188.
- [8] Ali Borji, Dicky N Sihite, and Laurent Itti. 2011. Computational modeling of top-down visual attention in interactive environments.. In *BMVC*, Vol. 85. 1–12.
- [9] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. 2018. What do different evaluation metrics tell us about saliency models? *IEEE*

- transactions on pattern analysis and machine intelligence* 41, 3 (2018), 740–757.
- [10] Jonathan Chang, Masatoshi Uehara, Dhruv Sreenivas, Rahul Kidambi, and Wen Sun. 2021. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems* 34 (2021), 965–979.
 - [11] Yuying Chen, Congcong Liu, Lei Tai, Ming Liu, and Bertram E Shi. 2019. Gaze training by modulated dropout improves imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7756–7761.
 - [12] Pranav Singh Chib and Pravendra Singh. 2023. Recent advancements in end-to-end autonomous driving using deep learning: A survey. *IEEE Transactions on Intelligent Vehicles* (2023).
 - [13] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. 2018. End-to-end driving via conditional imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 4693–4700.
 - [14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3213–3223.
 - [15] Resul Dagdanov, Halil Durmus, and Nazim Kemal Ure. 2023. Self-Improving Safety Performance of Reinforcement Learning Based Driving with Black-Box Verification Algorithms. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 5631–5637.
 - [16] Tao Deng, Lianfang Jiang, Yi Shi, Jiang Wu, Zhangbi Wu, Shun Yan, Xianshi Zhang, and Hongmei Yan. 2023. Driving Visual Saliency Prediction of Dynamic Night Scenes via a Spatio-Temporal Dual-Encoder Network. *IEEE Transactions on Intelligent Transportation Systems* (2023).
 - [17] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. 2019. How do drivers allocate their potential attention? Driving fixation prediction via convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems* 21, 5 (2019), 2146–2154.
 - [18] Tao Deng, Hongmei Yan, Long Qin, Thuyen Ngo, and BS Manjunath. 2020. How Do Drivers Allocate Their Potential Attention? Driving Fixation Prediction via Convolutional Neural Networks. *IEEE Transactions on Intelligent Transportation Systems* 21, 5 (May 2020), 2146–2154. <https://doi.org/10.1109/TITS.2019.2915540>
 - [19] Tao Deng, Kaifu Yang, Yongjie Li, and Hongmei Yan. 2016. Where does the driver look? Top-down-based saliency detection in a traffic driving environment. *IEEE Transactions on Intelligent Transportation Systems* 17, 7 (2016), 2051–2062.
 - [20] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. 2017. CARLA: An open urban driving simulator. In *Conference on robot learning*. PMLR, 1–16.
 - [21] Michel Failing and Jan Theeuwes. 2018. Selection history: How reward modulates selectivity of visual attention. *Psychonomic bulletin & review* 25, 2 (2018), 514–538.
 - [22] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, He Wang, and Sen Li. 2019. Dada-2000: Can driving accident be predicted by driver attention? analyzed by a benchmark. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 4303–4309.
 - [23] Jianwu Fang, Dingxin Yan, Jiahuan Qiao, Jianru Xue, and Hongkai Yu. 2021. DADA: Driver attention prediction in driving accident scenarios. *IEEE transactions on intelligent transportation systems* 23, 6 (2021), 4959–4971.
 - [24] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yariv Gal. 2020. Can autonomous vehicles identify, recover from, and adapt to distribution shifts?. In *International Conference on Machine Learning*. PMLR, 3145–3153.
 - [25] Scott Fujimoto, Herke Hoof, and David Meger. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*. PMLR, 1587–1596.
 - [26] Deepak Gopinath, Guy Rosman, Simon Stent, Katsuya Terahata, Luke Fletcher, Brenna Argall, and John Leonard. 2021. Maad: A model and dataset for “attended awareness” in driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3426–3436.
 - [27] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. 2018. Visualizing and understanding atari agents. In *International conference on machine learning*. PMLR, 1792–1801.
 - [28] Sorin Griborescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* 37, 3 (2020), 362–386. <https://doi.org/10.1002/rob.21918>
 - [29] Suna Sihang Guo, Ruohan Zhang, Bo Liu, Yifeng Zhu, Dana Ballard, Mary Hayhoe, and Peter Stone. 2021. Machine versus human attention in deep reinforcement learning tasks. *Advances in Neural Information Processing Systems* 34 (2021), 25370–25385.
