engineering proceedings

MDPI

Proceeding Paper

XAI-Interpreter: A Dual-Attention Framework for Transparent and Explainable Decision-Making in Autonomous Vehicles †

Candaş Ünal, Pelin Öksüz, Tolga Bodrumlu * and Musa Yazar

AVL Türkiye Research and Engineering, İstanbul, Türkiye; candas.unal@avl.com (C.Ü.); pelin.oksuz@avl.com (P.Ö.); musa.yazar@avl.com (M.Y.)

* Correspondence: tolga.bodrumlu@avl.com

† Presented at the 12th International Electronic Conference on Sensors and Applications (ECSA-12), 12–14 November 2025; Available online: https://sciforum.net/event/ECSA-12.

Abstract

Autonomous vehicles need to explain their actions to improve reliability and build user trust. This study focuses on enhancing the transparency and explainability of the decision-making process in such systems. A module named XAI-Interpreter is developed to identify and highlight the most influential factors in driving decisions. The module combines two complementary methods: Learned Attention Weights (LAW) and Object-Level Attention (OLA). In the LAW method, images captured from the ego vehicle's front and rear cameras in the CARLA simulation environment are processed using the Faster R-CNN model for object detection. GRAD-CAM is then applied to generate visual attention heatmaps, showing which regions and objects in the images affect the model's decisions. The OLA method analyzes nearby dynamic objects, such as other vehicles, based on their size, speed, position, and orientation relative to the ego vehicle. Each object receives a normalized attention score between 0 and 1, indicating its influence on the vehicle's behavior. These scores can be used in downstream modules such as planning, control, and safety. The module is currently tested in simulation. Future work will involve deploying the system on real vehicles. By helping the vehicle focus on the most critical elements in its surroundings, the Explainable Artificial Intelligence (XAI)-Interpreter supports more transparent and explainable autonomous driving systems.

Keywords: XAI-Interpreter; autonomous driving; visual attention; GRAD-CAM; object-level reasoning

Academic Editor(s): Name

Published: date

Citation: Ünal, C.; Öksüz, P.; Bodrumlu, T.; Yazar, M. XAI-Interpreter: A Dual-Attention Framework for Transparent and Explainable Decision-Making in Autonomous Vehicles. *Eng. Proc.* **2025**, *volume number*, x. https://doi.org/10.3390/xxxxx

1. Introduction

Autonomous vehicles play a key role in reshaping modern transportation systems in terms of safety, efficiency, and user experience. However, for these systems to be adopted, it is not enough for them to make correct decisions; they must also present these decisions in an explainable manner to the user. Transparency and clarity in decision-making processes are critical for building trust and ensuring the auditability of system behavior.

At this point, XAI methods offer solutions to increase user trust by making the internal workings of autonomous driving systems visible. Techniques such as visual attention maps, object-based importance scoring, and attention weight interpretation are frequently used to demonstrate which inputs influenced the system's decisions and how.

In the literature, various methods have been proposed to enhance the explainability of autonomous driving systems. Studies that utilize compact object-level representations instead of pixel-based visualizations have gained prominence. For example, the PlanT model proposed by Renz et al. enables faster and more efficient decision inference in the CARLA simulation environment compared to pixel-based approaches [1]. Chen and Krähenbühl introduced the "Learning from All Vehicles" approach, which improves driving performance by aggregating knowledge from different vehicles [2]. Nazat et al. presented the Multimodal-XAD model, which enhances explainability by combining bird's-eye view (BEV) representations with natural language explanations [3].

Furthermore, the XAI-ADS framework by Selvarajan et al. enables explainable anomaly detection based on time-series data [4], while Yuan et al. proposed the RAG-Driver model, which uses large language models to provide natural language explanations for driving actions [5]. Gao et al.'s VectorNet model improves explainability by offering graph-based driving predictions with object-level representations [6], and Kolekar and colleagues employed Grad-CAM after scene segmentation to produce visual explanations [7]. The work of Kim and Canny contributes to textual explanation of driving decisions and visualization of attention mechanisms [8,9]. Additionally, the systematic SafeX framework by Kuznietsov et al. helps classify existing XAI methods [10].

