ManeuverVLM: A Novel Multimodal Fusion of Scene Images and Temporal Signals for Maneuver Prediction

Roksana Yahyaabadi[1] , **Soodeh Nikan**[1] ,

[1]Department of Electrical and Computer Engineering, Western University, London, Ontario, Canada,
{ryahyaab, snikan}@uwo.ca

Abstract

Maneuver prediction in modern vehicles enhances safety by anticipating driver actions, enabling advanced driver assistance systems (ADAS) to provide proactive support and accident prevention. This research presents 'ManeuverVLM', a vision-language model (VLM) that integrates scene images and dynamic signals for maneuver prediction. The model employs a vision encoder to extract spatial-visual embeddings, a temporal encoder for dynamic signals, and a large language model (LLM) for maneuver classification. We evaluated the proposed model on our collected dataset that covers five maneuvers: straight, left/right turn, and left/right lane change. Experimental results demonstrate that ManeuverVLM with T5-mini achieves superior performance with micro- and macro-accuracy (99%, 98%) and a macro F1-score of 97% on our collected driving dataset. Notably, ManeuverVLM effectively handles challenging minority maneuvers, such as turning and lane changing, outperforming both Temporal-Only and Spatial-Temporal models that do not integrate LLM. The proposed model, with 36.5 million parameters, 61.2 Giga floating point operations (GFLOPs), and requiring only 163 MB of memory, is deployable on compact embedded processors on the vehicle.

1 Introduction

According to the World Health Organization, traffic accidents worldwide result in more than 1.35 million fatalities annually [Dahl *et al.*, 2023]. As vehicle manufacturers continue to ADAS technology, their goal is to enhance driver safety and traffic efficiency [Murtaza *et al.*, 2023]. A key part of improving ADAS is accurately predicting driving maneuvers, such as turning, changing lanes, and straight driving, allowing the system to respond effectively and assist the driver in real-time [Khairdoost *et al.*, 2020]. Accurate perception of the environment is crucial for predicting maneuvers in intelligent vehicles, as it enables the system to understand the surrounding conditions and make informed driving decisions [Liu *et al.*, 2021]. Although traditional models primarily rely on kinematic and dynamic features [Khairdoost *et al.*, 2020] [Jiang *et al.*, 2022], [Li *et al.*, 2024], [Zhang *et al.*, 2022], scene images, despite their complexity, provide critical contextual information about road structures, traffic conditions, and obstacles. In modern autonomous vehicles, which cameras are widely used [Wang *et al.*, 2022], integrating scene images' information with dynamic signals enhances maneuver prediction accuracy by providing a richer environmental context [Gao *et al.*, 2022].

This study proposes ManeuverVLM, a multimodal strategy in which a vision language model (VLM) integrates dynamic signals and scene images for maneuver prediction. The architecture includes a vision encoder to extract spatial-visual embeddings from scene images and a temporal encoder for dynamic embeddings. These embeddings are fused into a shared latent space using a vision language adapter, with a large language model (LLM) acting as the classifier. This unified approach enables robust and effective prediction of driving maneuvers using both spatial and temporal characteristics. The paper is organized as follows. In Section 2, the related works are reviewed. Sections 3 and 4 present the collected dataset and methodology, while Section 5 evaluates the model performance and discusses the results. The paper concludes in Section 6.

2 Related Works

In the recent literature, several deep learning models have been proposed for maneuver prediction, each using different approaches, data sources, and features. Gao et al. [Gao *et al.*, 2022] introduced an attention-based deep neural network that combines spatial features of road scene images with temporal patterns from driver physiological signals. However, their model is highly computationally complex and demands resources due to the integration of Global Context blocks, Channel-wise Attention, and dual attention-based LSTM. Khairdoost et al. [Khairdoost *et al.*, 2020] proposed a long short-term memory (LSTM) model for predicting five maneuvers (left turn, right turn, left lane change, right lane change, and straight driving) using vehicle dynamics and driver behavior without incorporating scene context. Jiang et al. [Jiang *et al.*, 2022] proposed a dynamic Bayesian network trajectory prediction algorithm that models driver intention, maneuvering behavior, and vehicle dynamics for lane keeping and lane change prediction, but does not incorporate scene

information. In another study, Li et al. [Li *et al.*, 2024] proposed a Transformer encoder model with a multi-head attention mechanism to integrate spatio-temporal gaze attention inputs from drivers with vehicle dynamics for lane changing maneuver prediction. However, their model does not incorporate scene images and predicts only three maneuvers: lane keeping and left/right lane change.

