Gaze-Informed Vision Transformers: Predicting Driving Decisions Under Uncertainty

Sharath Koorathota*† Columbia University New York, USA sk4172@columbia.edu

Nikolas Papadopoulos* Columbia University New York, USA np2832@columbia.edu

Jia Li Ma Columbia University New York, USA jm5299@columbia.edu

Shruti Kumar Columbia University New York, USA sk5111@columbia.edu

Xiaoxiao Sun Columbia University New York, USA xs2362@columbia.edu

Arunesh Mittal Columbia University New York, USA arunesh.mitl@gmail.com

Patrick Adelman Georgia Institute of Technology Atlanta, USA adelmanp@gmail.com

Paul Sajda Columbia University New York, USA psajda@columbia.edu

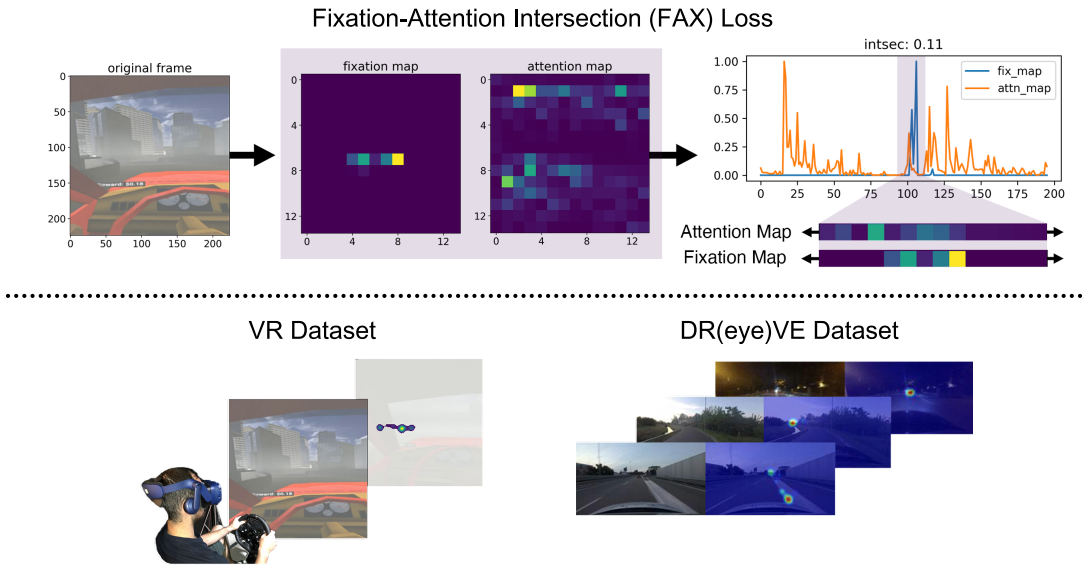

Figure 1: Our method integrates human gaze into Vision Transformers (ViT) to improve the prediction of driving decisions (left-right turns). We propose training ViT with a modified loss function, Fixation-Attention Intersection (FAX) loss, which calculates the intersection (dot product) of the model's attention map with the human fixation map. Experiments with both virtual reality and real-world datasets demonstrate that gaze integration enhances ViT accuracy in uncertain conditions (e.g., fog, bad weather/light conditions). When trained with FAX loss, the model's attention aligns with human gaze. Additionally, we release a novel dataset of human driving decisions collected in virtual reality to study turn behavior.

Abstract

Vision Transformers (ViT) have advanced computer vision, yet their efficacy in complex tasks like driving remains less explored. This study enhances ViT by integrating human eye gaze, captured via eye-tracking, to increase prediction accuracy in driving scenarios under uncertainty in both real-world and virtual reality scenarios.

*Both authors contributed equally to this research
†Corresponding author

ICMI Companion '24, November 04–08, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-0463-5/24/11
https://doi.org/10.1145/3686215.3688381

First, we establish the significance of human eye gaze in left-right driving decisions, as observed in both human subjects and a ViT model. By comparing the similarity between human fixation maps and ViT attention weights, we reveal the dynamics of overlap across individual heads and layers. This overlap demonstrates that fixation data can guide the model in distributing its attention weights more effectively. We introduce the fixation-attention intersection (FAX) loss, a novel loss function that significantly improves ViT performance under high uncertainty conditions. Our results show that ViT, when trained with FAX loss, aligns its attention with human gaze patterns. This gaze-informed approach has significant potential for driver behavior analysis, as well as broader applications in human-centered AI systems, extending ViT's use to complex visual environments.

CCS Concepts

• **Human-centered computing** → *Interaction design*; *Virtual reality*; *HCI theory, concepts and models*;

Keywords

eye-tracking, gaze, human-guided machine learning, vision transformers, driving decisions

ACM Reference Format:

Sharath Koorathota, Nikolas Papadopoulos, Jia Li Ma, Shruti Kumar, Xiaoxiao Sun, Arunesh Mittal, Patrick Adelman, and Paul Sajda. 2024. Gaze-Informed Vision Transformers: Predicting Driving Decisions Under Uncertainty. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI Companion '24), November 04–08, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3686215.3688381

1 Introduction

The performance of Vision Transformers (ViT) [9], has exceeded human performance across various visual tasks. ViT have exhibited state-of-the-art performance in tasks such as image recognition, action classification, and even autonomous driving [21]. The success of ViT has recently been attributed to their ability to process visual scenes like humans. This is particularly evident in their broader receptive fields compared to other model architectures and the distinct patterns of errors they exhibit [28]. Yet, utilizing ViT in real-world situations like driving poses challenges stemming from their limited interpretability and the absence of frameworks for direct human guidance.

We propose a novel approach [1] to tackle these challenges: incorporating eye-tracking data into ViT. Eye fixations offer a reliable measure of visual behavior and are often used to analyze human perception of intricate scenes [5]. Moreover, the attention mechanism intrinsic to ViT has been leveraged to study their interpretation of images and videos [4, 12]. We conduct experiments on two datasets involving human turn-taking decisions in virtual and real-world scenarios. Our primary objective is to uncover the relationship between human fixations and model attention to enhance the precision and reliability of decisions achieved by their combination.

First, we highlight differences in decision-making between humans and transformers in driving scenarios (left and right turning choices) under various types of uncertainty, expressed as opacity or contrast of the visual scene. Humans tend to mitigate localized uncertainty by fixating on fewer scene regions for extended durations. We then propose the fixation-attention intersection (FAX) loss, which calculates the intersection (dot product) of the model's attention map with the human fixation map. Experiments with both virtual reality and real-world datasets demonstrate that gaze integration enhances ViT accuracy in uncertain conditions (fog, bad weather/light conditions). When trained with FAX loss, the model's attention aligns with human gaze patterns, showing the potential for predicting human gaze. Additionally, we release a novel dataset of human driving decisions collected in virtual reality to study turn behavior.

