Gaze Supervision for Mitigating Causal Confusion in Driving Agents

Abhijat Biswas[1], Badal Arun Pardhi[1], Caleb Chuck[3],
Jarrett Holtz[2,3], Scott Niekum[4], Henny Admoni[1], and Alessandro Allievi[2,3]

***Abstract*— Imitation Learning (IL) algorithms such as behavior cloning are a promising direction for learning human-level driving behavior. However, these approaches do not explicitly infer the underlying causal structure of the learned task. This often leads to misattribution about the relative importance of scene elements towards the occurrence of a corresponding action, a phenomenon termed *causal confusion* or *causal misattribution*. Causal confusion is made worse in highly complex scenarios such as urban driving, where the agent has access to a large amount of information per time step (visual data, sensor data, odometry, etc.). Our key idea is that while driving, human drivers naturally exhibit an easily obtained, continuous signal that is highly correlated with causal elements of the state space: eye gaze. We collect human driver demonstrations in a CARLA-based VR driving simulator, DReyeVR, allowing us to capture eye gaze in the same simulation environment commonly used in prior work. Further, we propose a contrastive learning method to use gaze-based supervision to mitigate causal confusion in driving IL agents — exploiting the relative importance of gazed-at and not-gazed-at scene elements for driving decision-making. We present quantitative results demonstrating the promise of gaze-based supervision improving the driving performance of IL agents.**

I. INTRODUCTION

Imitation learning (IL) is a popular method for learning urban driving policies due to its ease of implementation and de-coupling of the data collection/action step and the training step by allowing offline learning of control, among other factors. However, it does not explicitly model the underlying causal structure of the task, instead inferring causality from strongly correlated elements of the state space that occur before specific actions are performed. This results in a policy that does the right things for the wrong reasons in the training distribution and thus doesn't generalize well at test time.

While a policy is a causal function mapping observations to particular actions, it is often difficult to identify which underlying state variables caused the policy to take a particular action. This is made more complex by the fact that instead of observing the state directly, the policy uses observations—i.e camera images, which are themselves a function of the state variables. It is then difficult to attach a particular state variable to any particular pixel. Causal confusion then arises when an undesirable state variable triggers a particular action.

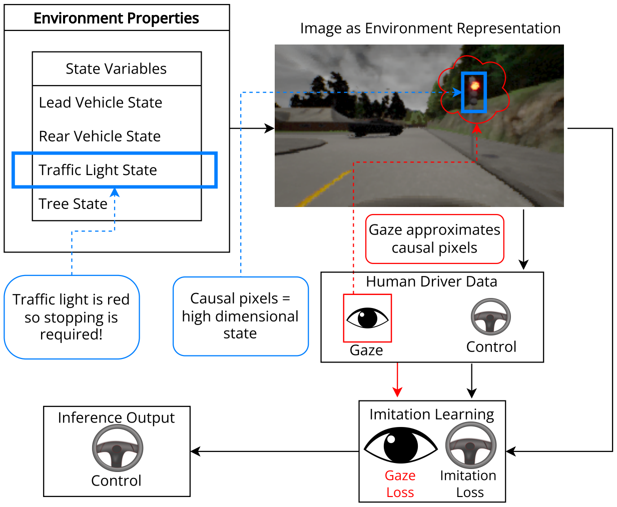


Fig. 1: Outline of training for imitation learning with gaze as a supervisory signal for approximating causal information in the scene. Here gaze pixels represent a noisy approximation of the causal pixels in the RGB image space, while the causal pixels are an abstraction of the causal scene elements. Gaze is only required at train time.

In the original paper identifying causal confusion [1], the authors use an example of a realistic driving setting to illustrate this phenomenon where, counter-intuitively, access to more information yields poorer task performance by the imitation learning agent. In the aforementioned example, an IL agent learns from demonstration images from inside the cab of a vehicle with and without a brake light indicator on the dash. When a brake light is present (and always on when the brake is applied), the agent may learn to brake only when the brake light indicator is on. This is an undesirable misattribution of cause and effect.

Several additional works exist in which history-based imitation models perform worse than their counterparts without access to this historical information, especially during driving (see Sec 4.7 from [2] for a review). For instance, in [3], imitation learning policies are trained with and without ”history” information about the car's past trajectory as input. The model with history has better performance on held-out demonstration data but much worse performance when deployed, indicating that causal confusion is occurring. Another example is in [4], where they identify a failure mode where the training data displays a tendency to stay static at a full stop, leading to a strong correlation between low speed and

[1] These authors are with the Robotics Institute, Carnegie Mellon University. Biswas (abhijat@cmu.edu) is the corresponding author.
[2] Holtz and Allievi are with Robert Bosch LLC.
[3] These authors are affiliated with The University of Austin, Texas.
[4] Niekum is with the University of Massachusets, Amherst.

not accelerating in the final policy. This leads to the final policy staying stopped at stop lights even after they turn green.

The most straightforward resolution to the causal confusion problem would be to simply learn the correct underlying causal structure. De Haan *et al.* propose methods using targeted interventions (DAgger-like expert queries [5] or GAIL-like environment interaction [6]) to prune a set of 2^N causal hypotheses where N is the dimensionality of the state space. This is a large search space for visuomotor tasks like driving, where the state-space dimensionality is often in the millions. Moreover, expert queries or environment interaction in the training loop can often be too expensive to be a feasible solution.

Taking a complementary approach, we seek to use a signal that human drivers naturally exhibit while operating vehicles which is highly correlated with causal parts of the state space — eye gaze. Our idea is to use driver eye gaze as a supervisory signal, alongside driving control, to highlight the lower dimensional parts of the (very high-dimensional) visual state space that the driver fixated on before making their driving decision. Specifically, we use a contrastive learning formulation to encourage visuomotor IL driving policies to change driving decisions based on visual information in the fixated-at regions. This gaze supervision seeks to mitigate causal confusion by directing the causal function of the policy towards the variables of the observation (clusters of pixels), which correspond to an underlying state variable that the human believes is causal to the optimal behavior, as outlined in Fig. 1.

