Enhancing End-to-End Autonomous Driving Systems Through Synchronized Human Behavior Data

Yiqun Duan
yiqun.duan-1@uts.edu.au
University of Technology Sydney
Sydney, New South Wales, Australia

Zhuoli Zhuang
zhuoli.zhuang@uts.edu.au
University of Technology Sydney
Sydney, New South Wales, Australia

Jinzhao Zhou
jinzhao.zhou@student.uts.edu.au
University of Technology Sydney
Sydney, New South Wales, Australia

Yu-Cheng Chang
fred.chang@uts.edu.au
University of Technology Sydney
Sydney, New South Wales, Australia

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

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

Abstract

This paper presents a pioneering exploration into the integration of fine-grained human supervision within the autonomous driving domain to enhance system performance. The current advances in End-to-End autonomous driving normally are data-driven and rely on given expert trials. However, this reliance limits the systems' generalizability and their ability to earn human trust. Addressing this gap, our research introduces a novel approach by synchronously collecting data from human and machine drivers under identical driving scenarios, focusing on eye-tracking and brainwave data to guide machine perception and decision-making processes. This paper utilizes the Carla simulation to evaluate the impact brought by human behavior guidance. Experimental results show that using human attention to guide machine attention could bring a significant improvement in driving performance. However, guidance by human intention still remains a challenge. This paper pioneers a promising direction and potential for utilizing human behavior guidance to enhance autonomous systems.

CCS Concepts

• **Human-centered computing** → *Computer supported cooperative work*; • **Computing methodologies** → *Control methods*; **Vision for robotics**.

Keywords

Human-Guided Autonomous Driving; Brain-Computer Interfaces

1 Introduction

This position paper aims to pioneer the exploration of integrating granular human supervision into the burgeoning field of autonomous driving to enhance its performance. Presently, the majority of autonomous driving approaches, whether end-to-end or pipeline-based, rely heavily on expert trials. Such reliance is inadequate for considerations such as generalizability and earning human trust. This paper collects synchronous data from both human and machine driving scenarios to investigate this aspect.

The current trend in autonomous driving systems could be categorized into pipeline formations [14, 17, 28, 35] and End-to-End (E2E) approaches [10, 23, 31, 40]. These systems primarily transform raw sensory inputs into machine-readable representations using a variety of feature fusion techniques, such as BEV or range view. The goal is to develop autonomous driving models that can complete predetermined routes while safely navigating dynamic environments and complying with traffic rules, all through observational data-driven policies.

These approaches mentioned above are predominantly data-driven, relying heavily on models learning from expert examples. However, the pure data-driven formation also brings strong reliability to the data distribution. We suggest an augmentation of this methodology by incorporating fine-grained, immediate human labels as fine-tuning feedback to further enhance the robustness of the driving process. Our proposal involves a novel approach where humans and machines share the same driving scenarios. During these shared experiences, we aim to gather data from both human and machine drivers concurrently. This dual-data collection focuses on eye-gazing markers and brainwave data, offering a comprehensive view of the driving environment and decision-making processes. Incorporating human-guided data into the autonomous system is intuitively right as humans still perform better ability while dealing with a lot of scenarios. The utilization of human guidance could leverage human superior intelligence in complex scenarios (eg. correct attention in complex scenarios) intuitively benefiting from expert guidance in the learning process.

Introducing human guidance into the learning loop is a feasible and effective strategy, as evidenced by previous studies [5, 22, 29, 39]. However, this area has very limited research. Following the autonomous driving scenarios mentioned above, we propose the driving framework where the overall structure of our exploration is illustrated in Fig 1. It delineates the research into two main aspects
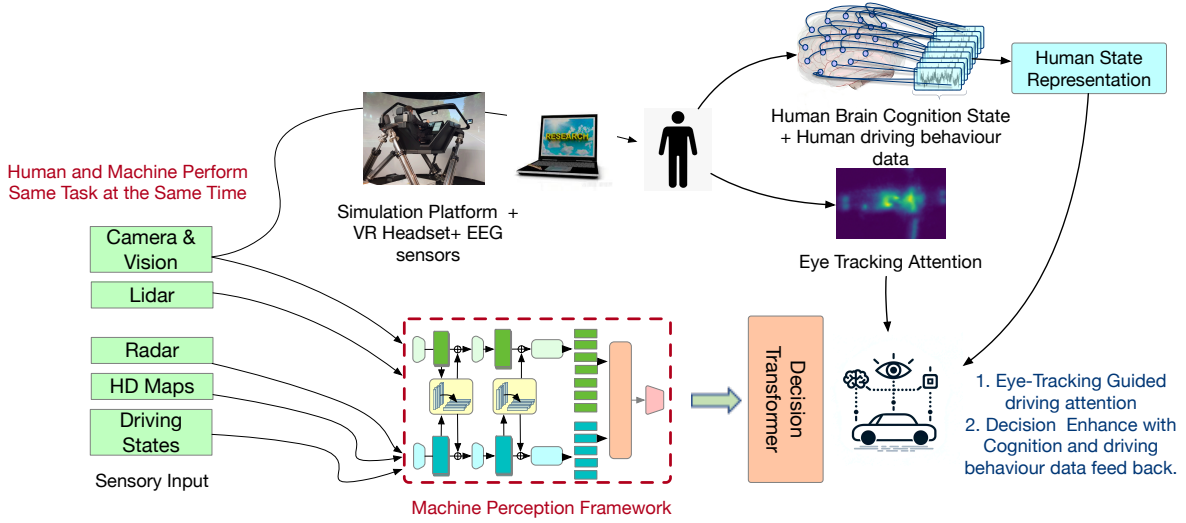
Figure 1: Visual schema of human enhanced autonomous driving. This model subjects both the machine and the human to identical driving scenarios along the same route. Throughout this process, we collect synchronous data on human behavior, including eye-tracking metrics, brainwave patterns, and brake signaling, capturing these elements in unison.

of exploring how human guidance could enhance the autonomous driving system: 1) eye tracking attention during human driving and 2) human cognition data generated through the driving procedure.