[1]Code and data available at: github.com/schko/fixatt

2 Related work

The parallel between human and machine vision has attracted considerable interest. Recent approaches have emphasized transformers' self-attention attributes and receptive fields, which mirror the human visual system [28], are robust to occlusions and perturbations [27], generalize to multiple problems, and highly accurate compared to convolutional networks [11, 18]. Transformer-based architectures have shown high accuracy in predicting various eye tracking measures, such as types of eye behavior, gaze paths, or saliency maps [16]. Previous research has primarily explored retrospective comparisons between human and model attention using transformer-based networks [17, 26], or methods like knowledge distillation through teacher-student model designs. An alternative approach [25] in the context of Visual Question Answering (VQA) uses saliency-predicting models to guide model attention by integrating human-like attention across both image and text modalities. While these approaches provide valuable insights into aligning machine with human attention, they do not directly integrate human gaze into the model training process. Our proposed method combines the advantage of larger receptive fields of ViT with the ability of the human visual system to gather task-relevant information from complex scenes quickly [23] during model training.

Eye-tracking in humans can provide many insights into the behavioral and neural dynamics that underlie the flexible decision-making required in tasks such as driving. Current research in driving behavior has focused on integrating eye tracking to driver monitoring systems to index the driver's attention and alertness as given by fixation coordinates and pupil dilation [5, 20]. Recent gaze-driven driving research has also been focused on identifying when a driver is distracted or not paying attention to the road. These systems can detect inattention by analyzing gaze direction and duration and provide warnings or interventions to maintain safety [1, 8]. Eye-tracking can aid in predicting the driver's intentions, such as lane changes or turns [7]. By analyzing gaze-based indices, an automated system can anticipate the driver's maneuvers, adjust its behavior accordingly, and potentially use gaze patterns to ensure safety and efficiency.

Humans are notably successful in performing sensorimotor decisions under uncertainty compared to their artificially intelligent counterparts [3, 6, 15]. In tasks such as making a right turn onto a street, humans can infer and integrate information across spatial, temporal, and sensory modes for optimal and efficient decisions [10].

Recent advances in state-of-the-art robotics aim to integrate templates of the processes that underlie sensorimotor decision-making in humans to improve existing flexibility in decision-making [19, 22]. Prior studies have revealed that visual attention is a critical cognitive process in performing sensorimotor decision-making tasks in the information processing stage but not in the motor planning stage [8, 24]. Visual attention is often evaluated with eye tracking technologies as visual information available to a subject depends on the field of view and the position of the pupil [23].

3 Proposed Methods

3.1 Baseline Vision Transformer

Following the original ViT architecture [9], the representative frame $\mathbf{x} \in \mathbb{R}^{H\times W\times C}$ from the premotor period prior to motor action is divided into N non-overlapping patches of size $P \times P$, which are then flattened to form $\mathbf{x}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, where (H, W, C) are the dimensions of the input frame and $N = HW/P^2$ is the total number of patches. Given that a, $a = 1 \ldots A$, represents the number of attention heads and l, $l = 1 \ldots L$, the number of layers in the ViT model, we choose to implement a ViT model with $L = 12$ layers and $A = 12$ attention heads. The weights for each attention head across layers are given by:

$$\mathcal{A}_{(l,a)} = \text{softmax}\left(\frac{\mathbf{q}_{(l,a)}\mathbf{k}^T_{(l,a)}}{\sqrt{D_h}}\right) \tag{1}$$

where $\mathcal{A}_{(l,a)} \in \mathbb{R}^{(N+1)\times(N+1)}$ and $D_h = D/A$, where D is the embedding size. We convert the attention weight matrix $\mathcal{A}_{(l,a)}$ into a vector $\mathbf{a}_{(l,a)} \in \mathbb{R}^N$ by averaging over patches, while excluding the CLS token. The resulting vector $\mathbf{a}_{(l,a)}$ illustrates how the model, at head a and layer l, assigns attention to different image patches and is used to visualize attention maps.

In our specific application, which centers around predicting left or right turns based on the premotor period frame, we employ binary cross entropy loss ($\mathcal{L}_{BCE}$) as the loss function for the baseline ViT model.

$$\mathcal{L}_{BCE} = -c_1 \cdot \log(m_1) - c_2 \cdot \log(m_2) \tag{2}$$

where c_1, $c_2 \in \{0, 1\}$ denotes the two classes (left, right) and m_1, $m_2 \in [0, 1]$ represent the predicted probabilities for the left and right class respectively.

3.2 Fixation Maps

Fixation maps $\mathbf{f} \in \mathbb{R}^{H\times W}$ represent the aggregate eye gaze during the premotor period and match the size of input frames. We define flattened patches of fixation map $\mathbf{f}_p \in \mathbb{R}^{N\times(P^2\cdot C)}$, similar to the approach in the baseline ViT model. Additionally, we resize and flatten the original fixation map $\mathbf{f}$ to produce the vector $\mathbf{f}_{red} \in \mathbb{R}^N$, which has same size as the vector $\mathbf{a}_{(l,a)}$ of the ViT model and is used to compute the similarity between the attention and fixation maps.

3.3 Fixation-Attention Intersection (FAX) Loss

To better guide the baseline ViT model to simulate human attention, we introduce a novel fixation-attention intersection loss $\mathcal{L}_{FAX}$ to improve the model's ability to capture human-like attention patterns during training. This loss quantifies the average intersection ($\mathcal{I}$) as the dot product between ViT attention weights $\mathbf{a}_{(l,a)}$ of all heads across all layers and the reduced human fixation map $\mathbf{f}_{red}$. To ensure compatibility between $\mathcal{L}_{INT}$ and $\mathcal{L}_{BCE}$, we apply a sigmoid function to the average intersection $\mathcal{I}$. This prevents $\mathcal{I}$ from being too small, which could otherwise make $\mathcal{L}_{INT}$ disproportionately large and dominate $\mathcal{L}_{BCE}$. Restricting $\mathcal{I}$ to a range between 0.5 and 1 keeps both loss terms on a comparable scale, allowing their effective combination.

$$\mathcal{I} = \frac{\sum_{l}^{L}\sum_{a}^{A}\mathbf{a}_{(l,a)} \cdot \mathbf{f}_{red}}{L \cdot A} \tag{3}$$

$$\mathcal{L}_{INT} = \frac{1}{\text{sigmoid}(\mathcal{I})} \tag{4}$$

Finally, we define $\mathcal{L}_{FAX}$ by combining $\mathcal{L}_{INT}$ with the original classification loss $\mathcal{L}_{BCE}$:

$$\mathcal{L}_{FAX} = (1 - \lambda) \cdot \mathcal{L}_{BCE} + \lambda \cdot \mathcal{L}_{INT}, \tag{5}$$

where λ, $\lambda \in [0, 1]$, is the hyperparameter used for the weighted addition of the two losses. To determine the optimal value of λ for our experiments, we systematically evaluated a range of λ values, namely {0.01, 0.1, 0.2, 0.8, 1}.

3.4 Peripheral Masking of the Input

Peripheral masking involves the removal of regions outside the visual periphery within the frame. This is achieved by expanding the fixation area within fixation maps ($\mathbf{f}_p$) and zeroing all pixels outside of this area (Fig. 2). To study the importance of human-fixated regions, we transform input data using peripheral masking for both datasets and compare the performance with random rotation and translation of the mask in a "random masking" control.

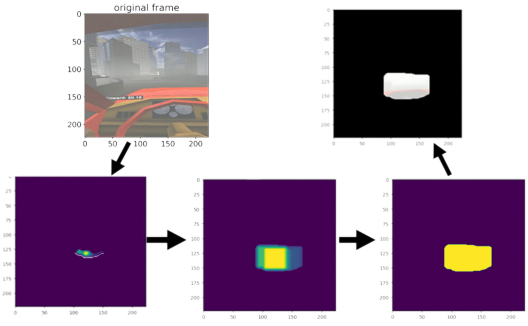

Figure 2: Peripheral masking of the input.