In this study, we present a module called the XAI-Interpreter, which integrates explainability as a core component of the system. This module utilizes two complementary methods to identify and visualize the factors influencing autonomous vehicle decision-making: LAW and OLA. In the LAW method, Grad-CAM is applied to objects detected via Faster R-CNN to generate attention maps, identifying the image regions that affect decisions [11,12]. The OLA method evaluates dynamic objects in the environment based on features such as velocity, position, and orientation. It then calculates a normalized attention score for each object, using these scores to inform the planning, control, and safety subsystems.

The following sections of the paper detail the explainability-oriented V-model architecture and describe the implementation and outcomes of the XAI-Interpreter module.

2. V-Model-Based Autonomous Driving Software Architecture

A V-model-based autonomous driving software architecture, which incorporates an explainable AI layer, is presented in Figure 1 [13]. This structure consists of five main layers, each designed hierarchically to include a perception and an actuator module. The Human-Machine Interface (HMI) and Monitoring Module, located at the lowest level, not only enables user interaction but also ensures that the system operates with minimal error. While the information flow between layers is defined sequentially, the HMI and Monitoring Module is uniquely designed to communicate directly with modules across all layers.

Based on the three main functional groups defined in the SAE J3016 standard by the Society of Automotive Engineers (SAE), the layers in the architecture can be classified as follows:

- Operational: Sensors—Actuators, Sensor Interface—Actuator Control Layers
- Tactical: Perception—Low-Level Controller, World Model—High-Level Controller Layers
- Strategic: Explainable AI (XAI) Interpreter—XAI Planner Layer

While the modules in the operational layer require high sampling rates, this requirement decreases for the tactical and strategic layers. In terms of computational load, the operational layer consumes fewer resources, whereas the strategic layer involves higher complexity and processing demand.

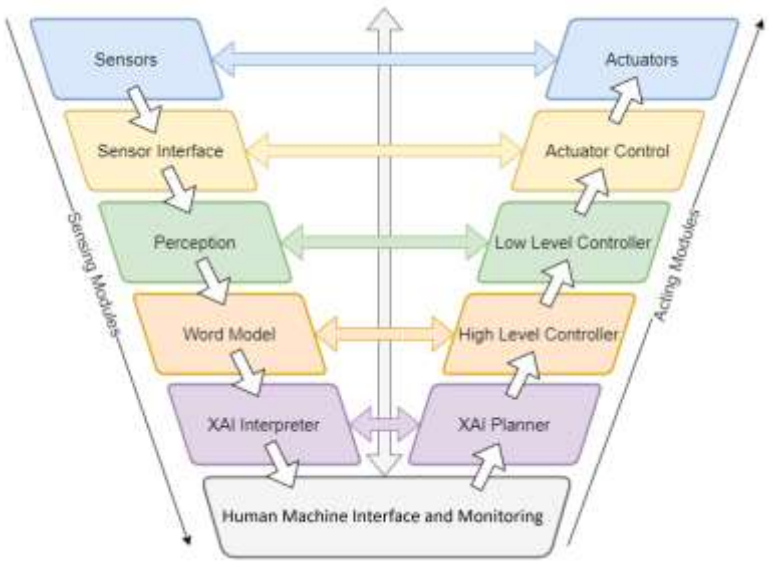

Figure 1. Autonomous Driving Software Architecture.

As a result, this architectural design enhances both the explainability of autonomous decision-making processes and the system-level utilization of the advantages provided by connected environments. This integration especially improves driving safety in complex traffic conditions and enables effective cooperative operation between connected and non-connected vehicles (CV and NCV).