We initially explore the concept of "observing like a human" in Section 3.5.1. This involves analyzing eye movements during driving to uncover deeper insights into the cognitive processes and attentional focus of human drivers. Leveraging human attention as guidance, this approach seeks to address overlooked aspects by machines. It allows for the correction of instances where objects of potential significance are ignored by automated systems. This human-guided attention is instrumental in highlighting objects that align with human empirical skills but may not be represented in expert training datasets. Such guidance is invaluable for machines, as it aids in recognizing and responding to critical elements in various driving scenarios, including sudden hazards or intricate traffic situations. By integrating these human-centric insights, we can significantly enhance the machine's perceptual and decision-making abilities in dynamic driving environments.

Secondly, we explore how the human cognition or behavior data could help guide the autonomous driving model in Section 3.5.2. We explore the human cognition data to train an additional reward criticizer to provide additional feedback for additional imitation learning. To realize this goal, we first train a simple EEG wave classifier to recognize whether humans will decide to emergency break according to the current situation. Then, we feed the simultaneously collected EEG waves while human driving to this classifier and use the break signal as the additional feedback to the reinforcement fine-tuning of the driving model.

Experimental results are conducted in Section 4, where we respectively explore how eye-tracking guidance and human cognition data guidance in Section 3.5.1. The results suggest leveraging human eye-tracking data as an auxiliary feedback mechanism in imitation learning has yielded positive improvements in driving

performance. Specifically, this approach enhanced the driving score on the Carla Long-Set 6 dataset from 50.63 to 51.29. Conversely, the integration of cognition data did not result in a significant direct enhancement of the machine's driving score. At the position paper stage, we think this issue is reasonable because of two issues: 1) The recognition accuracy of the human brain data is still limited. This limited the guidance accuracy of human intention for driving. 2) The data quality is still relatively low, where the human driver's reaction speeds vary, and their decision-making capability varies when facing dangerous situations. This outcome suggests that the effective application of cognition data in autonomous driving remains an area ripe for further investigation. However, the effectiveness of human-observing data still outlines future research integrating human guidance in a collaborative human-machine driving context. The contribution of this paper could be categorized into four folds.

- This position paper pioneers enhancing machine driving by two aspects 1) observing like human drivers, and 2) making decisions like human drivers.
- This paper collects the parallel human cognition and behavior data of simultaneous machine and human driving.
- Experimental results suggest effectiveness of introducing human-guidance into autonomous driving.

2 Related Works

End-to-End Autonomous Driving. Although rapidly developing, autonomous driving (AD) technology [24], it faces challenges in generalization and safety. To remedy this issue, a number of works have explored imitation learning (IL) and reinforcement learning methods to leverage human guidance for enhancing training efficiency and the safety of the learned of driving policy [5, 8, 9, 25, 30]. Current methodologies for integrating human guidance into autonomous driving (AD) systems primarily categorize based on their

impact on the AD system into two distinct approaches: 1) leveraging human expert knowledge to train reliable driving policies, and 2) incorporating human oversight during model execution. Techniques such as Conditional Imitation Learning (CIL) [4], and Generative Adversarial Imitation Learning (GAIL) [13] rely on human expert trajectories for training the driving policy model. However, these methods face significant challenges, including the inefficiency of data collection and distribution shifts. Alternatively, a different set of approaches focuses on interpreting human biosignals during the evaluation phase, enabling the model to maneuver the car more safely under human guidance or feedback [11]. Subsequent research has proposed strategies for employing real-time human guidance to enhance the safety and performance of the human-AI co-driving system [12, 15, 36, 37]. These strategies involve dynamic control transfer between human drivers and AI agents to facilitate timely intervention, thereby improving the co-driving experience. This evolution in approach underscores the importance of human-AI collaboration in enhancing the effectiveness and safety of autonomous driving systems. To combine the merits of the above methods, our proposed methods also make use of human biosignals to improve the training efficiency of AI models while also being used in driving interventions.

Real-Time Human Guidance for Autonomous systems. Current methods for introducing real-time human guidance in autonomous driving scenarios predominantly rely on head movement [20] and eye-tracking signals [1, 41]. However, these approaches, focus primarily on external indicators of attention and intent without considering the critical role of cognitive processes in driving. There are only limited works exploring the real-time monitoring of the cognitive process in the context of driving safety [2, 3, 26, 34], however, they did not consider using the cognitive brain signal as guidance to the AD system. To address this gap, we propose an innovative research direction that integrates human cognitive signals alongside eye-tracking data to provide a more comprehensive understanding of driver intentions. By combining these sources of information, we aim to achieve a more seamless integration of the driver's objectives and safety considerations with the performance of the autonomous driving (AD) system. This holistic approach promises to enhance the symbiosis between human drivers and AD technologies, paving the way for advancements in driving safety and efficiency.