The benefit of using eye gaze from human driving demonstrators is that it is essentially "free", *i.e.* it is a signal that is naturally exhibited by humans as they drive. Importantly, it does not require additional labeling or intervention from human experts and is non-intrusive, with gaze data being able to be collected with a pair of wearable glasses or even in-cabin sensors. In fact, some data-collection vehicles are already instrumented with cabin-facing visual or infrared sensors, that can be used to obtain traffic-scene registered eye gaze directly. For this work, we used eye gaze collected from a driver using a VR simulator with eye tracking built into the VR headset (Fig. 3).

We propose a gaze-based contrastive supervision method to incorporate driver gaze into policy training and show that finetuning a pre-trained IL driving policy using our method results in better driving performance than the pre-trained model. Our formulation encourages the trained policy network's driving actions to be affected by gazed-at regions. Further, the fine-tuned method's saliency better matches drivers' attention as indicated by their gaze. In summary, we investigate the utility of natural driver eye gaze-based supervision as a tool for mitigating causal confusion in imitation learning-based driving agents. We contribute:

- a novel dataset of human driving with actions and associated eye gaze in a CARLA-based VR simulation,
- a gaze-based contrastive supervision formulation for fine-tuning learned models for gaze, and
- experimental results quantifying the performance improvements in simulated driving scenarios with gaze supervision.

II. Related Work

Gaze is a common non-verbal physiological signal used to infer operators' intent. In robotics, gaze has been used widely from understand people's preferences over objects [7], [8] for facilitating smoother handovers, to facilitating shared autonomy by inferring goals in an assistive feeding task [9] or the next step in a multi-stage manipulation task [10]. Gaze has also been used as a supervisory signal. In work with simulated agents, the Atari-HEAD dataset represented a high-quality dataset with expert human performance and gaze [11] which subsequent works [12], [13] took advantage of to improve Atari game playing performance. In the driving context, Xia *et. al* show that gaze-informed models can predict ego vehicle speed in pedestrian-involved situations more effectively than non gaze models [14].

While the above survey indicates wide-ranging use of gaze as a supervisory signal, public datasets with driver gaze paired with driving action and sensor information are uncommon. The closest two works to ours in the driving domain, "A Gaze Model Improves Driving" [15] and Gaze Modulated Dropout [16] have several differences from us: First, both formulations require gaze information from the driver at test time. Since gaze is not available from a human source at test time, they use a learned gaze model to predict gaze maps, which introduces additional learnable parameters and sources of error. Second, their domain is limited to highway driving, which mostly involves lane maintenance and occasional overtaking, a far simpler driving paradigm than urban driving featured in our dataset. Third, the gaze data was collected by showing participants highway driving on a 23 inch screen which limits the naturalness of demonstrator gaze by constraining the field of view and not allowing head movement to change the viewpoint. In contrast, the data collected in the DReyeVR simulator allows participants to naturally move their viewpoint by moving their head in VR. Fourth, their gaze & driving data is not available publicly whereas our data is publicly available in a replayable format that allows arbitrary visual sensing configurations from any perspective to be generated posthoc which can be used by models using varied sensor configurations.

The closest work to ours outside the driving domain is by Saran *et al* [13] where the authors explored the use of human gaze to guide imitation learning agents. Saran *et al.* propose a gaze prediction task as an additional head on an imitation learning agent. In their design, the gaze prediction head is attached to the last convolutional layer of the policy network and a $1x1$ convolution is used to predict the associated gaze map at that resolution (generally much smaller than input). While this method has the some similar properties as ours, such as not adding any additional learnable parameters for the deployed model, it has one major difference: it requires an intervention in the model architecture at the last convolutional layer. To understand the limits of such an intervention, consider an object-based driving policy (such as the recent PlanT [17]) where there are no convolutional layers at all and traffic and route elements are input as tokens. In such a setting, the gaze coverage loss would be difficult to apply, but our contrastive training could be performed by masking out tokens at not-gazed-at or gazed-at locations. We do not perform evaluations on PlanT since it directly takes in the state of relevant traffic lights from the simulator and is inadmissible on the CARLA leaderboard.

Their work studied the effect of gaze on improving performance on a larger gaze dataset and simpler domain, 600 minutes of player gaze data during ATARI gameplay, as opposed to about 70 minutes in the more complex and dynamic simulated sensor-based driving environment of CARLA/DReyeVR. Finally, while they did explore the effect of gaze supervision in mitigating causal confusion, the experiment has some differences from ours. That work relies on a manually constructed causal confusion trap, intentionally localizing historical state information to a specific and fixed area of the state space. In contrast, our work deals with a more dynamic domain, where state information that may lead to causal confusion is not confined to specific regions, posing a more complex and naturally-occurring problem.

III. Causal confusion in Imitation Learned driving agents

As an example algorithm for exploring causal confusion in IL-based driving agents, we consider the popular and well-established Learning by Cheating (LBC) [18] model for autonomous urban driving in CARLA. This method uses a two-step approach where a teacher model is first trained with access to ground truth, overhead-view semantic segmentation maps around the ego-vehicle (approaching perfect perception). Then, this agent is used as an oracle to train a sensorimotor agent that only has access to RGB (*left/center/right* views) sensor data and a high-level command from a global plan. We use the latest author-provided code and model [19], which is a slight deviation from the original paper [18]. Of particular note, the LBC sensorimotor model takes a ten-channel image as input where nine channels correspond to three RGB images (*left, center, right*) and the last is a heatmap with the only "hot" region being a Gaussian distribution centered at the next waypoint in the frame of the *center* camera.

As one may expect, the LBC model also shows symptoms of suffering from causal confusion. Anecdotal descriptions from the authors can be found in [20]: "*I believe that the network has some issues with starting/stopping*". These problems seem to occur primarily in the absence of surrounding vehicles which may be wrongly used as causal cues.