4 Datasets

4.1 VR Dataset

The VR dataset was collected as a part of a more extensive study on closed-loop brain-computer interface [14] (Fig. 7). 10 participants were recruited to complete a boundary avoidance task (BAT), presented by the HTC Vive Pro Eye VR headset, in a virtual city

Table 1: Overview of VR and DR(eye)VE datasets.

Dataset	Uncertainty	Train Set	Valid Set	Test Set	Left Turns(%)	Density	Contrast
VR	High	2015	356	599	48.9	0.65	-
VR	Low	2020	356	600	51.1	0.24	-
DR(eye)VE	High	236	55	73	51.0	-	0.11
DR(eye)VE	Low	236	55	73	46.7	-	0.39

environment with varying visual noise opacity. Using the Logitech G steering wheel, participants were instructed to drive a simulated car toward target locations. Steering wheel data and a video of the driving scene were recorded throughout the driving sessions. We identify motor actions through a simple peak and trough-detection technique on the steering wheel channel. Using a non-overlapping, look-behind window of 750ms, we assured that the peak we encountered was the true peak in steering wheel activity. The transformer models in our study were trained with frames corresponding to individual left-right turn motor actions. There are 6006 frames identified to be associated with a left or right turn, with 3293 left-turn frames. In addition to the video and steering information, eye-tracking data was collected using an HTC VIVE Pro Eye headset. Gaze coordinates from the eye tracker was used to construct the fixation map for each input frame. This map was computed by aggregating fixation data spanning a 3-second duration of the premotor period. We use the last frame in this premotor period as the input to all models.

4.2 DR(eye)VE Dataset

DR(eye)VE [2] is a publicly available driving dataset collected in real-world conditions across various landscapes, weather conditions, and times of day. The dataset contains gaze coordinates, driving speed, and course information for more than 500,000 frames. Geo-referenced locations are also available approximately every 25 frames. Because steering wheel data was unavailable, we used a combination of relative car positions, global positioning coordinates, and driving speed to identify left and right turn actions. To ensure the accuracy of our automated turn detection pipeline, at least two of the authors reviewed the videos manually and annotated frames corresponding to left or right turns. After review and validation of video frames, we identified 728 frames associated with a left or right turn, with 348 left-turn frames. Eye tracking data was collected using an SMI ETG 2w sensor. Fixation maps were computed using the same method for the VR dataset with a premotor period of 1 second. We use the first frame in this premotor period as the input to all models, driven by the relatively narrow field of view of the DR(eye)VE scene camera compared to the VR scene camera.

4.3 Uncertainty in Visual Scene

In our study, uncertainty refers to the visual conditions within a scene that hinder a driver's ability to clearly recognize objects. These conditions include various factors such as fog, weather, and lighting, which differ between the VR and real-world driving datasets. Although the specific definition of uncertainty varies between these datasets, the common factor is that high uncertainty corresponds to scenes where the driver has difficulty to perceive objects clearly, whether due to dense fog in the VR environment or adverse weather and lighting conditions in the real-world dataset.

In the VR dataset, we manipulated visual uncertainty on a trial-by-trial basis by adjusting the visual noise opacity parameter, as described in [14]. This adjustment simulates the type of white, 1/f noise typically found in visual search tasks, and participants perceived it through the varying density of fog in the virtual city environment. High noise opacity represented higher uncertainty, making it more difficult for participants to distinguish objects and navigate the scene.

In the DR(eye)VE dataset, which represents real-world driving scenarios, uncertainty was quantified by computing the average contrast across the entire image. This contrast was calculated by averaging the local pixel contrast, determined from the minimum and maximum luminance within a 5×5 kernel around each pixel. Lower average contrast indicates higher uncertainty, as it reflects poorer visibility conditions, such as bad weather or low-light situations, where distinguishing objects becomes more challenging.

We summarize the two datasets in Table 4. We also provide a video of an example turn from both datasets (see supplementary material).

5 Results

5.1 Comparing Human and Model Attention Under Uncertainty

To establish the importance of integrating eye gaze into ViT, we analyze the complementary nature of human and model attention under uncertainty. This is demonstrated through a comparison of human fixation maps (Fig. 3A) with the attention weights from the 12-layer baseline ViT model (Fig. 3B, C) in the VR dataset, selected for its larger sample size.

The distribution of fixations in both datasets (gray) indicates that viewing time is concentrated around the center of the frames, although relatively sparse, with a larger fixation area for DR(eye)VE than the VR dataset. This difference is likely because participants were goal-directed in the VR dataset, focusing on avoiding the boundary and navigating through the environment with different levels of visibility. In comparison, in the DR(eye)VE dataset, participants needed to remain vigilant about other vehicles, attend to road signs, and ensure a safe driving experience by exploring the scene for unexpected events. The high uncertainty conditions across both datasets result in longer relative fixation durations in fewer regions. This finding is consistent with existing literature [23] and suggests that humans minimize local spatial uncertainty in low-visibility

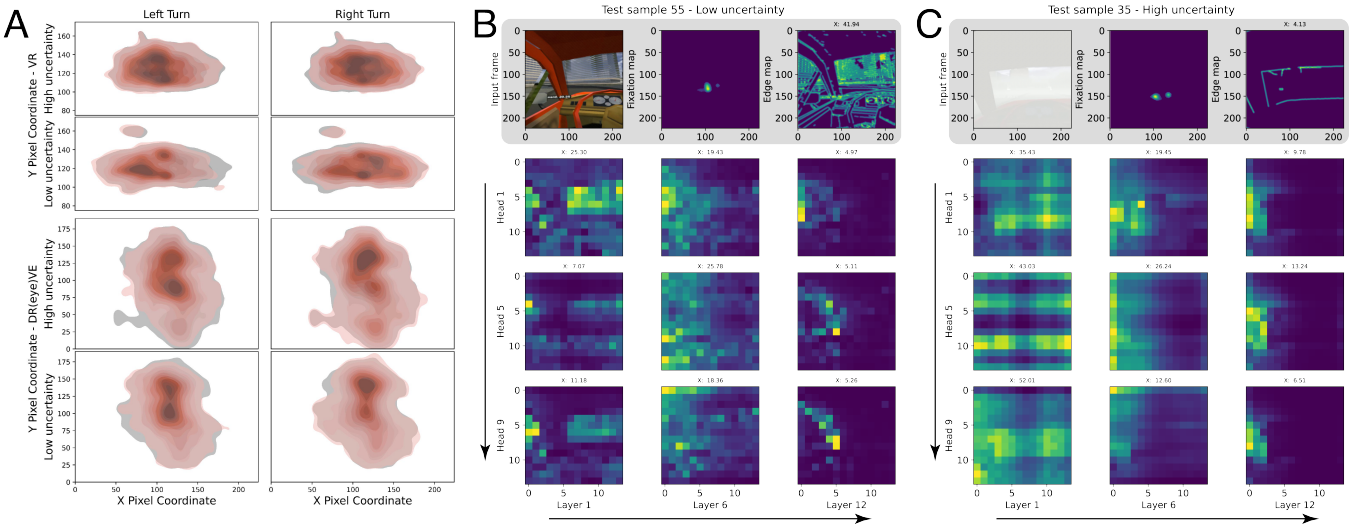

Figure 3: (A) KDE plot illustrating the distribution of fixations across pixel coordinates (x and y) across all test sample frames in the VR and DR(eye)VE datasets. Fixations are extracted from and aggregated over the premotor period prior to motor decisions. Higher density distribution indicates higher fixation duration. Class-specific (left or right) distributions are denoted in red; the overall distribution is gray. (B and C) Qualitative ViT results from two test samples corresponding to low (B) and high (C) uncertainty conditions in the VR dataset. X = dot product similarity between fixation and respective activation map. Only weights from 3 heads across 3 layers, corresponding to the first, middle, and last layers, respectively, are shown.

scenarios through longer fixation time in fewer areas rather than minimize global uncertainty through an exploration strategy.