3 Methodology

3.1 Overview

Problem Setup: We follow the previous widely accepted setting of E2E driving ([10, 32, 40]) that the goal is to complete a given route while safely reacting to other dynamic agents, traffic rules, and environmental conditions. Thus, the goal is to learn a policy π given observation. We choose the **Imitation Learning (IL)** approach to learn the policy. The goal is to obtain policy π by imitating the behavior of an expert π^*. Given an expert, the learning dataset $\mathcal{D} = \{(\mathcal{X}^i, \mathcal{W}^i)\}$ could be collected by letting the expert perform similar routes, where $\mathcal{X}^i = \{(\mathbf{x}_{im}^i, \mathbf{x}_{Li}^i)_t\}_{t=1}^T$ denotes image and LiDAR sensory observations of the current state, and $\mathcal{W} = \{(x_t, y_t)\}_{t=1}^T$ denotes the expert trajectory of waypoints. Here, x_t, y_t denotes the 2D coordinates in ego-vehicle (BEV) space. Thus, the learning

target could be defined as in Eq. 1.

$$\underset{\pi}{\arg\min} \ \mathbb{E}_{(\mathcal{X}, \mathcal{W}) \sim \mathcal{D}} \left[\mathcal{L}_{wp}(\mathcal{W}, \pi(\mathcal{X})) \right] \qquad (1)$$

where $\mathcal{L}_{wp}$ is the waypoint loss defined in Eq. 4, and $\pi(\mathcal{X})$ is the predicted waypoints given observation $\mathcal{X}$ through policy π to be learned.

Formation: In this paper, the policy $\pi(\mathcal{X})$ is realized by the combination of a hybrid fusion network (Sec. 3.2) and the decision transformer (Sec. 3.3), where the fusion network transfer multi-modality sensory inputs $\mathcal{X}$ into semantic tokens $\mathbf{F}_s$, and decision transformer predicts future goal points $\mathcal{W}$ given $\mathbf{F}_s$. The human guidance is injected by Then a PID Controller is applied on the waypoints $\mathcal{W}$ decision and decomposed into practical control, ie., steer, throttle, and brake.

3.2 Hybrid Fusion Transformer Encoder

For basic sensor fusion we utilize MaskFuser [8] as our main ourk. MaskFuser proposed a hybrid network shown in the lower part of Fig. 2 that combines the advantages of *early fusion* and *late fusion*. The network consists of two stages.

Early Fusion: At the first stage, we apply two separate CNN branches to extract shallow features respectively from monotonic image and LiDAR inputs. For the image branch, MaskFuser concatenates three front view camera inputs each with 60 Fov into a monotonic view and reshaped into shape $3 \times 160 \times 704$. For the LiDAR branch, MaskFuser reprocesses the raw LiDAR input with PointPillar [16] into BEV feature with shape $33 \times 256 \times 256$. Since the lower-level features still retain strong geometric relations, the separated encoder could extract tight local feature representation with fewer distractions. A novel monotonic-to-BEV translation (MBT) attention is applied to enrich each modality with cross-modality assistance. MBT attention translates both images and LiDAR features into BEV feature space and performs a more precise spatial feature alignment compared to previous element-wise approaches.

Late Fusion: At the second stage, the network respectively tokenizes [6] feature maps from Image and LiDAR stream into semantic tokens, respectively denoted by green and blue in lower part of Fig. 2. The late fusion is performed by directly applying a shared transformer encoder over the concatenated token representation. The shared encoder with position embedding could force the tokens from various modalities aligned into a unified semantic space. Also, by treating multi-sensory observation as semantic tokens, we could further introduce masked auto-encoder training mentioned below.

3.2.1 Monotonic-to-BEV Translation (MBT). MBT attention performs cross-modality attention more precisely by introducing human prior knowledge (BEV transformation). Inspired by Monotonic-Translation [21], we model the translation as a sequence-to-sequence process with a camera intrinsic matrix. The detailed structure of MBT attentions is shown in Fig. 3, where the feature map from image stream with shape $\mathbf{F}_{im} \in \mathbb{R}^{N \times C \times H \times W}$ is reshaped along width dimension W into image columns $\mathbf{F}_c \in \mathbb{R}^{NW \times \{H \times C\}}$. The column vectors are projected into a set of mediate encoding $\{\mathbf{h}_i \in \mathbb{R}^{H \times C}\}_{i=1}^{NW}$ through a transformer layer with multi-head self-attention. We treat mediate encoding of monotonic view $\mathbf{h}_i$ which projects the key and value representing the range view information to be
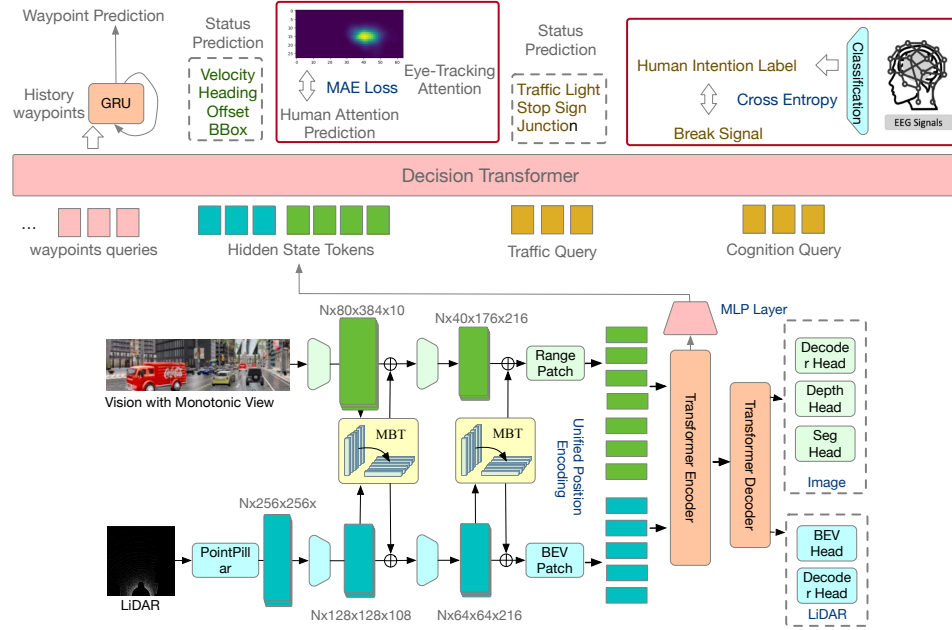
Figure 2: Framework of injecting human guidance into the autonomous system, taking autonomous driving as an example. The lower part is the hybrid fusion transformer encoder defined in Section 3.2. The learned machine state is fed into a decision transformer to come out with the final driving waypoint prediction. The decision transformer is supervised by waypoints prediction GT and other safety constraints. The human guidance is injected by adding two human-behavior data supervision branches. The decision transformer is required to reconstruct human behavior (eye-tracking attention and brake intention) jointly with other targets.
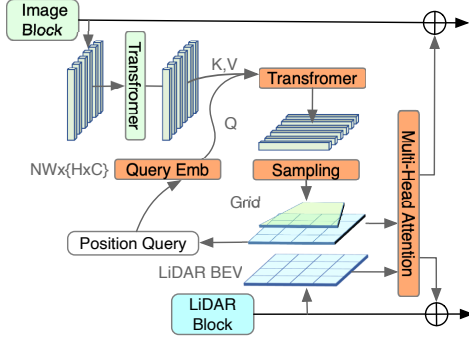