We especially notice traffic light infractions where the LBC agent either does not stop for a red light or fails to restart after stopping at a red light. We also observe cases where the agent stops at a red light but restarts when opposing traffic moves, even though the red light has not changed.

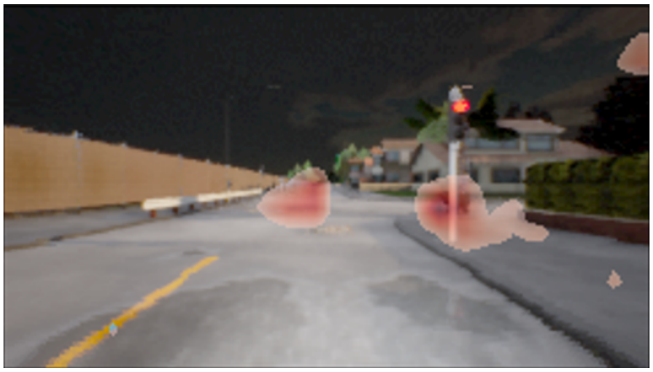


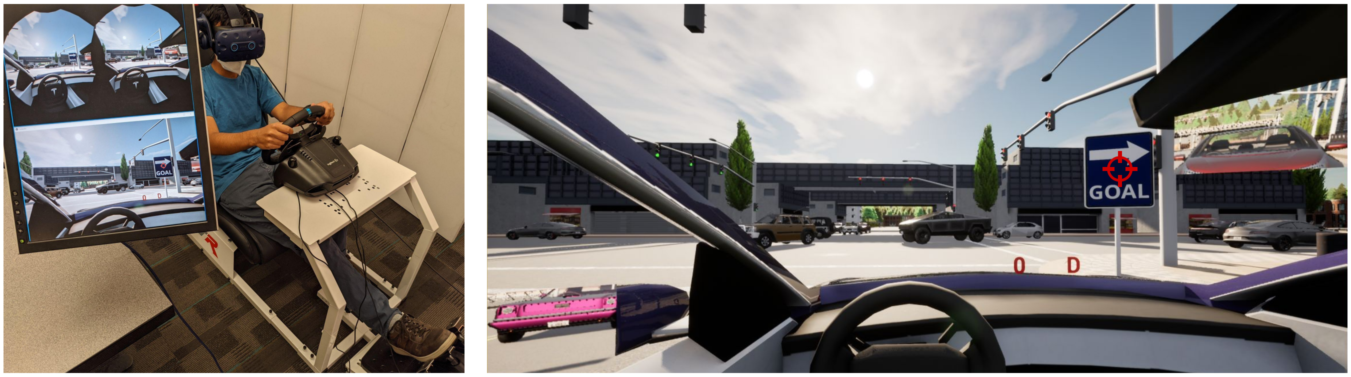
Fig. 2: Salience map generated by using blur-based saliency (details: Sec. III-A) for the pre-trained Learning by Cheating IL model [18]. The center input image to the model is shown with salience overlaid. This scene depicts the instant a vehicle comes to a stop after which it fails to restart. Notably, the bulk of salience weight is at the traffic light's base — a non-causal region which indicates causal misattribution.

In more recent works such as Transfuser [21] and PlanT [17], we see special heuristics or inputs to tackle problems due to causal confusion. The Transfuser agent uses a "creeping" behavior that gives the ego vehicle a small velocity if it has not moved in 55 seconds. It is used along with a safety heuristic that stops the car if there is a lidar hit in front of the vehicle. The PlanT model uses a privileged input to deal with this problem: it directly reads the state of the relevant traffic light from the simulator which is a lower dimensional causal signal. Consequently, it is also not validated on the official leaderboard. In this work, we are focused on exploring gaze as a supervisory signal that can act as a proxy for high dimensional causal variables, so we performed our experiments with a popular, validated method (LBC) which does not have associated heuristics.

A. Saliency-based causal confusion diagnosis

To investigate the relative importance of regions of the input state space in making decisions, we used a saliency method to investigate the decision-making process of the LBC model. Specifically, we used the blur-based saliency method by Greydanus *et al.* [22]. This method is designed to identify which spatial regions of the visual input are most salient for the action produced by a given deep visuomotor policy network. The method is network architecture agnostic and works by blurring different regions of the given visual input and measuring the difference in output with the original input. It reasons that input image regions, which, when blurred, cause the greatest difference in the agent's policy network output, are the most salient.

The original method was designed to work with Atari playing agents with inputs that were a single image per timestep. A uniform grid is sampled over each input image and a blur centered at each sample location is performed independently to construct the modified inputs for salience calculation. However, in the typical CARLA-based driving



(a) Physical setup with participant driver in driving pose, alongside experimenter's setup monitoring the simulation.

(b) First person DReyeVR simulator perspective during the same episode with eye reticle (red crosshair) denoting eye gaze on in-world navigational sign that gives drivers route direction. The crosshair is only for illustration (not shown in VR).

Fig. 3: Experimental setup during gaze data collection.

formulation, at each timestep, multiple images with varying amounts of overlap are used to represent the scene. Consequently, we modify the aforementioned method such that the blur salience is calculated only for the middle image. However, when a portion of the middle image is blurred, we also blur the corresponding region in the left (or right, as applicable) image so that the target space is blurred when input to the policy network.

Using this blur-based saliency measure, we can generate saliency maps for the LBC method, such as in Fig 2. The figure shows a vehicle stopped at a red light the frame before it turns green. Here, most of the salience lies erroneously on non-causal parts of the input image, such as the base of the traffic light.