To address these gaps, we proposed the ManeuverVLM which uses both scene images and dynamic signals to integrate spatial-visual and temporal features for maneuver prediction, covering five maneuvers: straight, left/right turn, and left/right lane change.

3 Description of the Data

The data set was collected by the Automotive and Surface Transportation team at the National Research Council of Canada (NRC). Data collection was carried out in London, Ontario, Canada using a 2021 RAV4 hybrid vehicle as shown in Fig. 1. The vehicle was equipped with a ZED stereo camera, a FLIR thermal camera, a Velodyne LiDAR with 32 channels, and an onboard Inertial Measurement Unit (IMU), all mounted on the roof. Data from all the mentioned sensors were collected using the Robot Operating System (ROS) framework and stored in bag-file format.

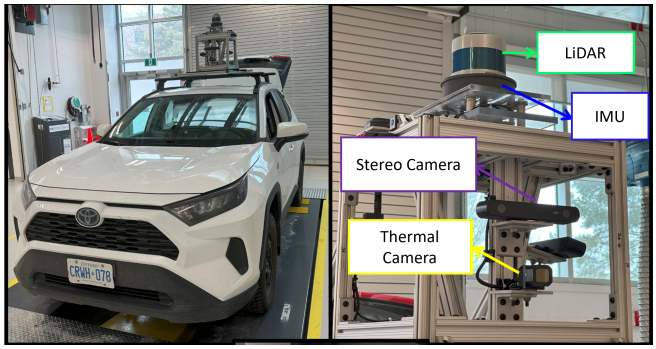

Figure 1: Equipped vehicle and sensors in data collection.

The dataset collected from four drivers (two females and two males), with approximately 30 minutes of driving data per driver. To ensure synchronization across sensors (LiDAR, cameras, and IMU), the LiDAR's frame rate of 10 Hz was chosen as the reference. Dynamic signals include 3D GNSS (global navigation satellite system) coordinates (latitude, longitude, altitude), 3D linear and angular velocity, and 3D linear and angular acceleration, totaling 15 features. These signals were recorded using an IMU with an embedded GNSS. The dataset comprises 81,190 samples labeled manually into five maneuver classes: 'straight', 'left turn', 'right turn', 'left lane change' and 'right lane change'. In addition, 85% of the samples belong to the straight class, the left and right turn classes account for 6% and 5% of the samples, respectively, and each lane change class comprises 2% of the samples.

4 Methodology

We proposed ManeuverVLM takes advantage of both spatial-visual and temporal features for maneuver prediction. To assess the impact of scene images and then the benefit of integrating an LLM, we also implemented two Temporal-Only model and a Non-VLM Spatial-Temporal model, as detailed in the following subsections.

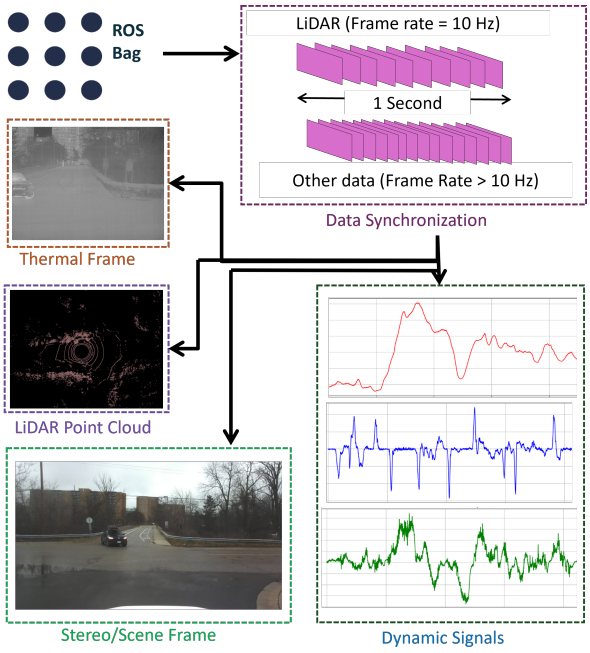

Figure 2: Overview of synchronization process of our dataset.