Figure 3: The structure of the MBT attention module, where the features from monotonic view are projected into BEV space through a sequence-to-sequence formation.

translated.

$$K(\mathbf{h}_i) = \mathbf{h}_i W_K, \quad V(\mathbf{h}_i) = \mathbf{h}_i W_V \quad (2)$$

A grid matrix is generated indicating the desired shape of the target BEV space. We generate position encoding $\{g_i \in \mathbb{R}^{r \times C}\}_{i=1}^{NW}$ along side each radius direction inside the grid with depth r. The grid position g_i is tokenized into query embedding and query the $\mathbf{h}_i$ with radius directions [1] as defined in Eq. 3 and generate the translated

feature map $\mathbf{F}_{mbt}$.

$$s(\mathbf{g_i}, \mathbf{h_j}) = \frac{Q(\mathbf{g}_i)K(\mathbf{h}_j)^T}{\sqrt{D}}, \quad Q(\mathbf{g}_i) = \mathbf{g}_i W_Q$$

$$\mathbf{F}_{mid} = \sum_i \frac{\exp(s(\mathbf{g_i}, \mathbf{h_j}))}{\sum_j \exp(s(\mathbf{g_i}, \mathbf{h_j}))} V(\mathbf{h}_i) \quad (3)$$

where $s(\mathbf{g_i}, \mathbf{h_j})$ is the scaled dot product [33] regularized by dimension D. Yet, since the monotonic view only suggests information inside certain FOV (Field of view), we apply a sampling process $\mathbf{F}_{mbt} = P(\mathbf{F}_{mid})$ to sample points inside FOV decided by camera intrinsic matrix into BEV feature map $\mathbf{F}_{mbt}$. The translated feature map $\mathbf{F}_{mbt}$ and feature map from LiDar $\mathbf{F}_{Li}$ is reshaped by flatten along width W and depth r into vectors and concatenate into sequence $\mathbf{F}_{in} = \text{cat}(\mathbf{F}_{mbt}, \mathbf{F}_{Li})$. The transformer layer is applied on $\mathbf{F}_{in} \in \mathbb{R}^{N^\star \times C}$ to perform self multi-head attention ([33]) between each token in $N^\star$ dimension.

3.3 Decision Transformer

Inspired by the previous work InterFuser [23], we use a similar transformer decoder as the decision layer. The decoder follows the standard transformer architecture, transforming some query embeddings of size d using K layers of multi-headed self-attention mechanisms. Five types of input are designed: $\{w_i\}_{i=1}^t$ previous waypoints query, R^2 density map queries (to query current vehicle status), human attention query, one traffic rule query, and one human intention query. These queries are concatenated into a sequence of tokens with shape $\mathbf{q} \in \mathbb{R}^{d \times N}$ which with the same shape

---

[1]The radius coordinates are calculated by given camera intrinsic matrix, FOV, and prefixed depth length.

with $\mathbf{z}$. To indicate the query order we add a standard positional embedding into the query $\mathbf{q}$ when doing the concatenation. In each decoder layer, we employ these queries to inquire about the desired information from the multi-modal multi-view features via the attention mechanism. More specifically, in each transformer layer, we treat perception state $\mathbf{z}$ as key $\mathbf{K}$ and value $\mathbf{V}$, and we treat the mentioned queries $\mathbf{q}$ as the query $\mathbf{Q}$. This structure allows the decision transformer query perception state accordingly with different queries inside the sequence, resulting in the output independently decoded into waypoints, one density map, human attention prediction, traffic status, and the human intention by the following prediction headers.

3.4 Basic Prediction Headers

The transformer decoder is followed by five parallel prediction modules to predict the waypoints, the object density map, human attention, traffic rule, and human intention respectively.

3.4.1 Waypoints Prediction. For the waypoints prediction, following the mentioned waypoints prediction network defined [24], we take a single layer GRU to auto-regressively predict a sequence of three future waypoints $\{w_{t+l}\}_{l=1}^{3}$. The GRU predicts the t+1-th waypoints by taking in the hidden state from the t-th decoded waypoint embedding from the transformer decoder related to the waypoints queries, and previous inputs from the recorded w_{t-2}, w_{t-1}, w_t. Also to inform the waypoints GRU predictor of the ego vehicle's goal location, we concatenate the GPS coordinates of the goal location at the beginning of the input sequence. More specifically, the loss function is defined in Equation 4 as follows:

$$\mathcal{L}_{wp} = \sum_{t=1}^{T} \left\| \mathbf{w}_t - \mathbf{w}_t^{gt} \right\|_1 \qquad (4)$$

where the $\mathbf{w}_t^{gt}$ is the ground truth from the expert route.