Simulation-Based Autonomous Driving Architecture

The diagram presented in Figure 2 demonstrates how the developed software architecture is integrated in a unified manner. This simulation-based structure allows cooperative interactions between connected and non-connected vehicles to be modeled both in virtual environments and in parallel with physical hardware. The integrated architecture includes several core modules, which are detailed below:

- **Carla Sensors:** Provides virtual sensor data that detects objects around the vehicle within the simulation environment.
- **V2X Communication Layer:** Merges V2X data (e.g., position, speed, turn signals) from other vehicles in the simulation with Carla sensor data to improve situational awareness.
- **Vehicle Dynamics Model:** Represents the physical motion of the vehicle, executing commands from perception and planning modules to steer the vehicle.
- **Carla Environment:** Refers to the virtual environment where the simulation takes place, including road infrastructure and traffic elements.
- **Traffic Generator:** Dynamically creates CVs, NCVs, and other vehicles in the environment to test different traffic scenarios.
- **XAI Interpreter:** Interprets environmental conditions using perception and V2X data. It generates semantic information about traffic signs, surrounding vehicles, and environmental factors.
- **XAI Planner:** Makes strategic decisions based on the output of the interpreter, such as yielding to NCVs at intersections, lane changing, or speed adaptation.
- **Motion Planning and Control:** Converts strategic decisions from the XAI Planner into trajectory and velocity profiles, and forwards the result to the vehicle control module.
- **Human-Machine Interface (HMI):** Represents interaction with the human driver in simulation scenarios by providing signals, information, and alerts.

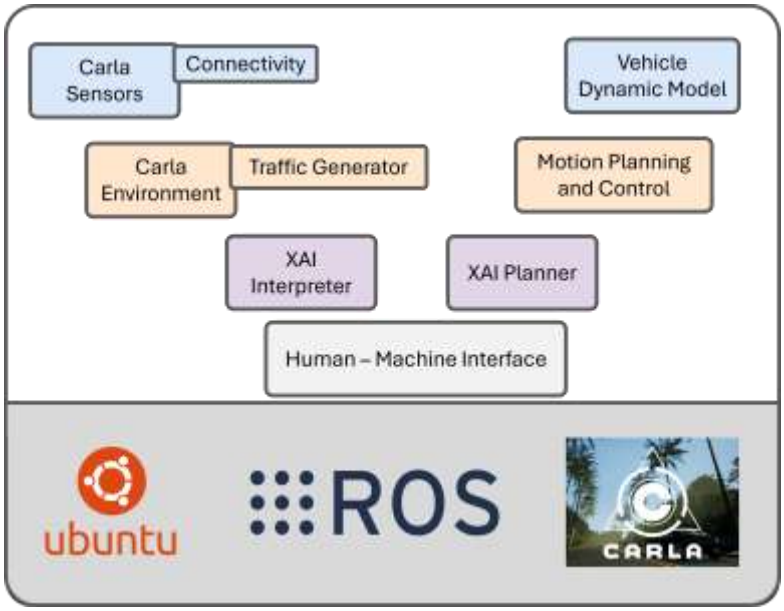

Figure 2. Autonomous Driving Architecture in Simulation Environment.

This architecture is integrated with the Carla simulation environment, ROS-based communication infrastructure, and real-time modules running on the Ubuntu operating system. Although simulated data is used instead of real-world sensor data, the V2X components of the architecture are developed to be fully compatible with physical hardware.

Following this architectural overview, the next section discusses the detailed structure of the **XAI Interpreter** module, one of the core components responsible for ensuring the system's explainability.

3. XAI-Interpreter Module

The XAI Interpreter is responsible for generating a semantic understanding of the environment by processing characteristic information such as traffic flow, inter-object relationships, and traffic signs located within the world modeling layer of the autonomous system. This module interprets environmental information by abstractly analyzing traffic rules, road topology, and dynamic environmental conditions.

In the interpretation process, not only real-time observations but also historical traffic patterns are considered to predict future scenarios. By doing so, the potential movement direction and position of every object in the dynamic environment can be estimated, allowing the system to foresee possible collisions or dangerous situations in advance. Using this information, the system can issue early warnings to the driver or decision-making modules, thereby enhancing safety and contributing to more optimal driving decisions.

The XAI Interpreter generates these predictions using probabilistic models and directly transmits the results to the Planning module. The planning process, in turn, can produce more informed and safer routes based on the risk assessment outputs.

Additionally, the objects that most influence the system's decision-making process are visually presented to developers or users through attention heatmaps (focus maps). This enables transparent observation of which objects or environmental elements were taken into account and to what extent.

With this structure, the XAI Interpreter functions not only as an environmental perception component but also as a core module that directly affects the explainability level of autonomous driving.