4.1 ManeuverVLM

As shown in Fig. 3, in the proposed architecture, MobileViT-small was selected as the vision encoder due to its efficiency in capturing both local and global scene information by combining lightweight convolutional and transformer models. Temporal features were modeled using a TCN consisting of two 1D convolution layers with ReLU activations, followed by adaptive average pooling and a linear projection to capture temporal patterns. The TCN model efficiently captures long-range temporal dependencies using a combination of causal and dilated convolutions, where causal convolutions ensure each time step depends only on past data, preserving temporal order, while dilated convolutions expand the receptive field without increasing parameters. The model architecture consists of two 1D convolutional layers with 64 and 128 filters, kernel size 3, padding 1, ReLU activation, adaptive average pooling, and a fully connected layer with 128 units. To fuse the extracted visual and temporal embeddings, a vision language adapter concatenates and passes them through a linear layer followed by LayerNorm and ReLU activation, mapping them into a joint multimodal embedding space. To balance efficiency and accuracy, our ManeuverVLM was evaluated using three T5 variants: T5-small, T5-mini, and T5-tiny.

4.2 Temporal-Only model

The Temporal-Only models in this study include a TCN and an LSTM network to capture the temporal patterns within the sequential dynamic signals. The TCN model uses two 1D

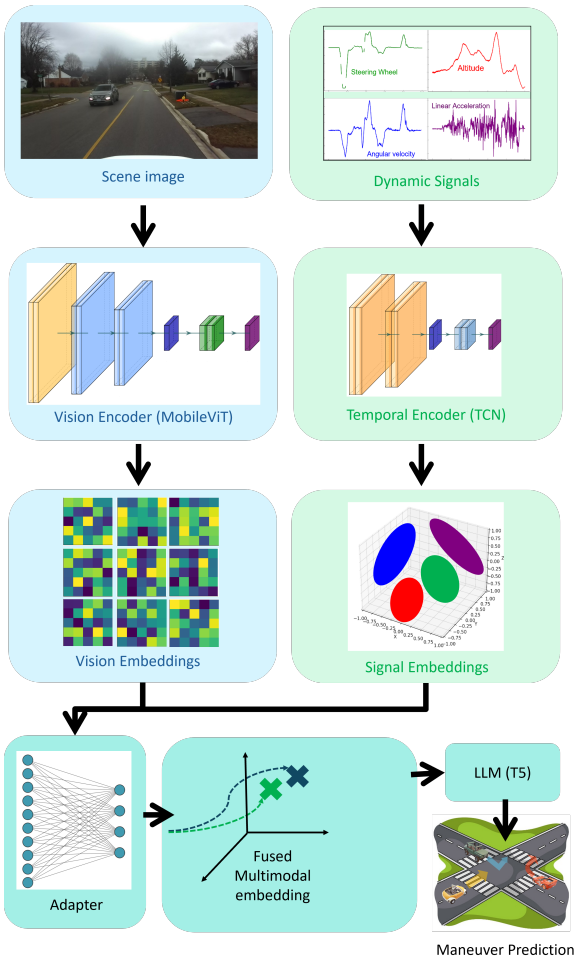

Figure 3: Overview of ManeuverVLM.

convolutional layers with 64 and 128 filters, kernel size 3, padding 1, ReLU activation, adaptive average pooling, and a fully connected layer with 128 units. Also, the LSTM model consists of three stacked LSTM layers with 128 hidden units, followed by a fully connected layer.