3.4.2 Density Map Status Prediction. The density map prediction forces the model to learn to predict the current vehicle status on a density map. We basically follow the original setting of Inter-Fuser [23] as this is not our research target. Please refer to appendix A for detailed definitions of loss $\mathcal{L}_{map}$.

3.4.3 Traffic Rule Prediction. For traffic rule prediction, the corresponding embedding from the transformer decoder is passed through a single linear layer to predict the state of the traffic light ahead, whether there is a stop sign ahead, and whether the ego vehicle is at an intersection. When predicting the traffic information $\mathcal{L}_{tf}$, we expect to recognize the traffic light status $\mathcal{L}_l$, stop sign $\mathcal{L}_s$, and whether the vehicle is at a junction of roads $\mathcal{L}_j$. All these three statuses are represented as a one-hot 0-1 label acquired from the CARLA simulator. These losses are simply calculated using cross entropy between prediction and ground truth as below:

$$\mathcal{L}_{tf} = \lambda_l \mathcal{L}_l + \lambda_s \mathcal{L}_s + \lambda_j \mathcal{L}_j, \qquad (5)$$

where λ balances the loss terms, which are calculated by binary cross-entropy loss.

3.5 Human-Guidance Headers

3.5.1 Human Eye-Tracking Attention Prediction. We design the eye tracking attention queries with shape $\mathbf{Q}_{eye} \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times d}$, where the $\frac{H}{16} \times \frac{W}{16}$ denotes the 16 times down-sampled range-view ratio of the camera, which has the exactly same ratio with the human attention ground truth map with ratio $H \times W$. The query $\mathbf{E}_{eye}$ is fed into the decision transformer by flattening $\frac{H}{16} \times \frac{W}{16}$ into a fixed length and outputs the prediction embedding the same shape. Then we use a transpose convolution to upsample the embedding from shape $\frac{H}{16} \times \frac{W}{16} \times d$ into human attention prediction $\mathbf{E}_{eye}$ with shape $H \times W \times 1$. We define the eye tracking prediction loss $\mathcal{L}_{eye}$ by calculating reconstruction loss between prediction $\mathbf{E}_{eye}$ and ground truth human attention $\bar{\mathbf{E}}_{eye}$:

$$\mathcal{L}_{eye} = \|\mathbf{E}_{eye} - \bar{\mathbf{E}}_{eye}\|^2, \qquad (6)$$

where we use mean square error (MSE) loss to supervise the human attention predictor head.

3.5.2 Human Intention Prediction. Given the synchronized collected data human intention EEG data, we use a pre-trained classifier [7, 42, 43] to determine whether the human has the intention to break the car we respectively represent the EEG classified label as $\mathbf{I}_{EEG}$. Meanwhile given the break behavior signal from the human driver's pedal as $\mathbf{I}_{brake}$, we propose to let the transformer jointly predict the human intention score. Similar to traffic rule prediction, we utilize a combined binary cross-entropy loss to supervise the decision transformer. The loss function is given below:

$$\mathcal{L}_{hb} = \lambda_{EEG} \mathcal{L}_{EEG} + \lambda_b \mathcal{L}_{brake}, \qquad (7)$$

where the human intention is supervised by calculating the binary cross-entropy combination.

3.6 Training Loss Combination

The loss function is designed to simultaneously predict multiple targets including, waypoints ($\mathcal{L}_{pt}$), object density map ($\mathcal{L}_{map}$), human attention ($\mathcal{L}_{eye}$), traffic rule ($\mathcal{L}_{tf}$), and human break intention ($\mathcal{L}_{hb}$).

$$L = \lambda_{pt} \mathcal{L}_{pt} + \lambda_{map} \mathcal{L}_{map} + \lambda_{map} \mathcal{L}_{map} + \lambda_{tf} \mathcal{L}_{tf} + \lambda_{hb} \mathcal{L}_{hb}, \quad (8)$$

where λ balances the three loss terms. Here, the waypoints ($\mathcal{L}_{pt}$), object density map ($\mathcal{L}_{map}$), and traffic rule ($\mathcal{L}_{tf}$) consists of the pure data driving supervision for autonomous driving similar Inter-Fuser [23] and the two additional loss items human attention ($\mathcal{L}_{eye}$) and human break intention ($\mathcal{L}_{hb}$) indicates the human guidance.

4 Experiments

4.1 Data Collection

The data are collected by collecting data simultaneously from humans and machines driving under the same route and simultaneously collecting the human data including human eye-tracking, human brain waves, and braking behaviors. The simultaneous human collection process is realized by letting human subjects drive using the Logitech G920 driving force wheels, pedals, and driving chair. The human visual attention is collected by letting humans wear a HTC Vive Pro Eye virtual reality gear. We use the widely
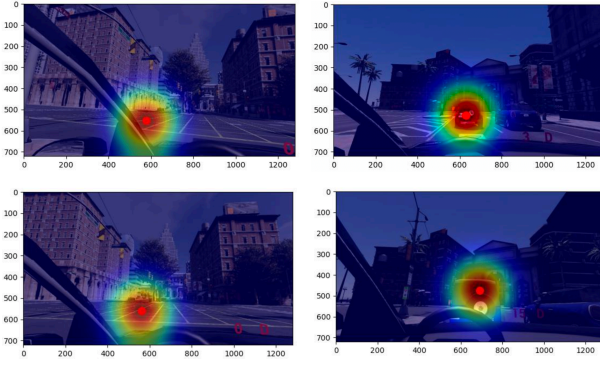
Figure 4: Visualization example of the collected combined human eye-tracking attention data.