B. *Relationship Between Gaze Supervision and Causal Confusion*

The causal relationship described in Fig. 1 can be formalized in terms of a Structural Causal Model (SCM). The state space $\mathcal{S}$ can be represented as a set of n causal variables $\mathcal{S} := \{\mathcal{S}_1, \ldots, \mathcal{S}_n\}$, representing other vehicles, pedestrians, time of day, trees, etc., while the random variable S represents the distribution over particular state values. The observation space $\mathcal{O}$, is determined by a causal function f of the underlying state $f : \mathcal{S} \to \mathcal{O}$. Here, the random variable O represents the distribution of observation values. Given a state, a particular observation variable O_i will not have equivalent counterfactual dependence on all state variables. This means there exists some causal variable S_i where an intervention to s_i' will have a significant effect on the distribution of observation variable O_i: $P(o_i|s, \text{do}(S_i = s_i)) \neq P(o_i|s, \text{do}(S_i = s_i'))$. There also exists causal variable S_j which has little effect on the distribution of O_i: $P(o_i|s, \text{do}(S_j = s_j)) \neq P(o_i|s, \text{do}(S_j = s_j'))$. In practice, this is the assumption we make about gaze: that the cluster of gaze pixels $g \in \mathcal{G}$, which is a collection of $\{o_i \quad \forall i \in \mathcal{G}\}$ is causally dependent on a state variable that the optimal policy should depend on.

Next, we can observe the causal relationship between the policy π and the observation O. A policy is a mapping from observation to actions: $\pi : \mathcal{O} \to \Phi$, where Φ is the space of actions. We abuse notation by representing the random variable for actions as Φ. For a cluster of gaze pixels $g \in \mathcal{G}$, our method applies two interventions in the form of perturbations on the observation $p_{+/-} : \mathcal{O} \times \mathcal{G} \to \mathcal{O}$, where p_- applies the perturbation to the gaze regulated components of the observation g, and p_+ applies the perturbation to the gaze unregulated components $\bar{g}$. By applying these interventions, the loss ensures that $P(\Phi = \pi(o)|\text{do}(O = p_+(o, g))) - P(\Phi = \pi(o)|\text{do}(O = p_-(o, g))) > \alpha$. Intuitively, this enforces the policy to be robust to perturbations of the non-gazed variables and sensitive to perturbations of the gaze variables.

Combining these insights, our operation enforces both interventional dependence on the state variables that generate the gaze and independence on the state variables not gazed at. In other words, for a gazed state variable $S_i = s_i$, $P(\Phi = \phi|\text{do}(S_i = s_i)) \neq P(\Phi = \phi|\text{do}(S_i = s_i'))$ for some values of s_i', and for ungazed state variable $S_k = s_k$, $P(\Phi = \phi|\text{do}(S_k = s_k)) = P(\Phi = \phi|\text{do}(S_k = s_k'))$. By aligning which causal state variables the policy is dependent on to salient features, our method reduces causal confusion.

IV. Method

Our method seeks to use human drivers' natural eye gaze as a supervisory signal for imitation-learned driving agents to help mitigate causal confusion. We collect driving demonstrations and driver eye gaze in a VR based driving simulation and incorporate gaze supervision as an auxiliary contrastive loss to existing driving imitation policies.

A. *Gaze data collection*

Human demonstration data was collected in the DReyeVR simulator [23], a modified version of the CARLA simulator to enable human driving in VR. DReyeVR also enables the collection of driver eye gaze as they use the simulator. Drivers were tasked with completing a navigational sign following task (see Fig. 3b), and their driving actions (steering, throttle, brake), as well as eye gaze movements, were recorded.

Eye gaze was collected at the simulator rate, about 50Hz. Eye gaze can be a noisy and high-frequency signal. To correct for this, we performed pre-processing in the following manner. First, driver eye gaze movements were classified into low-velocity fixations and high-velocity saccades using I-BMM, an off-the-shelf gaze event classifier [24]. Then, saccades were discarded (during saccades, drivers are moving their eyes between fixation points and cannot pay attention to the point of regard). Finally, fixations were aggregated into attention maps by initializing a Gaussian distribution ($\sigma = 2$) centered at each fixation point and aggregating these across a 15 second window of gaze history. LBC uses data from the simulator at 2Hz meaning each frame's associated attention map was composed of a maximum of 30 gaze points (some are discarded due to being saccades). This eye gaze was obtained in the form of 3D gaze coordinates in the virtual CARLA world, allowing us to project the gaze point-of-regard to virtual cameras in the world (such as the *left, center, right* images taken in as input by LBC (see Fig. 6).

We use data from $N = 7$ drivers, all of whom had held a US driver's license for more than one calendar year. Each participant drove five routes, with the first being for acclimatization to the VR simulator (this data was not used). However, three participants were unable to complete all four experimental routes due to motion-sickness in the simulator and four routes had to be discarded due to improper data recording. This data collection was approved by the relevant institutional review board. In total, we used 17 routes with about four minutes of driving data each or about 70 minutes of data. This is much lower than the auto-generated data used to train the LBC models, which is upwards of 350 minutes. We will release our collected gaze and driving data in a format that will make them fully replayable in the CARLA-based DReyeVR simulator allowing future users to generate data from arbitrary virtual sensor configurations with associated gaze. In rest of this paper, this dataset is referred to as the DRVR dataset.

We also experiment with synthetic gaze generated using a heuristic policy. Human gaze is expensive to collect, so simulated gaze allows us to investigate the efficacy of our gaze-based supervision method with synthetic datasets. Since the human gaze signal is modeled with a Gaussian distribution centered at the fixation point, the heuristic method generates fixation points using a probabilistic state machine (Fig. 4). Intuitively, the simulated gaze checks which objects are in the scene, fixates on an object for some time, or until the object leaves the frame, and then fixates on a new object, favoring more important objects. Specifically, fixation stays in a state with initial probability $p = 0.99$ that decays by 0.01 every timestep of the fixation. If the simulated gaze switches, it will choose the next fixation from a hierarchy: [”lead vehicle”, ”pedestrian”, ”stop sign”, ”traffic light”, ”vanishing point”, ”oncoming vehicle”], where it selects the highest priority object in the scene with probability $u = 0.8$, otherwise uniformly sampling from the lower priority objects. If the object of fixation moves beyond a maximum distance of 20m the agent resamples. The simulated gaze is generated at 2Hz, the same frame rate as the data collection. Using this state machine, we generated a driving dataset with synthetic gaze (and hence, attention maps) of the same size and towns as the DRVR dataset.