4.3 Non-VLM Spatial-Temporal model

The Non-VLM Spatial-Temporal model evaluates the impact of scene image context and visual features on maneuver prediction by combining spatial and temporal characteristics. It uses MobileViT-small for visual feature extraction and two temporal architectures: TCN and LSTM. The TCN model includes two 1D convolutional layers, ReLU, adaptive pooling, and a fully connected layer, while the LSTM model has three layers with a hidden size of 128. Visual and temporal embeddings are concatenated into a 256-dimensional feature vector, followed by a final fully connected layer for classification. An overview of the Non-VLM Spatial-Temporal model implemented is shown in Fig. 4.

5 Experimental Results

Our implementations were carried out on a 12th Gen Intel® Core™ i7-12700 processor, 2.10 GHz, supported by 32.0 GB RAM and a 12.0 GB NVIDIA GeForce RTX 3060 GPU.

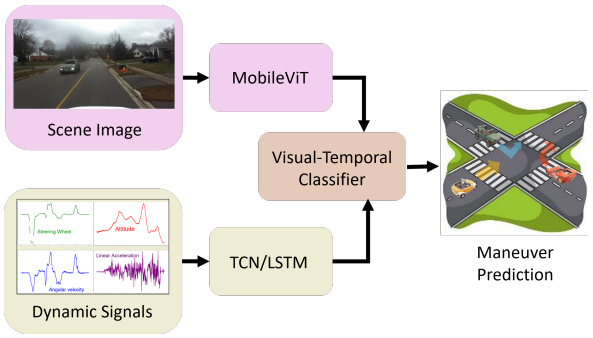

Figure 4: Overview of the Non-VLM Spatial-Temporal model.

5.1 Evaluation protocol

To ensure a comprehensive evaluation process, we followed a leave-one-driver-out cross-validation protocol. In this approach, for each fold, one driver's data is considered as the validation set, while the data from the remaining three drivers is treated as the training set.

5.2 Implementation Details

The Temporal-Only model was tested using two architectures: TCN and LSTM, both designed to process dynamic features for maneuver prediction. The models take dynamic signals with 15 features as input. All models were trained with the AdamW optimizer, a learning rate of 3×10^{-4}, a batch size of 32, and CrossEntropy loss for 40 epochs. For all models, evaluation was performed using micro and macro accuracy, and macro F1 score.

5.3 Results and Discussions

Table 1 presents a comparative analysis of the implemented models, including ManeuverVLM with different LLM variants, focusing on accuracy, F1-score, inference time, and resource requirements. The results highlight that integrating spatial-visual features from scene images in Non-VLM Spatial-Temporal models slightly enhances model performance compared to Only-Temporal, but significantly increases the number of parameters, GFLOPs, inference time, and inference memory. Among the evaluated LLM variants, the ManeuverVLM with T5-mini demonstrates the best overall performance, achieving 99% micro accuracy, 98% macro accuracy, and a 97% F1-score. These values are 3%, 1%, and 3% higher, respectively, compared to the ManeuverVLM with T5-tiny. Moreover, ManeuverVLM with T5-mini outperforms the T5-small variant by 3% in macro accuracy and 2% in macro F1-score, despite having fewer parameters. This indicates that for our dataset, the ManeuverVLM with T5-mini not only offers excellent performance but also strikes a balanced trade-off between efficiency and accuracy. It achieves results comparable to the T5-small variant while maintaining a lower parameter count, making it a practical and efficient choice for real-world applications.

Although the dataset is highly imbalanced, the ManeuverVLM perform exceptionally well in predicting minority

Model	Micro Accuracy	Macro Accuracy	Macro F1-score	Number of Parameters	Inference Time	GFLOPs	Required Memory
Temporal-Only (TCN)	0.92	0.56	0.54	29 K	2.86 ms	0.001	0.12 MB
Temporal-Only (LSTM)	0.90	0.48	0.48	342 K	0.28 ms	0.011	1.31 MB
Non-VLM Visual-Temporal (TCN)	0.93	0.75	0.61	5.1 M	21.81 ms	60.16	43.3 MB
Non-VLM Visual-Temporal (LSTM)	0.91	0.52	0.52	5.38 M	20.63 ms	60.17	44.5 MB
ManeuverVLM (T5-tiny)	0.96	0.97	0.94	20.8 M	10.45 ms	60.44	103 MB
ManeuverVLM (T5-mini)	**0.99**	**0.98**	**0.97**	36.5 M	10.88 ms	61.20	163 MB
ManeuverVLM (T5-small)	**0.99**	0.95	0.95	65.9 M	15.69 ms	62.10	275 MB

Table 1: Comparison between various implemented models and ManeuverVLM with different LLMs in terms of accuracy and resource requirements.