During the decision-making process concerning left or right turns, the attention weights (as depicted in Fig. 3B, C) exhibit broader scene coverage than human fixation maps. This phenomenon is characterized by increased attention across the entire frame, particularly in high-uncertainty scenarios. Furthermore, the overall attention across ViT layers differs with depth. In shallower layers, attention is dispersed, primarily capturing edge-related scene details, while in deeper layers, attention is more concentrated, integrating contextual information. Our results suggest that transformer models may seek to minimize global over local uncertainty, the opposite of human strategy. Thus, human gaze may provide the model with information on which regions of the frame may be more relevant for resolving uncertainty, allowing more accurate learning.

5.2 Layer Pruning in Vision Transformers Based on Similarity to Human Attention

To understand how model attention overlaps with edges and fixations from the scene, we measure overall attention and compute the similarity between fixation maps and layer-specific attention maps using the dot product described in Eq. 3. First, the total ViT attention is not notably different between low and high uncertainty test samples. In contrast, the total number of fixations and edges varies by uncertainty condition (Fig. 4A, B). In other words, high visibility (i.e., low uncertainty) results in a larger fixation area and more detected edges. Still, the model does not employ a different strategy in its attention overall. This suggests that the model may find fixation data beneficial for parsing uncertainty through better distributing its attention weights spatially in the scene.

Following the observation that total model attention does not differ by visual uncertainty, we explored layer-specific attention and its overlap with human fixation maps, as shown in Figures 4C, D. Observing a decline in this overlap beyond the fifth layer, we pruned the ViT model to its first five layers, creating a 5-layer model (5-ViT). We also tested a single-layer ViT (1-ViT), as a control.

For both DR(eye)VE and VR dataset, the accuracy of the 5-ViT model is not significantly different than 12-ViT (Table 3), suggesting that reducing the model's complexity by half does not impact accuracy. However, reducing the model to a single layer (1-ViT) significantly lowers performance compared to 12-ViT across both datasets. This outcome demonstrates that models pruned to retain layers with the highest alignment to human attention can maintain or even enhance performance, especially in scenarios with smaller datasets like DR(eye)VE, where the 5-ViT model shows an improvement, though not statistically significant, over the 12-ViT model.

5.3 Assessing the Impact of Human Eye Gaze on Task Performance

Here, we investigate the contribution of human eye gaze information to the task of predicting left or right turns. In particular, we evaluate whether gaze data during the premotor period suffices for accurately predicting a left or right turn. Our experimental setup includes a baseline model, a 12-layer Vision Transformer (12-ViT), processing the entire image frame. This model's performance was benchmarked against two approaches: peripheral masking (Section 3.4) and a dummy classifier that predicts the turn direction based on the assumption that drivers' gaze towards the right or left

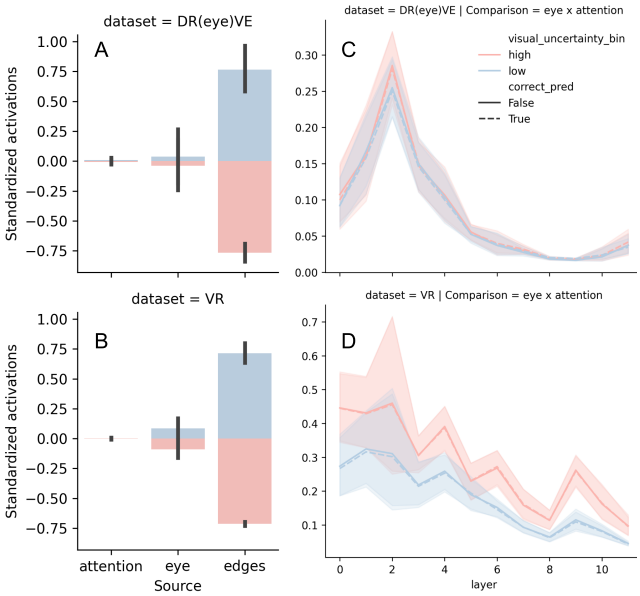

Figure 4: (A, B) Total, standardized sum of activations, by uncertainty split, for both datasets. We define total activation in the baseline ViT as the sum of attention weights across layers and heads. Total fixation refers to the pixel-wise sum of fixation maps, a measure of the overall fixation area. Total edge activation refers to the pixel-wise sum of edge maps. (C, D) The similarity between attention weights across layers and fixation maps, using Eq. 3. Results are aggregated from all test samples on the best-performing, 12-layer baseline ViT. Line color shows the uncertainty split of the test samples, while line style shows whether ViT classified the motor action correctly. Error band shows the 95% CI.

side of the frame indicates the corresponding turn direction (see Appendix).

Considering the task's reliance on spatial scene information, we evaluated the efficacy of a dummy classifier in leveraging eye gaze data for turn prediction. The performance of this classifier established a benchmark for evaluating our proposed methods, achieving 50.48 ± 4.19% accuracy in the DR(eye)VE dataset and 64.67 ± 1.33% in the VR dataset (Table 2). These results indicate that gaze information alone can accurately predict driving decisions in the VR dataset for a significant proportion of cases. The effectiveness in the VR context is attributed to its streamlined design, featuring roads without complex navigational choices or directional changes, where drivers' turn directions align with their gaze. Conversely, the real-world driving environment presents complex scenarios with external factors, such as pedestrians and other vehicles, thereby complicating the direct correlation between gaze direction and turning decisions observed in the VR environment.

We also used peripheral masking to assess whether focusing solely on the driver's fixation area within the scene yields enough information for accurate task completion. Across both datasets, our analysis revealed no significant difference in total accuracy between the full-frame baseline model (12-ViT) and the model trained on peripherally masked frames (Peripheral-ViT), accounting for both high and low uncertainty conditions (Table 3). These results suggest that while the fixation area is critical for classification, the surrounding scene holds valuable information as well. Consequently, we opted for the FAX loss approach, which rewards the model for focusing more on the fixation area without neglecting the surrounding scene information.