Figure 5: Example of the statistical distribution of left and right eye's pupil position and diameter size change during one episode.

used CARLA simulator [2] version 0.9.13 as the driving simulator The driving scenarios are projected into embodied perspective for the human driver. Thanks to the previous work DReyeVR [3] [27], we directly utilize the proposed modified CARLA simulator to complete the data collection. To visualize the human eye's attention, we provide examples of collected eye-tracking data in Figure 4. Simultaneously, we collect the EEG data through a 64-channel collection device, where excluding the ground, reference, CB1, and CB2, the available count should be 60 out of 64. We collect 12 human subjects using the default routes under Town 4 and Town 7 for human guidance.

For the machine perception dataset, we employed a rule-based expert agent that adheres to the methodologies outlined in TransFuser [10] and InterFuser [23]. This agent was deployed across a diverse range of eight urban layouts and varying weather conditions, operating at a frequency of 2 frames per second. Through this process, we amassed an extensive expert dataset consisting of 3 million frames, which equates to approximately 410 hours of data. This dataset served as the foundation for the initial pretraining phase of our machine perception model.

The eye-tracking is projected into 2D positions while training. For further illustration of the collected data distribution, we further visualize the statistical distribution of the eye data in Figure 5.

4.2 Implementation Details

The model training is divided into two stages, where stage 1 follows the normal autonomous model training, and stage 2 utilizes the human-guided data to further finetune the decision model.

Machine State Pretraining: For the feature extractor part, we follow the setting proposed in Section 4.2, the Hybrid Network, which employs cameras and LiDAR as dual modalities. Camera inputs are merged into a 120-degree FOV and reformatted to (160, 704), while LiDAR data is transformed to a 256×256 BEV format [16]. For feature extraction, ImageNet-trained RegNet-32 [38] is utilized on both image and BEV LiDAR data. Training includes angular augmentation by ±20 degrees on LiDAR, akin to Transfuser [10],
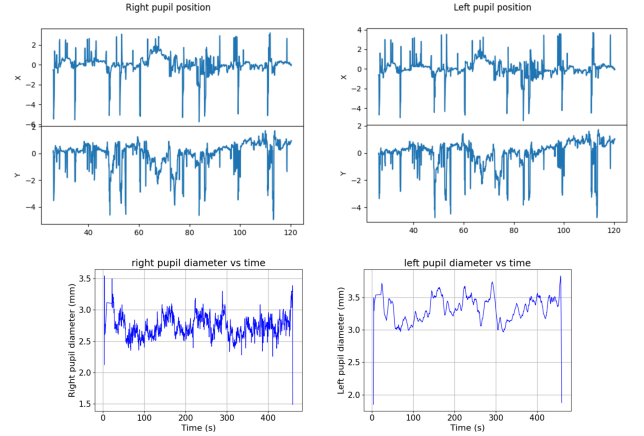
with corresponding label adjustments. Inter-modal bridging transformers (MBTs) integrate features post-initial convolutions, at resolutions ($C_1, 40, 176$) and ($C_2, 20, 88$), where $C_1 = 72$ and $C_2 = 216$. The MBT comprises two transformer layers with 512 hidden dimensions and 4 heads. Following established polar ray grid sampling [21], MBT attends to varying depth ranges, translating to a real-world coverage of up to 30.5 meters. Post-MBT, feature maps of $20 \times 88 \times 512$ and $32 \times 32 \times 512$ are sectioned into 4×4 patches, combined into a token sequence for a unified ViT encoder with four layers and 4 attention heads. This results in a semantic sequence of 174 tokens, maintaining the 512-dimensionality from early fusion.

For the decision transformer, the number of layers K in the transformer decoder and the transformer encoder is 6, and the feature dimension d is 256. The dimension of semantic token sequence output from the HybridNetwork is projected into 256 and fed into the decision transformer. We train our models using the AdamW optimizer [19] with a cosine learning rate scheduler [18]. The initial learning rate is set to $5 \times 10^{-4} \times \frac{\text{BatchSize}}{512}$ for the transformer encoder & decoder, and $2 \times 10^{-4} \times \frac{\text{BatchSize}}{512}$ for the encoders. The weight decay for all models is 0.07. All the models are trained for a maximum of 35 epochs with the first 5 epochs for warm-up [18]. For data augmentation, we used random scaling from 0.9 to 1.1 and color jittering.

Human-Guided Finetune Upon completion of the pretraining with the 3 million frames derived solely from machine-generated data, we transitioned to a phase of targeted refinement. In this stage, the perception model's parameters were fixed, and we exclusively fine-tuned the decision-making module of the transformer. For the EEG wave classifier, we pre-train a "human intention to brake" classifier based on the simultaneously collected EEG wave and the human behavior to brake label. This refinement utilized a smaller, yet highly nuanced dataset comprising 12,000 frames of human-derived data. The objective was to integrate human decision-making nuances into the machine perception model, thereby enhancing its ability to interpret complex driving scenarios. When we conduct the human-guided driving training, we use the initial learning rate

Table 1: Driving Evaluation on LongSet6 with Human Guidance, where DT denotes decision transformer.

Human Guidance	DT	Network	DS ↑	RC ↑ (%)	IS↑
×	×	TransFuser	46.95±5.81	89.64±2.08	0.52±0.08
×	×	MaskFuser	49.05±6.02	92.85±0.82 ×	0.56±0.07
×	✓	InterFuser	49.86±4.37	91.05±1.92	0.60±0.08
×	✓	MaskFuser + DT	<u>50.63±5.98</u>	**92.89±0.80**	0.62±0.08
Eye-Attention	✓	MaskFuser + DT	**51.39±5.33**	<u>92.01±0.97</u>	**0.65±0.09**
Intention	✓	MaskFuser + DT	50.06±5.81	90.97±1.80	<u>0.63±0.09</u>
Eye-Attention + Intention	✓	MaskFuser + DT	50.59±6.12	91.39±0.80	<u>0.63±0.09</u>
Eye-Attention + Fake Intention (GT)	✓	MaskFuser + DT	51.28±6.17	92.03±0.98	<u>0.64±0.08</u>
		Expert	75.83±2.45	89.82±0.59	0.85±0.03

as $1 \times 10^{-4} \times \frac{\text{BatchSize}}{512}$ for the decision transformer. The weight decay is kept the same as 0.07.