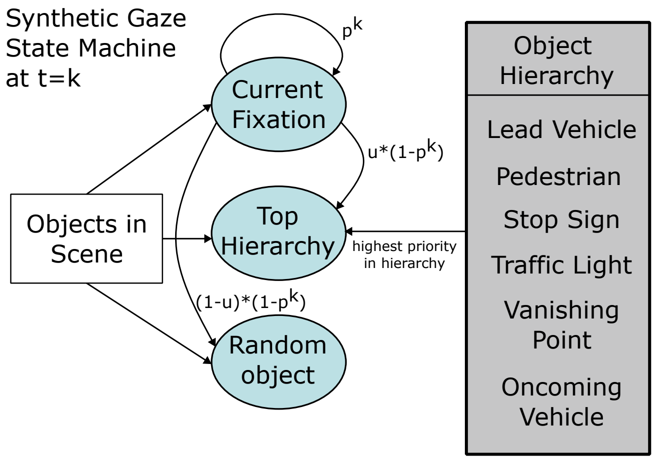


Fig. 4: Synthetic gaze generated by a state machine, shown at time $t = k$: after a transition to a new object, it becomes the current fixation. The process repeats at every timestep.

B. Gaze-based supervision via contrastive loss

As Fig. 2 shows, a large portion of the blur-based salience lies on the base of the traffic light – i.e. when blur is applied to this region, this causes the largest change in the LBC model's predicted output.

Our idea to provide gaze supervision comes from correcting this misplaced salience. We use a triplet loss and gaze-based salience as shown in Fig. 5. Given a policy network ϕ and a triplet of inputs (anchor x_a, positive sample x_+, negative sample x_-), a typical triplet loss formulation is below:

$$\begin{aligned} L_t(x_a, x_+, x_-) = \max ||\phi(x_a) - \phi(x_+)||_2 - \\ ||\phi(x_a) - \phi(x_-)||_2 + \alpha,\ 0) \end{aligned} \tag{1}$$

Here, α is the margin being enforced between the positive and negative inputs in the output space of the network ϕ. We follow the typical formulation where instead of computing (and hence enforcing) this margin in the output space, we compute it in the latent space of our policy network.

In our formulation, the original set of unblurred input images (*left, center, right*, waypoint) constitutes the anchor data point (above x_a, ours: x_{orig}). The negative input is constructed by applying Gaussian blur (same parameters as [22]) to important gazed-at scene locations (indicated by attention maps) in the same set of images (above x_+, our formulation: x_{gaze}). The corresponding positive point has the same blur applied to the unimportant scene regions (above x_-, our formulation: complement of attention maps $x_{!gaze}$). The reasoning for this formulation is as follows:

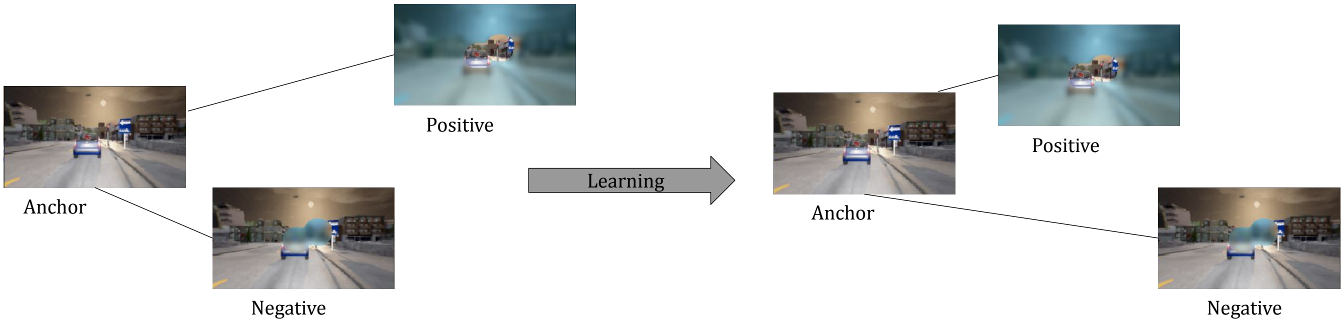


Fig. 5: Gaze-based supervision via triplet loss: during training, images with blur applied at non-gaze regions are moved closer to the anchor than those with blur applied at gazed-at regions. Input data points are represented here with center image but all three images are correspondingly blurred during training. See Fig 6 for the full input triplet in greater detail. Blue shading is added to illustrate the blurred region (not present in training input).

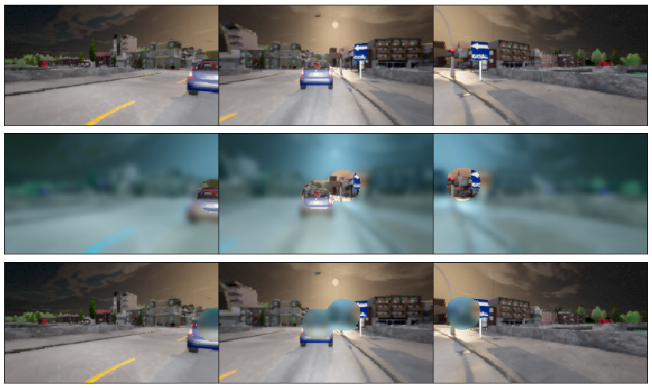


Fig. 6: Example triplet in more detail with gaze-contingent blur applied. In order (top-bottom): Anchor inputs (no blur), Positive image (blur in non-attention regions), Negative image (blur in attention regions). Blue shading is added to illustrate the blurred region (not present in training input).

the most important regions for decision making for actions lie in the gazed-at regions (as indicated by attention maps) and the non-gazed-at regions do not contain information that would change the driving decision. Hence, we can rewrite Equation 1 as follows:

$$\begin{aligned} L_t(x_a, x_+, x_-) = \max(\ & ||\phi(x_a) - \phi(x_+)||_2 \\ & - ||\phi(x_a) - \phi(x_-)||_2 + \alpha,\ 0) \end{aligned} \tag{2}$$

The triplet loss minimizes the distance between the anchor and the positive point while maximizing the distance between the anchor and the negative point. Hence, our loss enforces that visual inputs blurred in locations unimportant to driving should lead to a smaller change in network output than the same blur applied in important regions. An example triplet is shown in Fig. 6.