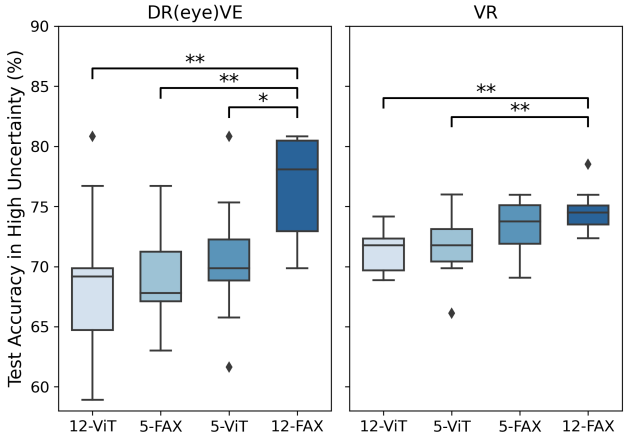

Figure 5: Boxplots displaying the test accuracy in high uncertainty of the top performing models on the DR(eye)VE and VR datasets. 12-ViT and 5-ViT denote Vision Transformer models with 12 and 5 layers, respectively; 12-FAX and 5-FAX represent equivalent ViT models trained with the FAX loss, with the optimal λ value in each case. The Mann-Whitney U test assesses statistical significance (* p < 0.05, ** p < 0.01, *** p < 0.001).

Table 2: Top performing models based on test accuracy (mean ± std). Models are compared against the baseline accuracy of the dummy classifier.

	Model	High Uncertainty	Low Uncertainty	Total
DR(eye)VE	*Dummy*	49.73 ± 8.12	51.23 ± 4.80	50.48 ± 4.19
	12-ViT	68.77 ± 6.42	64.11 ± 7.36	66.44 ± 6.19
	5-FAX	69.18 ± 4.44	70.82 ± 4.96	70.00 ± 4.14
	5-ViT	70.55 ± 5.18	69.45 ± 6.68	70.00 ± 5.35
	12-FAX	76.58 ± 4.26	71.23 ± 4.03	73.90 ± 3.39
VR	*Dummy*	71.15 ± 1.66	58.38 ± 2.03	64.67 ± 1.33
	12-ViT	69.75 ± 6.50	60.47 ± 4.67	65.13 ± 5.34
	5-ViT	71.68 ± 2.67	61.89 ± 2.47	66.71 ± 1.71
	5-FAX	73.38 ± 2.25	61.24 ± 1.70	67.31 ± 1.22
	12-FAX	74.66 ± 1.72	61.25 ± 1.97	67.88 ± 1.31

5.4 Gaze Integration Improves ViT Accuracy in Uncertain Conditions

Next, we validate our findings from 5.1 that human eye gaze can resolve uncertainty. To do so, we trained both the full-layer and

Table 3: Pairwise Mann-Whitney U Test Comparisons.

Model 1	Model 2	DR(eye)VE				VR			
		Total		High Uncertainty		Total		High Uncertainty	
		p-value	reject	p-value	reject	p-value	reject	p-value	reject
12-FAX	12-ViT	0.005	True	0.007	True	0.064	False	0.002	True
12-FAX	5-ViT	0.110	False	0.016	True	0.161	False	0.007	True
12-FAX	5-FAX	0.044	True	0.003	True	0.405	False	0.326	False
5-ViT	5-FAX	0.970	False	0.518	False	0.677	False	0.173	False
12-ViT	5-ViT	0.161	False	0.378	False	0.570	False	0.850	False
12-ViT	1-ViT	0.008	True	0.002	True	0.003	True	0.003	True
12-ViT	Periph.-ViT	0.000	True	0.000	True	0.006	True	0.140	False
Dummy	Periph.-ViT	0.677	False	0.621	False	0.058	False	0.096	False

the ablated ViT models using the FAX loss to drive the model's attention towards areas aligned with human gaze. For both datasets, the top performing models include: 12-FAX, 5-FAX, 5-ViT, 12-ViT (Fig. 5). In the DR(eye)VE dataset, the 12-FAX model demonstrated a significant accuracy improvement over the 12-ViT model, with a mean increase of 7.46%. Furthermore, the 12-FAX model exhibited a distinct, though not statistically significant, accuracy improvement over the 5-ViT model and a significant improvement over the 5-FAX model. Notably, in high uncertainty conditions, the 12-FAX model consistently outperformed the 12-ViT and 5-ViT models, underscoring the benefits of gaze integration. These results are consistent with our previous observations that human eye gaze contributes to resolving uncertainty and enhances ViT performance, especially in high uncertainty conditions.

In the VR dataset, performance differences among models were not statistically significant in low uncertainty conditions. However, in high uncertainty scenes, the 12-FAX model significantly outperformed both the 12-ViT and 5-ViT models, further highlighting the importance of human gaze in resolving uncertainty. Additionally, there is no significant difference in accuracy between 5-FAX and 5-ViT across both datasets and all levels of uncertainty (Table 3), suggesting that shallower networks do not benefit as significantly as the full-layer model from the integration of human gaze data. Overall, our results support the conclusion from 5.1 that human gaze enhances the model's attention, thereby improving ViT model performance in uncertain conditions.

5.5 FAX Loss Aligns Model Attention with Human Gaze

Lastly, we present qualitative results displaying the average attention map across all heads for each layer of the ViT model with varying λ values, from $\lambda = 0$ (where only $\mathcal{L}_{BCE}$ is used) to $\lambda = 1$ (where only $\mathcal{L}_{INT}$ is used), as illustrated in Fig. 6. An increase in λ (from left to right in the figure) correlates with attention maps increasingly resembling human fixation patterns, except at $\lambda = 1$, where the model disperses attention across the frame to mimic human fixations. In the earlier layers, attention is distributed globally across the image, while subsequent layers progressively shift to more precise attention regions correlating with human fixations, especially at optimal λ settings, such as 0.2 or 0.8. Notably, at $\lambda = 0.8$, which aligns with the highest accuracy in the DR(eye)VE dataset, there is the highest alignment between attention maps and human fixation data. To quantitatively support these observations, we calculated the Intersection over Union (IoU) metric between attention and human fixation maps for all λ values (Fig. 6 D). The IoU was calculated by converting heatmaps to binary masks using a threshold of 0.4, computing the intersection and union of these masks, and averaging the IoU across all heads and layers for all test set samples of the two datasets.

The results demonstrate that ViT, when trained with FAX loss at an optimal λ value, can adopt an attention strategy similar to humans. By integrating gaze, ViT learn to focus on specific regions, reducing local uncertainty without compromising their broader receptive field. This significantly enhances model performance in driving contexts and could also apply to other areas where transformers are employed for image and video tasks.

6 Limitations and Future Work

While our study demonstrates the potential of integrating human gaze into Vision Transformers (ViT) for driving decision-making under uncertainty, several limitations warrant further investigation. Currently, our approach predicts turning decisions based on single frames. However, driving is inherently a dynamic task that requires processing sequences of frames to understand the broader context. Expanding our model to handle dynamic video data could enhance its applicability and performance in real-world scenarios. Future research should explore incorporating temporal information to better align the model with the sequential nature of human decision-making.

Moreover, while this work focuses on driving, the integration of human gaze has broader implications. Future studies should explore applying this approach in other domains where gaze data is valuable, such as medical imaging. For instance, training models on datasets that include radiologists' gaze data could enhance diagnostic accuracy and decision-making in clinical contexts [13].