To ensure explainability in its operation, the XAI Interpreter employs two complementary mechanisms: Learned Attention Weights (LAW) and Object-Level Attention (OLA). Each of these methods is described in detail below.

3.1. Learned Attention Weights (LAW)

In deep learning models, not every pixel from camera images contributes equally during the prediction phase. Instead, the model selectively focuses on the most informative regions of the input, known as attention maps, which highlight the areas deemed most relevant for decision-making. This is useful for image classification and object detection. Thus, it makes the prediction made when looking at the model's output more interpretable and explainable. In line with these explanations, the combination of Faster RCNN and Grad Cam applied to an image obtained from the vehicle's camera looks like in Figure 3.

Figure 3. Grad-Cam and Faster RCNN output image applied on camera image.

Grad-CAM was used to extract attention maps. The Grad-CAM application generated attention maps by correlating them with the gradients of the activation maps in the last convolutional layer. In this study, the prediction performance and confidence scores of different architectures were compared. Faster R-CNN and Grad-CAM are integrated to obtain visual explanations of model predictions. Different versions of the model yield varying results depending on the scenario.

In this study, we aim to compare these variants to identify the most suitable architecture for our specific task. **ResNet-50** provided a good balance between interpretability and performance, achieving an average confidence score of **92.85%** and a prediction success rate of **96.42%**. The **MobileNetV3** Large model produced less accurate and scattered attention maps, with an average prediction rate of **87.14%** and an average confidence score of **69.47%.** While the **ResNet-101** model achieved the highest prediction performance with **98.31%** accuracy and **99.74%** confidence, the delay in inference time can be limiting in real-time applications.

Considering all these results, the **ResNet-50** model was determined to be the most suitable model for this study due to its performance in real-time scenarios and its high level of explainability.

3.2. Object-Level Attention (OLA)

The Object-Level Attention (OLA) module is designed as a submodule that determines how important the surrounding objects are to the driving behavior of the EGO vehicle. This submodule processes information about all vehicles in the same or adjacent lanes as the EGO vehicle as it moves along its trajectory and assigns each one a value between 0 and 1. A value of 1 indicates that the vehicle is highly important, whereas a value of 0 means it has very low importance.

The structure of the submodule is shown in Figure 4. Data coming from the world model is first acquired, then processed through data preprocessing and augmentation, and finally passed through the PlanT [1] model. The output is then published for the XAI Planner to use.

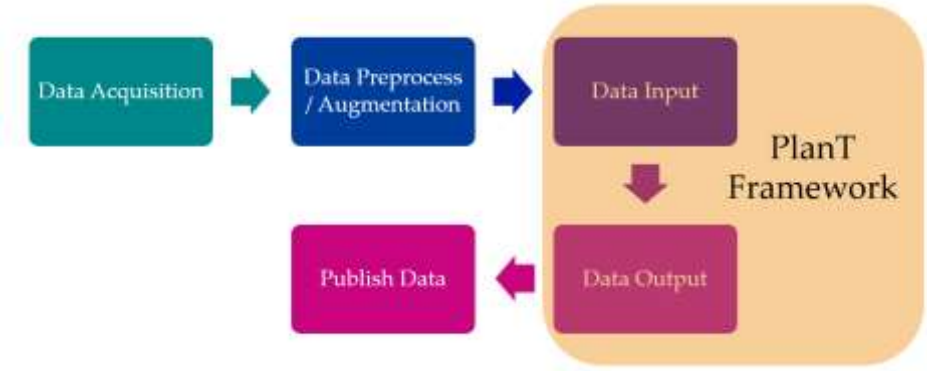

Figure 4. OLA Framework.

The attention values are calculated using a machine learning model called PlanT. This model, which is based on the Transformer architecture and trained via "Imitation Learning", learns by mimicking the behavior of another expert model with the support of Supervised Learning. As input, it uses features such as the surrounding vehicles' IDs, width, length, speed, heading, and distance to the EGO vehicle, along with the EGO vehicle's trajectory points. The model then produces attention values between 0 and 1 for each surrounding vehicle (Figure 5).