As explained in the paper [18], LBC training takes place in two steps: first, by learning a privileged agent that learns to drive with perfect sensor information and then, by using it to supervise a sensorimotor model that learns to "see" via RGB images. In this work, we focus primarily on mitigating causal confusion in the sensorimotor model since that is the one that learns the task with sensory inputs (and greater potential for causal confusion) and because it is the final deployed model. We also focus on fine-tuning the sensorimotor model rather than training from scratch since the amount of data with gaze supervision is much lower than the auto-generated driving data. We perform an ablation study by fine-tuning the sensorimotor agent using either the driving control supervision loss used by the LBC authors [19] (LBC) or via our proposed gaze-based triplet loss (Triplet).

In models that use the LBC loss, the choice of privileged model to use as the teacher to the sensorimotor model during training is another consideration. Here, we use the best-performing privileged agent model released by the authors of LBC (LBC best).

Finally, we must consider the data used for finetuning the pre-trained model. We use two primary datasets: data from the rule-based expert (RBE) and data from the human drivers in the DReyeVR simulator (DRVR). The RBE dataset is the one used by the original LBC training, containing about 600 minutes of trajectories driven by their handcrafted expert autopilot that leverages the internal state of the simulator to navigate through fine-grained hand-designed waypoints. The RBE data did not contain any associated human gaze since it was driven algorithmically. The DRVR dataset is much smaller, totaling about 70 minutes, and was driven by licensed US drivers. Every trajectory in this dataset had human gaze associated with it. However, the DRVR route contained only five unique routes, with repetitions of each route by different drivers. To mitigate the catastrophic forgetting effect of fine-tuning solely on the small DRVR dataset, we also explore a mixed paradigm — RBE+DRVR in Table. I. However, since the RBE data does not have associated gaze labels, when we finetune on the mixed data only the LBC control loss is active for trajectories sampled from this dataset. The DRVR dataset trajectories have both the LBC control supervision and our gaze-based supervision enabled. When the two losses are combined, the total loss is given by:

$$\texttt{Loss}_{tot} = \texttt{Loss}_{LBC} + \lambda \times \texttt{Loss}_{Triplet}$$

Empirically, we found $\lambda = 0.1$ to lead to the best performance and use it throughout our experiments below.

V. Experimental Results

We evaluate the effectiveness of our gaze-based supervision method in two parts: First, we show that after applying gaze supervision, the saliency of the model matches the attention maps from human drivers' gaze more closely. Second, we show that the imitation agent's driving performance improves after applying gaze supervision.

A. Model saliency

We wish to first investigate the effect of gaze-based supervision on the driving agent's saliency. Here, we use the same definition of model saliency as Sec. III-A and use the modified version of Greydanus *et al.* [22] described therein to compute saliency maps. For this experiment, we use routes driven by human subjects since those have associated ground truth gaze (and hence, attention) available for comparison. To run the imitated driving agents on the human drivers' driven routes, we replay those routes in the DReyeVR simulator while storing the required sensor outputs. Then, we use those sensor outputs as inputs to the driving agents to calculate their saliency in an off-policy fashion. This is done for both the pre-trained and the gaze-supervised model to evaluate their salience comparatively. For this evaluation, both human gaze-based attention maps and saliency maps are first binarized and then compared to each other. We use Intersection-over-Union as the metric to evaluate both the sensitivity and the specificity of the salience maps compared to the gaze-based attention maps. Higher IOUs indicate more similarity of the agents' salience (*i.e.*, scene elements that affected the agents' actions) with the human gaze.

In our results (Table I), we see that fine-tuning with gaze supervision does indeed improve the IOU of model saliency maps (*i.e.*, they better match the true attention maps from human demonstrators). While fine-tuning solely with human data and gaze supervision improves IOU the most, it also leads to severely diminished driving performance. Fine-tuning with Mixed data and both gaze and control supervision achieves a good balance of both IOU and driving performance. Lastly, using synthetic gaze does lead to good driving performance but does not match the human demonstrators gaze as well which is to be expected as they are signals with different characterstics (synthetic gaze has less noise and causal fixations with more probability).

Some qualitative examples of the LBC model's salience before and after gaze-based supervision can be seen in Fig. 7 (fine tuned on Mixed data). This shows that more of the policy network's actions are dictated by causal sets of pixels (as indicated by gaze) post gaze supervision, as expected.

B. Driving performance

To evaluate the driving performance of our fine-tuned models, we used the Longest6 benchmark [21]. The test set contains 36 driving routes each with a unique combination of weather and daylight conditions. The routes are spread over 6 virtual towns, of which 2 are unseen in the training data. We used the Longest6 benchmark instead of the CARLA Leaderboard since the latter eval set is only available through their online portal which restricts teams to 2 evaluations per month making it unsuitable for ablation studies.