7 Conclusion

Our paper establishes the critical role of human eye gaze in enhancing Vision Transformer (ViT) models for driving decision-making under uncertainty. A systematic comparison between human and

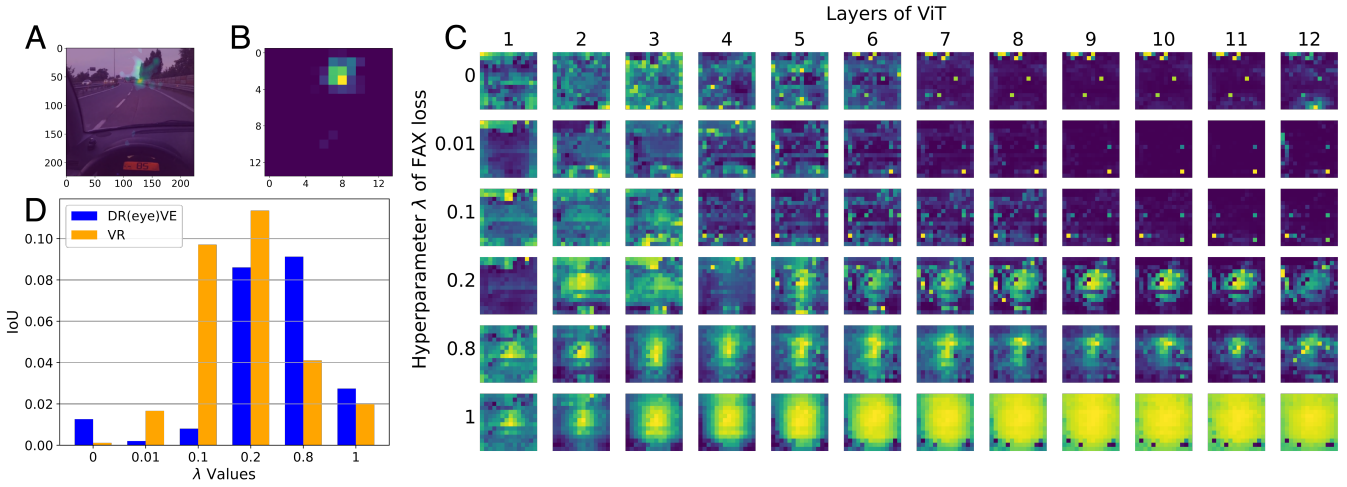

Figure 6: Training with FAX loss aligns model attention with human gaze. This figure shows the impact of FAX loss on the ViT model's attention maps for test set samples. (A) Input frame with overlaid human fixation map. (B) Human fixation data aligned to ViT attention map dimensions. (C) Average attention maps across all heads for each ViT layer, for distinct λ values in FAX loss (Eq. 5), showing increasing resemblance to human fixation patterns with higher λ values. (D) Intersection over Union (IoU) metric between attention and fixation maps for the test set samples for all λ values, quantifying alignment. These results demonstrate that optimal λ values in FAX loss (e.g., $\lambda = 0.2, 0.8$ for DR(eye)VE and $\lambda = 0.1, 0.2$ for VR) lead to attention maps better resembling human fixation area, indicating the model's ability to predict human gaze in driving scenarios.

model attention characteristics provided the groundwork for this integration. We introduced the fixation-attention intersection (FAX) loss, a novel methodology for incorporating gaze data into ViT. Notably, the application of FAX loss led to a significant performance improvement in ViT, particularly under high uncertainty conditions. These advancements demonstrate the potential of human-guided transformer models to refine driver behavior analysis and improve human-vehicle interaction, with broader implications for human-centered AI systems and human-computer interaction.

Acknowledgments

The authors would like to thank Pawan Lapborisuth, Josef Faller, and Ziheng Li for their valuable contributions and support during the course of this study. The study was funded by a Vannevar Bush Faculty Fellowship from the U.S. Department of Defense (N00014-20-1-2027), a Cooperative Agreement with the Army Research Laboratory (W911NF-23-2-0067), and a Center of Excellence grant from the Air Force Office of Scientific Research (FA9550-22-1-0337).

References

[1] Christer Ahlstrom, Katja Kircher, and Albert Kircher. 2013. A Gaze-Based Driver Distraction Warning System and Its Effect on Visual Behavior. *IEEE Transactions on Intelligent Transportation Systems* 14, 2 (June 2013), 965–973. https://doi.org/10.1109/TITS.2013.2247759

[2] 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 *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, Las Vegas, NV, USA, 54–60. https://doi.org/10.1109/CVPRW.2016.14

[3] I Barabás, A Todoruţ, N Cordoş, and A Molea. 2017. Current challenges in autonomous driving. *IOP Conference Series: Materials Science and Engineering* 252 (Oct. 2017), 012096. https://doi.org/10.1088/1757-899X/252/1/012096

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? arXiv:2102.05095 [cs]

[5] James F. Cavanagh, Thomas V. Wiecki, Angad Kochar, and Michael J. Frank. 2014. Eye Tracking and Pupillometry Are Indicators of Dissociable Latent Decision Processes. *Journal of Experimental Psychology: General* 143, 4 (2014), 1476–1488. https://doi.org/10.1037/a0035813

[6] José Fernando Sabando Cárdenas, Jong Gyu Shin, and Sang Ho Kim. 2020. A Few Critical Human Factors for Developing Sustainable Autonomous Driving Technology. *Sustainability* 12, 7 (April 2020), 3030. https://doi.org/10.3390/su12073030

[7] Jan Willem De Gee, Tomas Knapen, and Tobias H. Donner. 2014. Decision-Related Pupil Dilation Reflects Upcoming Choice and Individual Bias. *Proceedings of the National Academy of Sciences* 111, 5 (Feb. 2014). https://doi.org/10.1073/pnas.1317557111

[8] T. D'Orazio, M. Leo, C. Guaragnella, and A. Distante. 2007. A Visual Approach for Driver Inattention Detection. *Pattern Recognition* 40, 8 (Aug. 2007), 2341–2355. https://doi.org/10.1016/j.patcog.2007.01.018

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2020). https://doi.org/10.48550/ARXIV.2010.11929

[10] Jason P. Gallivan, Craig S. Chapman, Daniel M. Wolpert, and J. Randall Flanagan. 2018. Decision-making in sensorimotor control. *Nature Reviews Neuroscience* 19, 9 (Sept. 2018), 519–534. https://doi.org/10.1038/s41583-018-0045-9

[11] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2022. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. http://arxiv.org/abs/1811.12231 arXiv:1811.12231 [cs, q-bio, stat].

[12] Jacob Gildenblat. 2020. Exploring Explainability for Vision Transformers. http://jacobgil.github.io/deeplearning/vision-transformer-explainability.

[13] Bulat Ibragimov and Claudia Mello-Thoms. 2024. The Use of Machine Learning in Eye Tracking Studies in Medical Imaging: A Review. *IEEE Journal of Biomedical and Health Informatics* 28, 6 (2024), 3597–3612. https://doi.org/10.1109/JBHI.2024.3371893

[14] Sharath Koorathota, Jia Li Ma, Josef Faller, Linbi Hong, Pawan Lapborisuth, and Paul Sajda. 2023. Pupil-linked arousal correlates with neural activity prior to sensorimotor decisions. *Journal of Neural Engineering* 20, 6 (dec 2023), 066031. https://doi.org/10.1088/1741-2552/ad1055

[15] John W Krakauer and Pietro Mazzoni. 2011. Human sensorimotor learning: adaptation, skill, and beyond. *Current Opinion in Neurobiology* 21, 4 (Aug. 2011), 636–644. https://doi.org/10.1016/j.conb.2011.06.012

[16] Jianxun Lou, Hanhe Lin, David Marshall, Dietmar Saupe, and Hantao Liu. 2022. TranSalNet: Towards perceptually relevant visual saliency prediction. *Neurocomputing* 494 (July 2022), 455–467. https://doi.org/10.1016/j.neucom.2022.04.080 arXiv:2110.03593 [cs].