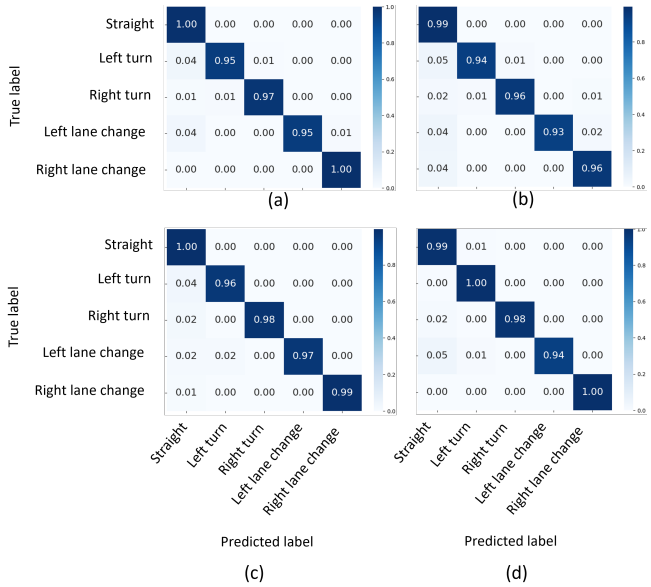

Figure 5: Confusion matrices for five maneuver classes across four-fold cross-validation using ManeuverVLM with T5-mini: (a) driver 1 as validation, (b) driver 2 as validation, (c) driver 3 as validation, and (d) driver 4 as validation.

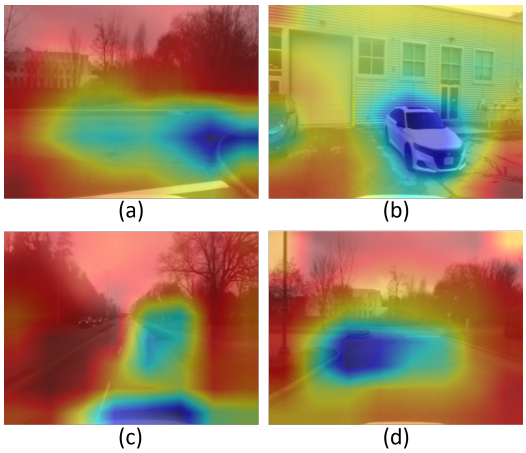

Figure 6: Examples of attention heatmaps extracted from the visual encoder (MobileViT-small) corresponding to: (a) Left turn, (b) Right turn, (c) Left lane change, and (d) Right lane change.

classes (turning and lane changing), significantly outperforming both Temporal-Only and Non-VLM Spatial-Temporal models, as evidenced by higher Macro accuracy and Macro F1-score. Although Temporal-Only models (TCN and LSTM) are lightweight in terms of parameters and resource usage, ManeuverVLMs demonstrate superior handling of imbalanced classification tasks. The increased parameter count and memory footprint in ManeuverVLMs stem from the dense embedding and attention mechanisms of transformer-based architectures. However, the optimized T5 models and GPU acceleration ensure comparable GFLOP and inference time with ManeuverVLMs, making them suitable for real-time performance. ManeuverVLM with T5-mini, comprising 36.5 million parameters, 61.2 GFLOPs, and 163 MB of inference memory, is efficient for deployment on resource-constrained embedded platforms such as Jetson Nano [Tailor *et al.*, 2023].