TABLE I: Driving performance and Model saliency IoU on the Longest6 [21] benchmark. Base model for all rows is LBC [18]. Abbreviation guide: RBE = demonstrations from LBC's rule based expert driver; DRVR = demonstrations from human drivers in the DReyeVR simulator; Mix = RBE + DRVR; Synth = RBE demonstrations with synthetic gaze

Training approach	Training data	Loss used	DS (↑)	IoU (↑)
Pre-trained [18]	RBE	LBC	7.01	0.13
Human (gaze only)	DRVR	Triplet	0.42	0.22
Human (control & gaze)	DRVR	LBC+Triplet	4.82	0.19
Mixed (control only)	Mix	LBC	7.81	0.12
Mixed (control & gaze)	Mix	LBC+Triplet	9.61	0.18
Synthetic (control & gaze)	Synth	LBC+Triplet	10.76	0.13

We use the $DrivingScore$ metric from the Carla Leaderboard, which is calculated as the average of $RouteCompletionPercentage \times InfractionScore$ per route. Here $RouteCompletionPercentage$ is the percentage of the route completed by the driving agent, and $InfractionScore$ is a number in $[0, 1]$ that encapsulates the number of infractions committed by the driving agent. $InfractionScore$ starts at 1 for each route and progressively decreases per infraction (we refer readers to [25] for details). Hence, the maximum achievable $DrivingScore$ would be 100.

From experiments investigating the agent's driving performance, the first noticeable trend is driving score degradation due to fine-tuning on solely the human demonstrator driving data (DRVR) with either solely gaze supervision and a combination of gaze supervision and LBC driving control supervision. This may be due to the much smaller size and different characteristics of the DRVR data compared to that auto-generated by the rule-based expert (RBE) which can cause catastrophic forgetting of the original training of the LBC model.

Expectedly, using just control supervision on the Mixed (RBE + DRVR) dataset does improve performance over the vanilla pre-trained LBC model since it sees more training data than just the RBE dataset. However, finetuning using Mixed data using the both gaze and control losses in conjuction leads to the better driving performance than using control loss alone.

Finally, using the Synthetic dataset (with gaze generated by the Synthetic gaze state machine in Sec. IV-A) in the combined loss paradigm leads to better performance than the real gaze dataset. This can be explained by considering the role of gaze as a causal signal proxy as discussed in Sec. III-B. Namely, both real and synthetic gaze are a proxy for causal signals but inherently the synthetic gaze is a ”cleaner”

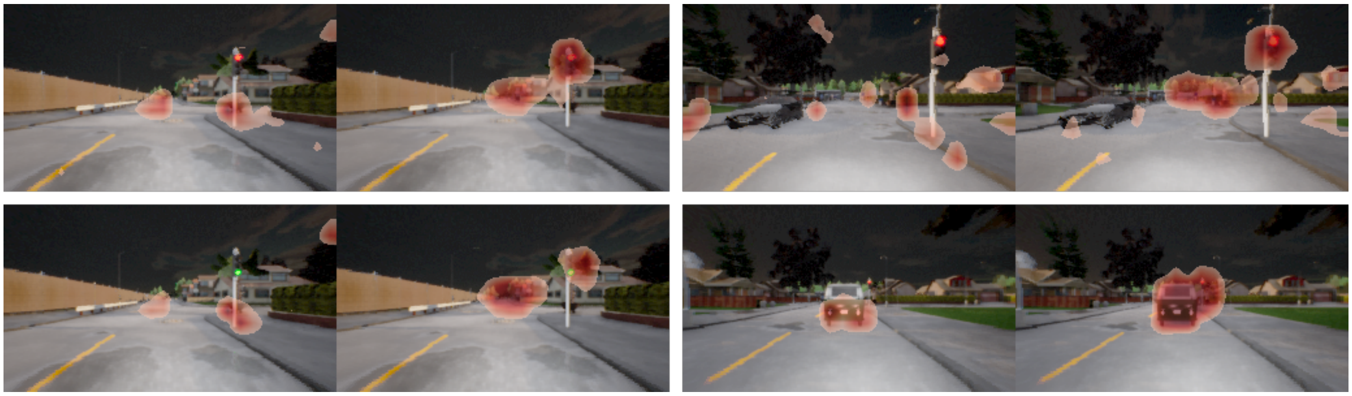


Fig. 7: Model blur-based saliency example pairs (see Sec. III-A for salience computation details) showing LBC model salience before and after gaze-based supervision. Within each pair, left images correspond to the pre-trained model with no gaze-based training and right images correspond to our model fine-tuned on the "Mix" dataset with gaze and control supervision. Overall, the gaze supervised model's salience lies more on causal scene elements such as traffic lights and the space in front of the vehicle.

proxy since it is explicitly generated by us with object fixations in mind compared to real gaze which is naturally exhibited by human drivers. However, since synthetic gaze is generated using privileged simulator information (exact 3D location and semantic labels of objects) while generating the training routes, it cannot be made available in the real world with the same accuracy while real gaze can.

VI. Discussion & Future work

There are limits to the formulation of gaze-based causal confusion mitigation. For instance, we assume that a collection of local pixels is generated by a particular causal variable, though some causal variables, such as lighting, cannot be easily attached to any particular pixel or group thereof. Additionally, we assume that the intervention of blurring is a sufficient operation to ensure robustness, though this pushes the perturbed images out of the normal image space. In the construction of triplets for gaze supervision, deleting objects that are not gazed at is a more direct way of reflecting their absence in the positive sample than simply blurring them out. This deletion could be done via techniques such as partial convolutions to block out image regions as in [26]. Finally, drivers can use their peripheral vision to monitor stimuli that are away from their foveal vision [27]. Our method currently does not account for drivers using solely their peripheral vision to attend to and monitor causal groups of pixels, which is unlikely compared to a mix of foveal and peripheral attention.