[17] Paria Mehrani and John K. Tsotsos. 2023. Self-attention in vision transformers performs perceptual grouping, not attention. *Frontiers in Computer Science* 5 (June 2023). https://doi.org/10.3389/fcomp.2023.1178450

[18] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Intriguing Properties of Vision Transformers. http://arxiv.org/abs/2105.10497 arXiv:2105.10497 [cs].

[19] Kuniaki Noda, Hiroaki Arie, Yuki Suga, and Tetsuya Ogata. 2014. Multimodal Integration Learning of Robot Behavior Using Deep Neural Networks. *Robotics and Autonomous Systems* 62, 6 (June 2014), 721–736. https://doi.org/10.1016/j.robot.2014.03.003

[20] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. 2018. Predicting the Driver's Focus of Attention: the DR(eye)VE Project. http://arxiv.org/abs/1705.03854 arXiv:1705.03854 [cs].

[21] Sayak Paul and Pin-Yu Chen. 2022. Vision Transformers Are Robust Learners. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 2 (June 2022), 2071–2081. https://doi.org/10.1609/aaai.v36i2.20103

[22] Luka Peternel, Nikos Tsagarakis, and Arash Ajoudani. 2017. A Human–Robot Co-Manipulation Approach Based on Human Sensorimotor Information. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25, 7 (July 2017), 811–822. https://doi.org/10.1109/TNSRE.2017.2694553

[23] Laura Walker Renninger, Preeti Verghese, and James Coughlan. 2007. Where to Look next? Eye Movements Reduce Local Uncertainty. *Journal of Vision* 7, 3 (Feb. 2007), 6. https://doi.org/10.1167/7.3.6

[24] Jeffrey D Schall and Narcisse P Bichot. 1998. Neural Correlates of Visual and Motor Decision Processes. *Current Opinion in Neurobiology* 8, 2 (April 1998), 211–217. https://doi.org/10.1016/S0959-4388(98)80142-6

[25] E. Sood, F. Kogel, P. Muller, D. Thomas, M. Bace, and A. Bulling. 2023. Multimodal Integration of Human-Like Attention in Visual Question Answering. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE Computer Society, Los Alamitos, CA, USA, 2648–2658. https://doi.org/10.1109/CVPRW59228.2023.00265

[26] Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020. Interpreting Attention Models with Human Visual Attention in Machine Reading Comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, Raquel Fernández and Tal Linzen (Eds.). Association for Computational Linguistics, Online, 12–25. https://doi.org/10.18653/v1/2020.conll-1.2

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve Jegou. 2021. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 10347–10357. https://proceedings.mlr.press/v139/touvron21a.html

[28] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. 2021. Are Convolutional Neural Networks or Transformers more like human vision? (2021). arXiv:2105.07197 [cs.CV]

A Appendix

A.1 Comparative Performance of Models

Comparative performance of all Vision Transformer (ViT) model variants on the VR dataset and on the DR(eye)VE dataset under both low and high uncertainty combined (Table 5). Metrics include test accuracy (%), Area Under the Receiver Operating Characteristic Curve (AUC), and F1-Score. Results are presented as mean ± standard deviation computed from 10 independent training runs with distinct data splits. Model variants include 1-, 5-, and 12-layer ViT (denoted as 1-ViT, 5-ViT and 12-ViT) trained on image data using binary cross entropy (BCE) loss, and equivalent models trained with FAX loss (denoted as 1-FAX, 5-FAX and 12-FAX). Peripheral-ViT and Random Peripheral-ViT denote the baseline 12-layer ViT trained with ablated input, as outlined in the 'Proposed Methods' section. For the models trained with FAX loss, eye gaze was not used during inference. Instead, we loaded the trained weights to the vanilla ViT architecture and proceeded with predictions on the test data.

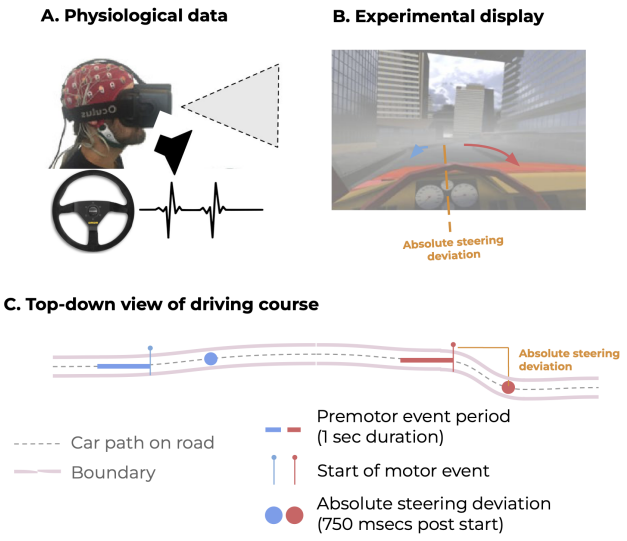

Figure 7: Adapted from [14]. Overview of how motor events were detected and premotor period is defined for the VR dataset. (A) We simultaneously collected neural data from EEG, autonomic measures using ECG, eye movements and pupil dynamics using a VR-headset embedded eye tracking system, and motor actions using a steering wheel. (B) Participants (n=10) performed 3 virtual reality driving task sessions, requiring boundary avoidance under time pressure and changing visual uncertainty. Their motor actions were recorded from the steering deviation as they were navigating a city environment. We analyzed direction-independent (i.e., absolute) steering deviation). Motor actions belong to a global trial with a set level of visual fog (opacity) in the environment that participants drove in. (C) The start of each motor action was marked using a peak detection method on the steering wheel data since this was most relevant to navigating the boundary avoidance task. The premotor periods of interest for this study were a fixed, 3-second interval before each event, and the intensity of the motor activity was determined by the post-event steer angle. Blue and red circles indicate events with low and high motor intensity, respectively.

A.2 Fixation Map and Edge Detection

The fixation maps for both VR and DR(eye)VE datasets are generated by aggregating fixation data spanning a different duration of the premotor period. The fixation map for DR(eye)VE datasets was generated by utilizing the fixation coordinate for 10 frames, which is 0.4 seconds, whereas the VR dataset uses fixation data for 3 seconds. The reason for such a difference was, the fixation map for DR(eye)VE tends to have a larger fixation area due to the number of uncertain factors to consider. A shorter duration in this case generates a fixation map with a reasonable size.

The fixation map is generated based on the dimension of the original frame. The gaze coordinates for the VR, in particular, need

Table 4: Configuration of top performing models.

Model	Architecture	Layers	Loss	λ	
				VR	DR(eye)VE
12-ViT	ViT	12	CrossEntr.	-	-
5-ViT	ViT	5	CrossEntr.	-	-
12-FAX	ViT	12	FAX	0.1	0.8
5-FAX	ViT	5	FAX	0.2	0.2

to be re-scaled to match the frame dimension before the fixation map gets generated. The generation of the fixation map starts with computing a 2D Gaussian matrix with pre-defined variables. These variables are being used to further re-scale the gaze coordinate so that the coordinate aligns with the Gaussian matrix. The fixation map is generated by adding up the portion of the Gaussian matrix that corresponds to those gaze coordinates. There is a duration parameter available that can be used to adjust the weight of the gaze coordinate. The fixation map for the DR(eye)VE dataset is generated in the same fashion.