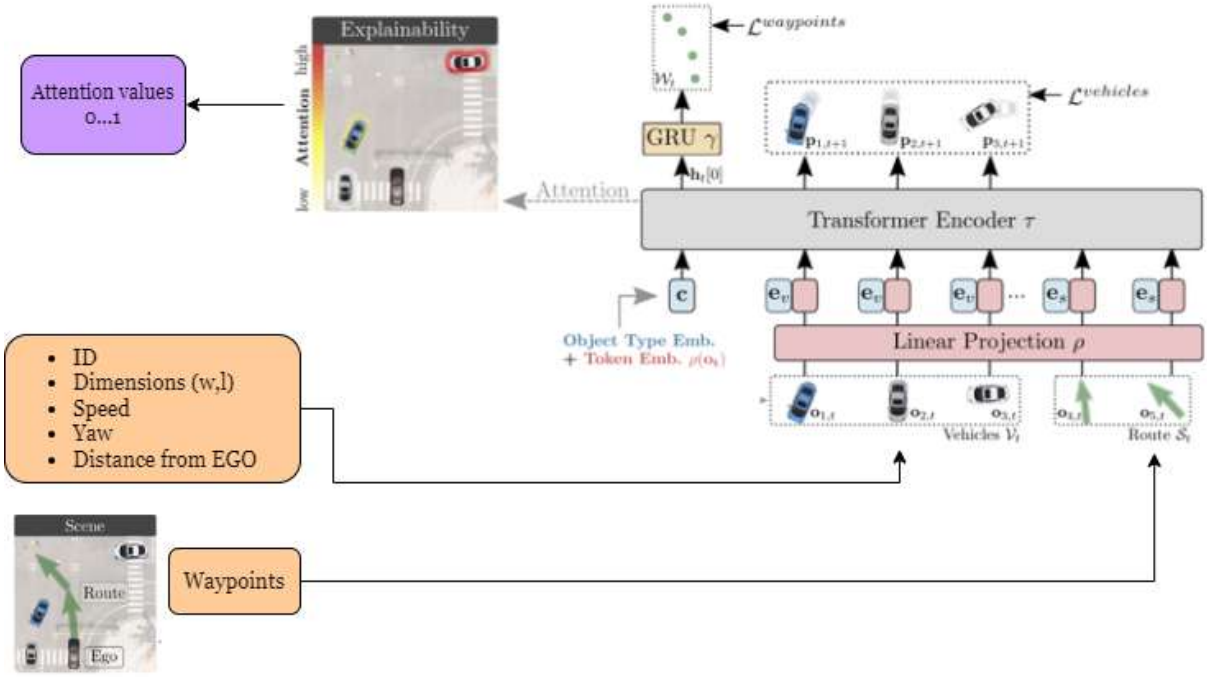

Figure 5. Framework of PlanT [1].

In the tests conducted in the CARLA simulation environment, the OLA submodule was evaluated under various autonomous driving scenarios (such as lane keeping, lane changing, and speed control). It was successfully operated using information from the world model, and the results were observed to be consistent with the expected attention values. In both lane-keeping and lane-changing scenarios, the submodule demonstrated

its ability to generate meaningful and usable attention values based on the EGO vehicle's waypoints as well as the distance, heading, and speed of the vehicles in the current and upcoming lanes.

In Figure 6 you can see the OLA values displayed on the surrounding vehicles during a lane-keeping scenario. During lane keeping, the OLA values for nearby vehicles in adjacent lanes are calculated to be high, while vehicles that are farther away from the EGO vehicle receive lower values compared to those nearby. In Figure 7, for an EGO vehicle performing both lane keeping and lane changing, the OLA submodule is able to correctly assess the complexity of the transition process and generate reliable attention values accordingly.