4.3   Evaluation with Human Guidance

Evaluation Benchmark: We directly keep the experimental settings the same as the pure machine driving evaluation on CARLA LongSet 6 [10]. We conduct our detailed ablation and comparison based on the Longeset6 Benchmark proposed by TransFuser [10], which chooses the 6 longest routes per town from the officially released routes from the CARLA Challenge 2019 and shares quite a similarity with the official evaluation.

Quantative Evaluation For both online evaluation and offline evaluation, we follow the official evaluation metrics to calculate three main metrics, **Route Completion (RC)**, **Infraction Score (IS)**, and **Driving Score (DS)**. The RC score is the percentage of route distance completed. Given R_i as the completion by the agent in route i, RC is calculated by averaging the completion rate $RC = \frac{1}{N} \sum_i^N R_i$. The IS is calculated by accumulating the infractions P_i incurred by the agent during completing the routes. The driving score is calculated by accumulating route completion R_i with infraction multiplier P_i as $DS = \frac{1}{N} \sum_i^N R_i P_i$. We also calculate the detailed infraction statistical details according to the official codes. We report the evaluation with the human-guided fine-tuning models on these metrics in Table 1.

It is observed from Table 1 that the integration of human-guided fine-tuning with the MaskFuser decision transformer (DT) leads to a noticeable improvement in Driving Score (DS). The improvement is mainly brought by human attention guidance, yet for the human intention guidance, we didn't observe a clear improvement. Here, the Eye-Attention guidance combined with the decision transformer achieves the highest mean driving score of 51.39, indicating the effectiveness of incorporating human attention to guide the driving model. Furthermore, the addition of intention data slightly reduces the driving score under both settings. This phenomenon is still reasonable because of two reasons. 1) The human intention recognition is not as accurate at this stage as the supervise label. During our experiments, the human intention classifier has an accuracy between $60\% - 70\%$ which is still compared low at this stage. 2) A long-existing problem appears that the simultaneously collected data has small time shifts between the human timestamp and the machine timestamp, this will lead to the wrong labels in rapidly changing situations. However, the improvement brought by the human attention guidance underscores the potential of leveraging nuanced human behavioral cues to enhance autonomous driving systems.

5   Limitation

In the experimental section, although using eye-tracking data has a positive impact on autonomous driving performance, using human-intention data does not. We attribute this issue to two main factors: First, human intention relies on a pre-trained EEG recognition model, which can only achieve a $60 - 70\%$ accuracy rate in identifying people's danger or braking intentions, thereby introducing noise. Second, compared to the abundant machine-generated autonomous driving data, collecting driver intention data is relatively costly. We utilized driving data from only 12 individuals, which may not be sufficient to significantly influence the training of large models. Moreover, intention data, compared to human attention, provides sparser supervision, necessitating more supervised data for training. This aspect warrants further discussion in future work.

6   Conclusion

In this paper, we delve into the utilization of human behavioral data to improve autonomous driving performance. We explore harnessing insights from human drivers to enhance the driving system's capabilities by two aspects 1) observing like a human, and 2) decision like a human. To achieve this, we collected eye-tracking and brake & cognition data from 12 human subjects by letting machines and humans drive the same route. The experimental results indicate that guiding machine attention with human attention can lead to a clear improvement in performance. However, the experiments did not demonstrate that human cognition data could significantly enhance outcomes. Integrating granular human supervision into machine driving merits further in-depth investigation. Such an approach is beneficial for increasing the machine's trustworthiness to humans while making its decision-making processes more anthropomorphic.

References

[1] Yaman Albadawi, Maen Takruri, and Mohammed Awad. 2022. A review of recent developments in driver drowsiness detection systems. *Sensors* 22, 5 (2022), 2069.

[2] Emad Alyan, Stefan Arnau, Julian Elias Reiser, Stephan Getzmann, Melanie Karthaus, and Edmund Wascher. 2023. Blink-related EEG activity measures

cognitive load during proactive and reactive driving. *Scientific Reports* 13, 1 (2023), 19379.

[3] Zehong Cao, Chun-Hsiang Chuang, Jung-Kai King, and Chin-Teng Lin. 2019. Multi-channel EEG recordings during a sustained-attention driving task. *Scientific data* 6, 1 (2019), 19.

[4] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. 2018. End-to-end driving via conditional imitation learning. In *IEEE International Conference on Robotics and Automation.* IEEE, 4693–4700.

[5] Felipe Codevilla, Eder Santana, Antonio M López, and Adrien Gaidon. 2019. Exploring the limitations of behavior cloning for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 9329–9338.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).

[7] Yiqun Duan, Charles Chau, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. 2024. Dewave: Discrete encoding of eeg waves for eeg to text translation. *Advances in Neural Information Processing Systems* 36 (2024).

[8] Yiqun Duan, Xianda Guo, Zheng Zhu, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. 2024. MaskFuser: Masked Fusion of Joint Multi-Modal Tokenization for End-to-End Autonomous Driving. *arXiv preprint arXiv:2405.07573* (2024).

[9] Yiqun Duan, Qiang Zhang, and Renjing Xu. 2024. Prompting Multi-Modal Tokens to Enhance End-to-End Autonomous Driving Imitation Learning with LLMs. *arXiv preprint arXiv:2404.04869* (2024).