Figure 5 presents the confusion matrices for the 4-fold cross-validations, where each fold uses one driver as the validation subject and the others for training. The straight class shows nearly 100% accuracy, while minority classes (turning and lane changing) have higher error rates. Notably, Left maneuvers (turning and lane changing) are predicted more accurately than right maneuvers, due to the sharper dynamics and more distinct features associated with left turns and lane changes.

Figure 6 illustrates the heatmaps for the average attention weights, extracted from MobileViT-small component of ManeuverVLM with T5-mini, across all samples of driver 1. The heatmaps show that in turning maneuvers, the model mainly focuses on the surroundings, particularly on the corresponding left or right sides. Also, for lane changing maneuvers, the attention is distributed across both sides, with a particular focus on the lane markings and approaching objects (such as cars).

6 Conclusion

In this work, ManeuverVLM was proposed for driving maneuver prediction on our collected dataset. ManeuverVLM utilizes both scene image and dynamic signals for extracting spatial-visual and also temporal features. The results indicate that ManeuverVLM with T5-mini not only achieves the best performance, but also offers high efficiency, making it suitable for vehicle deployment and real-time applications.

Acknowledgments

This research has been supported by the Artificial Intelligent for Logistic program of National Research Council Canada (NRC).

References

[Dahl *et al.*, 2023] John Dahl, Gabriel Rodrigues de Campos, and Jonas Fredriksson. Prediction-uncertainty-aware threat detection for adas: A case study on lane-keeping assistance. *IEEE Transactions on Intelligent Vehicles*, 8(4):2914–2925, 2023.

[Gao *et al.*, 2022] Jun Gao, Jiangang Yi, and Yi Lu Murphey. Attention-based global context network for driving maneuvers prediction. *Machine Vision and Applications*, 33(4):53, 2022.

[Jiang *et al.*, 2022] Yuande Jiang, Bing Zhu, Shun Yang, Jian Zhao, and Weiwen Deng. Vehicle trajectory prediction considering driver uncertainty and vehicle dynamics based on dynamic bayesian network. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):689–703, 2022.

[Khairdoost *et al.*, 2020] Nima Khairdoost, Mohsen Shirpour, Michael A Bauer, and Steven S Beauchemin. Real-time driver maneuver prediction using lstm. *IEEE Transactions on Intelligent Vehicles*, 5(4):714–724, 2020.

[Li *et al.*, 2024] Jingyuan Li, Titong Jiang, He Liu, Yingbo Sun, Chen Lv, Qingkun Li, Guodong Yin, and Yahui Liu. Lane changing maneuver prediction by using driver's spatio-temporal gaze attention inputs for naturalistic driving. *Advanced Engineering Informatics*, 61:102529, 2024.

[Liu *et al.*, 2021] Yongkang Liu, Ziran Wang, Kyungtae Han, Zhenyu Shou, Prashant Tiwari, and John HL Hansen. Vision-cloud data fusion for adas: A lane change prediction case study. *IEEE Transactions on Intelligent Vehicles*, 7(2):210–220, 2021.

[Murtaza *et al.*, 2023] Mohsin Murtaza, Chi-Tsun Cheng, Mohammad Fard, and John Zeleznikow. The importance of transparency in naming conventions, designs, and operations of safety features: from modern adas to fully autonomous driving functions. *AI & society*, 38(2):983–993, 2023.

[Tailor *et al.*, 2023] Manav Tailor, Jahangir Ali, Xinrui Yu, Won-Jae Yi, and Jafar Saniie. Application of machine learning and image recognition for driver attention monitoring. In *2023 IEEE International Conference on Electro Information Technology (eIT)*, pages 1–5. IEEE, 2023.

[Wang *et al.*, 2022] Chaoyang Wang, Xiaonan Wang, Hao Hu, Yanxue Liang, and Gang Shen. On the application of cameras used in autonomous vehicles. *Archives of Computational Methods in Engineering*, 29(6):4319–4339, 2022.

[Zhang *et al.*, 2022] Hailun Zhang, Rui Fu, Chang Wang, Yingshi Guo, and Wei Yuan. Turning maneuver prediction of connected vehicles at signalized intersections: A dictionary learning-based approach. *IEEE Internet of Things Journal*, 9(22):23142–23159, 2022.