For edge detection, the original frame first goes through color conversion from color to grayscale. Afterward, Gaussian smoothing is performed on the grayscale image with a 3x3 kernel. The purpose of this step is to improve the edge detection result. This blurred image finally feeds into the Canny edge detector with a lower threshold of 25 and an upper threshold of 50. The specific implementation details are provided in the `process_driving_video.py`.

A.3 Peripheral Masking

In our processing pipeline, we enhance the fixation heatmap to encompass a broader peripheral area, which we subsequently employ for generating masks. This augmentation of the heatmap accounts for the fact that peripheral regions can provide valuable insights into what is visually perceived around the central foveated point, as captured by eye tracking systems.

To achieve this, we perform a dilation operation on the fixation heatmap. Dilation involves expanding regions of high intensity, effectively enlarging the areas of interest. The dimensions of the kernel used for dilation are set empirically to 30x30 pixels within the context of a 224×224 input frame. This choice aligns with the approximate 30 degrees of visual angle that humans perceive in the mid-peripheral region. The specific implementation details are provided in the `create_peripheral_mask.py`.

A.4 Dummy Classifier

The dummy classifier is based on the simple assumption that a driver's gaze direction within an image frame -either right or left- indicates the intended turning direction. This classifier splits each image into two equal halves (left-right) and calculates the sum of pixel values for each side. A prediction for a right or left turn is then made based on which half of the image has a greater sum of pixel values, suggesting that a higher sum indicates the driver's focus area and, by extension, the turn direction. This approach establishes a baseline for comparison, illustrating the extent to which the spatial information inferred from gaze can address the task. Consequently, it allows for an evaluation of the enhancements brought by integrating gaze data into ViT models. The specific implementation details are provided in the `dummy_classifier.py`.

A.5 Implementation Details

The experiments were conducted on a Lambda Labs Vector Machine, equipped with Threadripper Pro 3990X v4 @ 4.3GHz (64 cores), 128 GB DDR4 RAM and 2x NVIDIA GeForce RTX 3090 (24 GB VRAM each). The implementation was carried out in PyTorch, using pretrained weights from ImageNet. The dataset was split into training, validation, and test sets with a ratio of 65:15:20 for the virtual reality (VR) and 68:12:20 for the real-world (DR(eye)VE) datasets to accommodate differences in dataset sizes. We report the mean and standard deviation of the test accuracy resulting from 10 distinct splits of our datasets for each type of run. The training was performed using the SGD optimizer with an initial learning rate of 0.001. The learning rate was adjusted using a scheduler to ensure convergence. Models were trained for a maximum of 100 epochs, with early stopping based on the validation loss (20 epochs) to prevent overfitting. For all models, gaze data was not used during inference. Trained weights were instead imported into the vanilla ViT architecture for predictions on the test dataset.

Table 5: Performance metrics of ViT variants on the VR and DR(eye)VE datasets under both low and high uncertainty combined, showing test accuracy (%), AUC, and F1-Score as mean ± standard deviation from 10 training runs with distinct data splits.

		VR Dataset			DR(eye)VE Dataset		
Model	λ	Accuracy (%)	AUC	F1	Accuracy (%)	AUC	F1
Dummy	-	64.67 ± 1.33	0.65 ± 0.01	0.62 ± 0.02	50.48 ± 4.19	0.49 ± 0.04	0.60 ± 0.05
1-ViT	-	61.74 ± 1.55	0.62 ± 0.02	0.65 ± 0.02	58.22 ± 4.21	0.57 ± 0.05	0.62 ± 0.08
5-ViT	-	66.71 ± 1.62	0.67 ± 0.01	0.65 ± 0.04	70.00 ± 5.08	0.70 ± 0.05	0.72 ± 0.07
12-ViT	-	65.13 ± 5.07	0.65 ± 0.05	0.61 ± 0.20	66.44 ± 5.87	0.66 ± 0.05	0.69 ± 0.08
12-FAX	0.01	66.17 ± 1.09	0.66 ± 0.01	0.67 ± 0.03	63.22 ± 6.03	0.63 ± 0.06	0.65 ± 0.10
»	0.1	67.88 ± 1.24	0.68 ± 0.01	0.70 ± 0.02	67.47 ± 4.84	0.66 ± 0.06	0.72 ± 0.05
»	0.2	67.36 ± 2.00	0.67 ± 0.02	0.68 ± 0.04	64.66 ± 8.00	0.64 ± 0.08	0.67 ± 0.15
»	0.8	65.80 ± 1.98	0.66 ± 0.02	0.68 ± 0.02	73.90 ± 3.22	0.74 ± 0.03	0.76 ± 0.05
»	1	49.73 ± 1.53	0.50 ± 0.00	0.13 ± 0.27	46.44 ± 3.04	0.50 ± 0.00	0.26 ± 0.32
5-FAX	0.01	66.11 ± 1.12	0.66 ± 0.01	0.66 ± 0.02	70.27 ± 4.04	0.70 ± 0.03	0.72 ± 0.07
»	0.1	66.53 ± 1.58	0.67 ± 0.02	0.67 ± 0.02	69.04 ± 5.27	0.69 ± 0.05	0.69 ± 0.12
»	0.2	67.31 ± 1.16	0.67 ± 0.01	0.67 ± 0.03	70.00 ± 3.93	0.70 ± 0.03	0.72 ± 0.06
»	0.8	63.83 ± 2.56	0.64 ± 0.03	0.63 ± 0.08	66.51 ± 3.99	0.66 ± 0.04	0.68 ± 0.06
»	1	49.83 ± 0.42	0.50 ± 0.00	0.25 ± 0.31	48.08 ± 4.53	0.50 ± 0.01	0.27 ± 0.33
1-FAX	0.01	58.89 ± 4.55	0.59 ± 0.05	0.58 ± 0.15	62.05 ± 4.56	0.61 ± 0.04	0.67 ± 0.06
»	0.1	51.63 ± 1.51	0.51 ± 0.02	0.54 ± 0.26	61.92 ± 4.41	0.61 ± 0.04	0.66 ± 0.07
»	0.2	50.76 ± 0.81	0.51 ± 0.01	0.53 ± 0.26	61.16 ± 5.95	0.61 ± 0.04	0.60 ± 0.13
»	0.8	50.42 ± 0.62	0.50 ± 0.00	0.53 ± 0.27	51.64 ± 4.18	0.51 ± 0.02	0.38 ± 0.28
»	1	49.88 ± 0.29	0.50 ± 0.00	0.33 ± 0.33	46.85 ± 3.61	0.50 ± 0.01	0.17 ± 0.26
Peripheral-ViT	-	63.38 ± 1.30	0.63 ± 0.01	0.63 ± 0.05	51.23 ± 4.13	0.51 ± 0.04	0.54 ± 0.10
Random Periph.-ViT	-	52.90 ± 1.10	0.53 ± 0.01	0.53 ± 0.08	48.42 ± 4.01	0.48 ± 0.04	0.50 ± 0.06