Lastly, the LBC model is now no longer a top performing model on the leaderboard. Ideally, we would like to perform our experiments with the newer, better performing models such as Transfuser [21] or Learning from All Vehicles [18]. However, those models are much more data hungry than LBC, requiring two magnitudes of data more than LBC and performing very poorly when trained on LBC-sized data (*e.g.* $DS = 0.53$ for Transfuser if trained with data comparable to LBC training dataset size). In addition, these models incorporate additional input modalities and special heuristics that complicate the data collection and integration while making it more difficult to attribute causal confusion mitigation to any one part of the algorithm. Ultimately, our goal is not to top the CARLA leaderboard but to show that gaze-based training can help IL agents better identify causal portions of high-dimensional state spaces in a complex task like driving. Our contribution is still valuable since in the real world, it is only marginally more expensive to instrument a modern vehicle with gaze-tracking hardware while collecting driving demonstrations to train an autonomous driving system. This is because eye gaze is exhibited by human drivers naturally while performing the driving task and it can be used directly (with straightforward pre-processing) in our method.

VII. Conclusion

We proposed a gaze-based supervisory formulation that improves driving performance of IL agents by mitigating causal confusion. Our method uses the insight that human drivers' gaze is associated with lower dimensional portions of otherwise high dimensionsional inputs which are also causally relevant. We collected a novel dataset of human drivers who used a VR driving simulator to provide demonstrations on CARLA routes while recording their eye gaze. This data was used to finetune an IL driving model with our gaze-based contrastive loss and improved its driving performance. We commit to making our dataset public and hope that it can spark new research directions combining physiological signals and learning given its tight integration with the widely used CARLA simulator. We see gaze based causal confusion mitigation as a promising direction, not just for driving but also for other domains where operator gaze is associated with causal sub-spaces of the state space.

References

[1] P. De Haan, D. Jayaraman, and S. Levine, "Causal confusion in imitation learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *arXiv preprint arXiv:2306.16927*, 2023.

[3] D. Wang, C. Devin, Q.-Z. Cai, P. Krähenbühl, and T. Darrell, "Monocular plan view networks for autonomous driving," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2876–2883, IEEE, 2019.

[4] F. Codevilla, E. Santana, A. M. López, and A. Gaidon, "Exploring the limitations of behavior cloning for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9329–9338, 2019.

[5] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 661–668, JMLR Workshop and Conference Proceedings, 2010.

[6] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[7] C.-M. Huang and B. Mutlu, "Anticipatory robot control for efficient human-robot collaboration," in *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*, pp. 83–90, IEEE, 2016.

[8] B. A. Newman, A. Biswas, S. Ahuja, S. Girdhar, K. K. Kitani, and H. Admoni, "Examining the effects of anticipatory robot assistance on human decision making," in *International Conference on Social Robotics*, pp. 590–603, Springer, 2020.

[9] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 2018.

[10] R. M. Aronson, N. Almutlak, and H. Admoni, "Inferring goals with gaze during teleoperated manipulation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7307–7314, IEEE, 2021.

[11] R. Zhang, C. Walshe, Z. Liu, L. Guan, K. Muller, J. Whritner, L. Zhang, M. Hayhoe, and D. Ballard, "Atari-head: Atari human eye-tracking and demonstration dataset," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 6811–6820, 2020.

[12] R. Zhang, Z. Liu, L. Zhang, J. A. Whritner, K. S. Muller, M. M. Hayhoe, and D. H. Ballard, "Agil: Learning attention from human for visuomotor tasks," in *Proceedings of the european conference on computer vision (eccv)*, pp. 663–679, 2018.

[13] A. Saran, R. Zhang, E. S. Short, and S. Niekum, "Efficiently guiding imitation learning agents with human gaze," in *Adaptive Agents and Multi-Agent Systems*, 2020.

[14] Y. Xia, J. Kim, J. Canny, K. Zipser, T. Canas-Bajo, and D. Whitney, "Periphery-fovea multi-resolution driving model guided by human attention," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1767–1775, 2020.

[15] C. Liu, Y. Chen, L. Tai, H. Ye, M. Liu, and B. E. Shi, "A gaze model improves autonomous driving," in *Proceedings of the 11th ACM symposium on eye tracking research & applications*, pp. 1–5, 2019.

[16] Y. Chen, C. Liu, L. Tai, M. Liu, and B. E. Shi, "Gaze training by modulated dropout improves imitation learning," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7756–7761, IEEE, 2019.

[17] K. Renz, K. Chitta, O.-B. Mercea, A. S. Koepke, Z. Akata, and A. Geiger, "Plant: Explainable planning transformers via object-level representations," in *Conference on Robot Learning*, pp. 459–470, PMLR, 2023.

[18] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Conference on Robot Learning*, pp. 66–75, PMLR, 2020.

[19] "Learning-by-cheating carla challenge 2020 submission." Accessed Sept 20, 2022. `https://github.com/bradyz/2020_CARLA_challenge/`.

[20] "Causal confusion in learning-by-cheating github issue." Accessed Sept 20, 2022. `https://github.com/bradyz/2020_CARLA_challenge/issues/16`.

[21] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *Pattern Analysis and Machine Intelligence (PAMI)*, 2022.

[22] S. Greydanus, A. Koul, J. Dodge, and A. Fern, "Visualizing and understanding atari agents," in *International conference on machine learning*, pp. 1792–1801, PMLR, 2018.

[23] G. Silvera, A. Biswas, and H. Admoni, "Dreyevr: Democratizing virtual reality driving simulation for behavioural & interaction research," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, p. 639–643, IEEE Press, 2022.

[24] E. Tafaj, G. Kasneci, W. Rosenstiel, and M. Bogdan, "Bayesian online clustering of eye movement data," in *Proceedings of the symposium on eye tracking research and applications*, pp. 285–288, 2012.

[25] "Carla leaderboard." Accessed Sept 20, 2022. `https://github.com/carla-simulator/leaderboard`.

[26] C. Li, S. H. Chan, and Y.-T. Chen, "Who make drivers stop? towards driver-centric risk assessment: Risk object identification via causal inference," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10711–10718, IEEE, 2020.

[27] A. Biswas and H. Admoni, "Characterizing drivers' peripheral vision via the functional field of view for intelligent driving assistance," in *2023 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1–8, IEEE, 2023.