 - [30] Shengran Hu and Jeff Clune. 2024. Thought cloning: Learning to think while acting by imitating human thinking. *Advances in Neural Information Processing Systems* 36 (2024).
 - [31] Hidenori Itaya, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyoshi, and Komei Sugiura. 2021. Visual explanation using attention mechanism in actor-critic-based deep reinforcement learning. In *2021 International Joint Conference On Neural Networks (IJCNN)*. IEEE, 1–10.
 - [32] Ho-Taek Joo and Kyung-Joong Kim. 2019. Visualization of deep reinforcement learning using grad-CAM: how AI plays atari games?. In *2019 IEEE Conference on Games (CoG)*. IEEE, 1–2.
 - [33] Isaac Kasahara, Simon Stent, and Hyun Soo Park. 2022. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*. Springer, 126–142.
 - [34] Iuliia Kotseruba and John K Tsotsos. 2021. Behavioral research and practical models of drivers’ attention. *arXiv preprint arXiv:2104.05677* (2021).
 - [35] Iuliia Kotseruba and John K Tsotsos. 2022. Attention for vision-based assistive and automated driving: A review of algorithms and datasets. *IEEE transactions on intelligent transportation systems* 23, 11 (2022), 19907–19928.
 - [36] Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. 2019. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in Neural Information Processing Systems* 32 (2019).
 - [37] Matthias Kümmeler, Thomas SA Wallis, and Matthias Bethge. 2015. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences* 112, 52 (2015), 16054–16059.
 - [38] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. 2020. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems* 22, 2 (2020), 712–733.
 - [39] Qixia Lai, Salman Khan, Yongwei Nie, Hanqiu Sun, Jianbing Shen, and Ling Shao. 2020. Understanding more about human and machine attention in deep neural networks. *IEEE Transactions on Multimedia* 23 (2020), 2086–2099.
 - [40] Timo Lajunen and Heikki Summala. 1995. Driving experience, personality, and skill and safety-motive dimensions in drivers’ self-assessments. *Personality and individual differences* 19, 3 (1995), 307–318.
 - [41] Olivier Le Meur and Thierry Baccino. 2013. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods* 45, 1 (2013), 251–266.
 - [42] Quanyi Li, Zhenghao Peng, and Bolei Zhou. 2022. Efficient learning of safe driving policy via human-ai copilot optimization. *arXiv preprint arXiv:2202.10341* (2022).
 - [43] Ji Hyoun Lim and Yili Liu. 2009. Modeling the influences of cyclic top-down and bottom-up processes for reinforcement learning in eye movements. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 39, 4 (2009), 706–714.
 - [44] Grace W Lindsay. 2020. Attention in psychology, neuroscience, and machine learning. *Frontiers in computational neuroscience* 14 (2020), 29.
 - [45] Congcong Liu, Yuying Chen, Lei Tai, Haoyang Ye, Ming Liu, and Bert Shi. 2019. A Gaze Model Improves Autonomous Driving. <https://doi.org/10.1145/3314111.3319846>
 - [46] Biao Luo, Zhengke Wu, Fei Zhou, and Bing-Chuan Wang. 2023. Human-in-the-loop reinforcement learning in continuous-action space. *IEEE Transactions on Neural Networks and Learning Systems* (2023).
 - [47] David Mascharka, Philip Tran, Ryan Soklaski, and Arjun Majumdar. 2018. Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4942–4950.
 - [48] Vishwali Mhasawade, Salman Rahman, Zoe Haskell-Craig, and Rumi Chunara. 2024. Understanding Disparities in Post Hoc Machine Learning Explanation. *arXiv preprint arXiv:2401.14539* (2024).
 - [49] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
 - [50] Alexander Mott, Daniel Zoran, Mike Chrzanowski, Daan Wierstra, and Danilo Jimenez Rezende. 2019. Towards interpretable reinforcement learning using attention augmented agents. *Advances in neural information processing systems* 32 (2019).