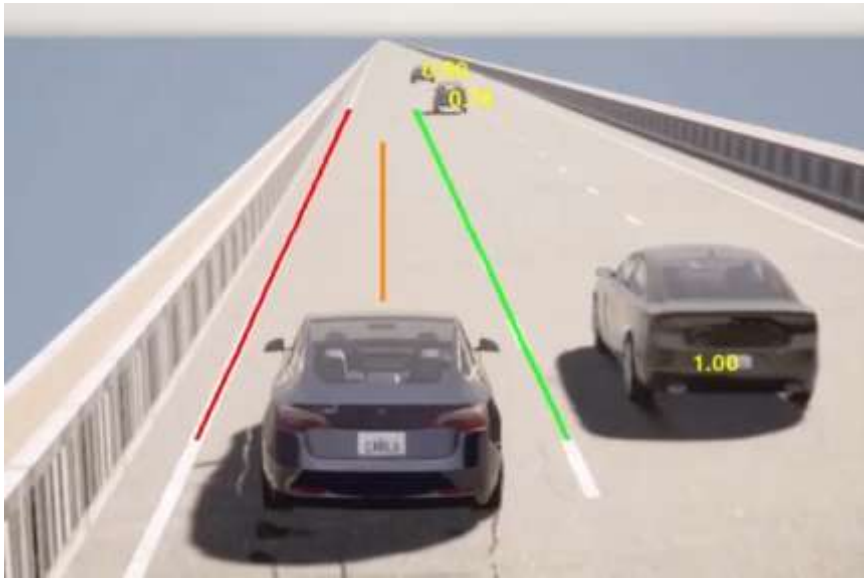

Figure 6. OLA Application on Lane Keep Scenario.

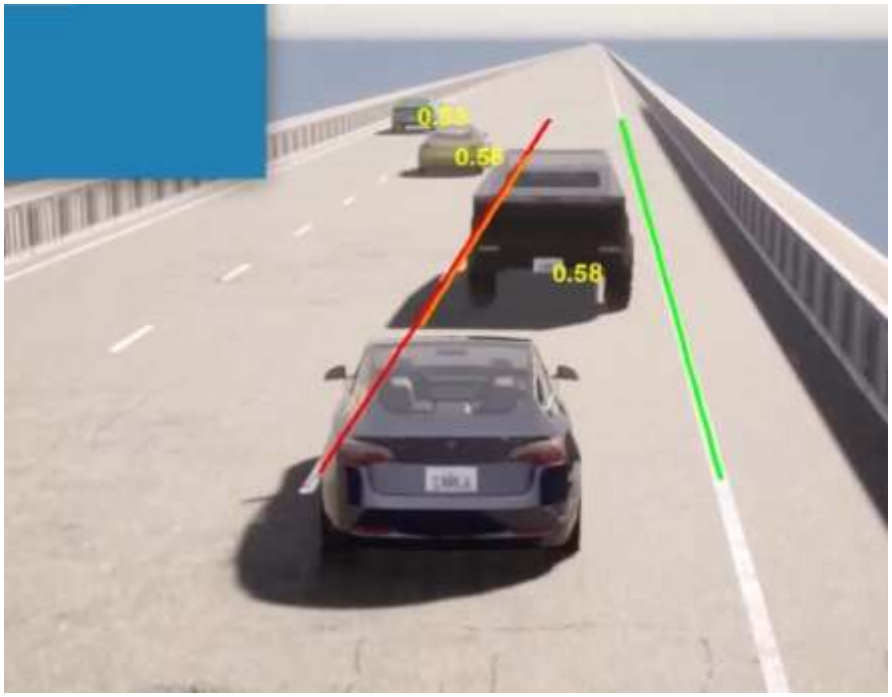

Figure 7. OLA Application on Lane Change Scenario.

4. Conclusions and Future Work

This study aimed to develop an XAI-supported decision-making infrastructure to address the requirements of safety, transparency, and user trust in autonomous vehicle systems. Within this scope, the proposed XAI-Interpreter module enhances the visibility and auditability of the system's internal operations by making the decision-making processes more understandable.

The developed solution is integrated into a V-model-based, multi-layered autonomous driving software architecture and has been modeled to work compatibly with both virtual and physical components.

The XAI-Interpreter comprises two main components that provide explainability: LAW and OLA. Within the LAW approach, attention maps were generated using the Grad-CAM method integrated with Faster R-CNN. These maps visually illustrate which regions of the image the model focuses on during decision-making. The ResNet-50 backbone demonstrated a balanced performance in terms of both attention map accuracy and inference speed, making it a suitable model for achieving interpretability alongside real-time operation.