[10] Prakash et al. [n. d.]. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. In *CVPR2021.* 7077–7087.

[11] Deepak Gopinath, Guy Rosman, Simon Stent, Katsuya Terahata, Luke Fletcher, Brenna Argall, and John Leonard. 2021. Maad: A model and dataset for" attended awareness" in driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 3426–3436.

[12] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. *Advances in neural information processing systems* 26 (2013).

[13] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems* 29 (2016).

[14] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790* (2021).

[15] W Bradley Knox and Peter Stone. 2011. Augmenting reinforcement learning with human feedback. In *ICML 2011 Workshop on New Developments in Imitation Learning (July 2011)*, Vol. 855. 3.

[16] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 12697–12705.

[17] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. 2022. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. *arXiv preprint arXiv:2205.13542* (2022).

[18] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[19] Ilya Loshchilov and Frank. Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[20] Mahdi Rezaei and Reinhard Klette. 2014. Look at the driver, look at the road: No distraction! no accident!. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 129–136.

[21] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. 2022. Translating images into maps. In *2022 International Conference on Robotics and Automation (ICRA).* IEEE, 9200–9206.

[22] Stefan Schaal. 1999. Is imitation learning the route to humanoid robots? *Trends in cognitive sciences* 3, 6 (1999), 233–242.

[23] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. 2022. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. *arXiv preprint arXiv:2207.14024* (2022).

[24] Hao Shao, Letian Wang, Ruobing Chen, Hongsheng Li, and Yu Liu. 2023. Safety-enhanced autonomous driving using interpretable sensor fusion transformer. In

[25] Jiamin Shi, Tangyike Zhang, Junxiang Zhan, Shitao Chen, Jingmin Xin, and Nanning Zheng. 2023. Efficient Lane-changing Behavior Planning via Reinforcement Learning with Imitation Learning Initialization. In *2023 IEEE Intelligent Vehicles Symposium (IV).* IEEE, 1–8.

[26] Justin Sia, Yu-Cheng Chang, Chin-Teng Lin, and Yu-Kai Wang. 2023. EEG-Based TNN for Driver Vigilance Monitoring. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI).* IEEE, 53–57.

[27] Gustavo Silvera, Abhijat Biswas, and Henny Admoni. 2022. DReyeVR: Democratizing Virtual Reality Driving Simulation for Behavioural & Interaction Research. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction.* 639–643.

[28] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. 2020. Pip: Planning-informed trajectory prediction for autonomous driving. In *European Conference on Computer Vision.* Springer, 598–614.

[29] Jingkai Sun, Qiang Zhang, Yiqun Duan, Xiaoyang Jiang, Chong Cheng, and Renjing Xu. 2024. Prompt, plan, perform: Llm-based humanoid control via quantized imitation learning. In *2024 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, 16236–16242.

[30] Lei Tai, Jingwei Zhang, Ming Liu, and Wolfram Burgard. 2018. Socially compliant navigation through raw depth inputs with generative adversarial imitation learning. In *2018 IEEE international conference on robotics and automation (ICRA).* IEEE, 1111–1117.

[31] Ardi Tampuu, Tambet Matiisen, Maksym Semikin, Dmytro Fishman, and Naveed Muhammad. 2020. A survey of end-to-end driving: Architectures and training methods. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[32] Marin Toromanoff, Emilie Wirbel, and Fabien Moutarde. 2020. End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7153–7162.

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems.* 5998–6008.

[34] Yu-Kai Wang, Tzyy-Ping Jung, and Chin-Teng Lin. 2015. EEG-based attention tracking during distracted driving. *IEEE transactions on neural systems and rehabilitation engineering* 23, 6 (2015), 1085–1094.

[35] Mingyun Wen, Jisun Park, and Kyungeun Cho. 2020. A scenario generation pipeline for autonomous vehicle simulators. *Human-centric Computing and Information Sciences* 10, 1 (2020), 1–15.

[36] Jingda Wu, Zhiyu Huang, Zhongxu Hu, and Chen Lv. 2023. Toward human-in-the-loop AI: Enhancing deep reinforcement learning via real-time human guidance for autonomous driving. *Engineering* 21 (2023), 75–91.

[37] Jingda Wu, Zhiyu Huang, Wenhui Huang, and Chen Lv. 2022. Prioritized experience-based reinforcement learning with human guidance for autonomous driving. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[38] Jing Xu, Yu Pan, Xinglin Pan, Steven Hoi, Zhang Yi, and Zenglin Xu. 2022. RegNet: self-regulated network for image classification. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[39] Qiang Zhang, Peter Cui, David Yan, Jingkai Sun, Yiqun Duan, Arthur Zhang, and Renjing Xu. 2024. Whole-body humanoid robot locomotion with human reference. *arXiv preprint arXiv:2402.18294* (2024).

[40] Qingwen Zhang, Mingkai Tang, Ruoyu Geng, Feiyi Chen, Ren Xin, and Lujia Wang. 2022. MMFN: Multi-Modal-Fusion-Net for End-to-End Driving. *arXiv preprint arXiv:2207.00186* (2022).

[41] Feng Zhou, X Jessie Yang, and Joost CF De Winter. 2021. Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE transactions on intelligent transportation systems* 23, 3 (2021), 2284–2295.

[42] Jinzhao Zhou, Yiqun Duan, Ziyi Zhao, Yu-Cheng Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. 2024. Towards Linguistic Neural Representation Learning and Sentence Retrieval from Electroencephalogram Recordings. *arXiv preprint arXiv:2408.04679* (2024).

[43] Jinzhao Zhou, Justin Sia, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2024. Masked EEG Modeling for Driving Intention Prediction. *arXiv preprint arXiv:2408.07083* (2024).