 - [51] Rakshit Naidu, Ankita Ghosh, Yash Maurya, Shamanth R. Nayak K, and Soumya Snigdha Kundu. 2020. IS-CAM: Integrated Score-CAM for axiomatic-based explanations. *CoRR abs/2010.03023* (2020). [arXiv:2010.03023](https://arxiv.org/abs/2010.03023) <https://arxiv.org/abs/2010.03023>
 - [52] Dmitry Nikulin, Anastasia Ianina, Vladimir Aliev, and Sergey Nikolenko. 2019. Free-lunch saliency via attention in atari agents. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 4240–4249.
 - [53] Liang Ou, Yu-Chen Chang, Yu-Kai Wang, and Chin-Teng Lin. 2023. Fuzzy Centered Explainable Network for Reinforcement Learning. *IEEE Transactions on Fuzzy Systems* 32, 1 (2023), 203–213.
 - [54] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. 2018. Predicting the Driver’s Focus of Attention: the DR (eye) VE Project. *IEEE transactions on pattern analysis and machine intelligence* 41, 7 (2018), 1720–1733.
 - [55] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. 2005. Components of bottom-up gaze allocation in natural images. *Vision research* 45, 18 (2005), 2397–2416.
 - [56] Martina Poletti, Michele Rucci, and Marisa Carrasco. 2017. Selective attention within the foveola. *Nature neuroscience* 20, 10 (2017), 1413–1417.
 - [57] Dean A Pomerleau. 1988. Alvin: An autonomous land vehicle in a neural network. *Advances in neural information processing systems* 1 (1988).

- [58] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. 2021. Stable-Baselines3: Reliable Reinforcement Learning Implementations. *Journal of Machine Learning Research* 22, 268 (2021), 1–8. <http://jmlr.org/papers/v22/20-1364.html>
- [59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 234–241.
- [60] Akanksha Saran, Ruohan Zhang, Elaine Schaertl Short, and Scott Niekum. 2020. Efficiently guiding imitation learning agents with human gaze. *arXiv preprint arXiv:2002.12500* (2020).
- [61] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2015. Prioritized experience replay. *arXiv preprint arXiv:1511.05952* (2015).
- [62] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR abs/1707.06347* (2017). [arXiv:1707.06347](http://arxiv.org/abs/1707.06347) <http://arxiv.org/abs/1707.06347>
- [63] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.
- [64] Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal* 27, 3 (1948), 379–423.
- [65] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. 2023. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In *Conference on Robot Learning*. PMLR, 726–737.
- [66] Hao Shao, Letian Wang, Ruobing Chen, Steven L Waslander, Hongsheng Li, and Yu Liu. 2023. ReasonNet: End-to-End Driving with Temporal and Global Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13723–13733.
- [67] Yuan Shen, Niviru Wijayarathne, Pranav Sriram, Aamir Hasan, Peter Du, and Katherine Driggs-Campbell. 2022. CoCAtt: A cognitive-conditioned driver attention dataset. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 32–39.
- [68] Brook A Shiferaw, David P Crewther, and Luke A Downey. 2019. Gaze entropy measures detect alcohol-induced driver impairment. *Drug and alcohol dependence* 204 (2019), 107519.
- [69] Brook A Shiferaw, Luke A Downey, Justine Westlake, Bronwyn Stevens, Shantha MW Rajaratnam, David J Berlowitz, Phillip Swann, and Mark E Howard. 2018. Stationary gaze entropy predicts lane departure events in sleep-deprived drivers. *Scientific reports* 8, 1 (2018), 2220.
- [70] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. 2017. Mastering the game of go without human knowledge. *nature* 550, 7676 (2017), 354–359.
- [71] Gustavo Silvera, Abhijat Biswas, and Henny Admoni. 2022. DReyeVR: Democratizing Virtual Reality Driving Simulation for Behavioural & Interaction Research. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 639–643.
- [72] Miriam Spering. 2022. Eye movements as a window into decision-making. *Annual review of vision science* 8 (2022), 427–448.
- [73] Neville A Stanton, Paul M Salmon, Guy H Walker, and Maggie Stanton. 2019. Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian. *Safety Science* 120 (2019), 117–128.
- [74] Michael J Swain and Dana H Ballard. 1991. Color indexing. *International journal of computer vision* 7, 1 (1991), 11–32.