In the OLA approach, dynamic objects in the environment are analyzed based on features such as position, velocity, and orientation, and each object is assigned a normalized attention score between 0 and 1. These scores, produced using a PlanT-based model, revealed the influence of the environment on the decision-making process at the object level showing that closer or more hazardous objects are assigned higher scores. Simulation-based tests demonstrated that this system produces consistent results across various driving scenarios and meaningfully contributes to behavior generation processes.

In conclusion, the architecture supported by the XAI-Interpreter module allows driving decisions to be monitored more transparently by both developers and users, enhances system reliability, and fosters more natural human-machine interaction.

Future work will focus on evaluating this architecture in hardware-supported test environments and conducting field tests with real vehicles. Additionally, the integration of the XAI approach into other domains such as driver alert systems is planned. In this direction, the XAI-Interpreter is positioned as one of the foundational building blocks of next-generation explainable systems in autonomous driving technologies, addressing ethical, safety, and user experience dimensions.

Author Contributions: Conceptualization, C.Ü., P.Ö., T.B. and M.Y.; methodology, T.B. and M.Y.; software, C.Ü. and P.Ö.; validation, C.Ü., P.Ö., T.B. and M.Y.; investigation, C.Ü., P.Ö., T.B., and M.Y.; supervision, T.B. and M.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data sharing Does not apply to this paper.

Acknowledgments: This work is supported by the Scientific and Technological Research Council of Turkey (TÜBİTAK) under the project Heterogeneous Integration for Connectivity and Sustainability (HiCONNECTS).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Renz, K.; Chitta, K.; Mercea, O.-B.; Koepke, A.S.; Akata, Z.; Geiger, A. PlanT: Explainable Planning Transformers via Object-Level Representations. In Proceedings of the Conference on Robot Learning (CoRL), Auckland, New Zealand, 14–18 December 2022.
2. Chen, D.; Krähenbühl, P. Learning from All Vehicles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022.
3. Feng, Y.; Feng, Z.; Hua, W.; Sun, Y. Multimodal-XAD: Explainable Autonomous Driving Based on Multimodal Environment Descriptions. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 19469–19481.
4. Nazat, S.; Li, L.; Abdallah, M. XAI-ADS: An Explainable Artificial Intelligence Framework for Enhancing Anomaly Detection in Autonomous Driving Systems. *IEEE Access* **2024**, *12*, 48583–48607.
5. Yuan, J.; Sun, S.; Omeiza, D.; Zhao, B.; Newman, P.; Kunze, L.; Gadd, M. RAG-Driver: Generalisable Driving Explanations with Retrieval-Augmented In-Context Learning in Multi-Modal Large Language Model. *arXiv* **2024**, arXiv:2402.10828.
6. Gao, J.; Sun, C.; Zhao, H.; Shen, Y.; Anguelov, D.; Li, C. VectorNet: Encoding HD Maps and Agent Dynamics from Vectorized Representation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
7. Kolekar, S.; Gite, S.; Pradhan, B.; Alamri, A. Explainable AI in Scene Understanding for Autonomous Vehicles in Unstructured Traffic Environments on Indian Roads Using the Inception U-Net Model with Grad-CAM Visualization. *Sensors* **2022**, *22*, 9677.
8. Kim, J.; Canny, J. Textual Explanations for Self-Driving Vehicles. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
9. Kim, J.; Canny, J. Interpretable Learning for Self-Driving Cars by Visualizing Causal Attention. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
10. Kuznietsov, A.; Gyevnar, B.; Wang, C.; Peters, S.; Albrecht, S.V. Explainable Artificial Intelligence for Autonomous Driving: A Systematic Review. *arXiv* **2024**, arXiv:2402.10086.
11. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
13. Yazar, M.N.; Ağın, B.; Öksüz, P.; Ünal, C.; Özdemir, C.E.; Özçelik, M.B.; Kaya, U. Açıklanabilir Yapay Zeka Katmanlı V-Model Otonom Sürüş Yazılım Mimarisi. In Proceedings of the Taşıt Araçları ve Teknolojileri Kongresi (TOK), İstanbul, Türkiye, 14–16 September 2023.