- [75] Salah Taamneh, Panagiotis Tsiamirytzis, Malcolm Dcosta, Pradeep Buddhharaju, Ashik Khatri, Michael Manser, Thomas Ferris, Robert Wunderlich, and Ioannis Pavlidis. 2017. A multimodal dataset for various forms of distracted driving. *Scientific data* 4, 1 (2017), 1–21.
- [76] Yujin Tang, Duong Nguyen, and David Ha. 2020. Neuroevolution of self-interpretable agents. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*. 414–424.
- [77] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. 2020. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 7153–7162.
- [78] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [79] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. 2020. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 24–25.
- [80] Zhe Wang, Helai Huang, Jinjun Tang, Xianwei Meng, and Lipeng Hu. 2022. Velocity control in car-following behavior with autonomous vehicles using reinforcement learning. *Accident Analysis & Prevention* 174 (2022), 106729.
- [81] Chuan Wen, Jierui Lin, Jianing Qian, Yang Gao, and Dinesh Jayaraman. 2021. Keyframe-Focused Visual Imitation Learning. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*. PMLR, 11123–11133.
- [82] Nathan J Wispinski, Jason P Gallivan, and Craig S Chapman. 2020. Models, movements, and minds: bridging the gap between decision making and action. *Annals of the New York Academy of Sciences* 1464, 1 (2020), 30–51.
- [83] Jeremy M Wolfe. 2010. Visual search. *Current biology* 20, 8 (2010), R346–R349.
- [84] Jingda Wu, Zhiyu Huang, Zhongxu Hu, and Chen Lv. 2023. Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering* 21 (2023), 75–91.
- [85] Jingda Wu, Zhiyu Huang, Wenhui Huang, and Chen Lv. 2022. Prioritized experience-based reinforcement learning with human guidance for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems* (2022).
- [86] Jingda Wu, Yanxin Zhou, Haoan Yang, Zhiyu Huang, and Chen Lv. 2023. Human-guided reinforcement learning with sim-to-real transfer for autonomous navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2023).
- [87] Ye Xia, Danqing Zhang, Jinkyu Kim, Ken Nakayama, Karl Zipser, and David Whitney. 2019. Predicting driver attention in critical situations. In *Computer Vision—ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part V* 14. Springer, 658–674.
- [88] Enze Xie, Wenhui Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems* 34 (2021), 12077–12090.
- [89] Lantao Yu, Tianhe Yu, Jiaming Song, Willie Neiswanger, and Stefano Ermon. 2023. Offline imitation learning with suboptimal demonstrations via relaxed distribution matching. In *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. 11016–11024.
- [90] Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, and Matthieu Cord. 2022. Explainability of deep vision-based autonomous driving systems: Review and challenges. *International Journal of Computer Vision* 130, 10 (2022), 2425–2452.
- [91] Vinicius Zambaldi, David Raposo, Adam Santoro, Victor Bapst, Yujia Li, Igor Babuschkin, Karl Tuyls, David Reichert, Timothy Lillicrap, Edward Lockhart, et al. 2018. Deep reinforcement learning with relational inductive biases. In *International conference on learning representations*.
- [92] Luxin Zhang, Ruohan Zhang, Zhuode Liu, Mary Hayhoe, and Dana Ballard. 2018. Learning attention model from human for visuomotor tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [93] Ruohan Zhang, Zhuode Liu, Mary M Hayhoe, and Dana H Ballard. 2017. Attention guided deep imitation learning. *Cognitive Computational Neuroscience (CCN)* (2017).
- [94] Ruohan Zhang, Zhuode Liu, Luxin Zhang, Jake A Whritner, Karl S Muller, Mary M Hayhoe, and Dana H Ballard. 2018. Agil: Learning attention from human for visuomotor tasks. In *Proceedings of the european conference on computer vision (eccv)*. 663–679.
- [95] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl Muller, Jake Whritner, Luxin Zhang, Mary Hayhoe, and Dana Ballard. 2020. Atari-head: Atari human eye-tracking and demonstration dataset. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 6811–6820.
- [96] Boyuan Zheng, Sunny Verma, Jianlong Zhou, Ivor W Tsang, and Fang Chen. 2022. Imitation learning: Progress, taxonomies and challenges. *IEEE Transactions on Neural Networks and Learning Systems* 99 (2022), 1–16.
- [97] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *Computer Vision and Pattern Recognition*.