Improving Visual Q/A in Autonomous Driving Scenerio's Using LLM

Beena*
Dr. Chalavadi Vishnu*
{cs22m104,chalavadivishnu}@iittp.ac.in
Indian Institute of Technology Tirupati
Tirupati, Andhra Pradesh, India - 517619

ABSTRACT

In this study, we use Large Language Models (LLMs) and Vision Transformers (ViT) to investigate advances in Visual Question Answering (VQA) with the nu-scenes dataset of autonomous driving. Our methodology combines ViT for detailed visual feature extraction with LLMs for sophisticated textual interpretation to enable a comprehensive understanding of driving events. Our model intends to dramatically enhance accuracy in answering a wide range of questions relating to traffic signs, road conditions, and driver behaviors through pre-training on several multi-modal datasets and subsequent fine-tuning for task-specific adaptation. In order to improve interpretability and performance in autonomous systems, this project addresses the complexities of real-world problems and advances the integration of computer vision and natural language processing.

KEYWORDS

Visual Question Answering (VQA); Vision Transformers (ViT); Large Language Models (LLMs); Multi-modal Datasets; Autonomous Driving .

ACM Reference Format:
Beena and Dr. Chalavadi Vishnu. 2024. Improving Visual Q/A in Autonomous Driving Scenerio's Using LLM. In . ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The goal of autonomous driving technology is to build cars that can not only drive themselves across roadways but also recognize and react to the subtleties of human speech. To realize this vision, this project combines state-of-the-art technology in language processing and vision, particularly Large Language Models (LLMs) and Vision Transformers (ViT).

For autonomous cars to function safely and effectively in challenging conditions, they need to be able to process enormous volumes of textual and visual input. By processing visual data, Vision Transformers play a vital role in enabling the car to "see" and comprehend a wide range of information, including pedestrian movements, traffic signals, street signs, and other vehicles' actions. With the help of this technology, the vehicle can precisely sense and interpret its environment, making it safer to drive through busy and changing environments.

Large Language Models (LLMs) are used in parallel to analyze and produce human-like replies based on observations made by the vehicle and interactions with its human occupants. By processing natural language, these models enable the car to comprehend passenger requests and questions, as well as carry on meaningful discussions. For example, a traveler would inquire, "What's the

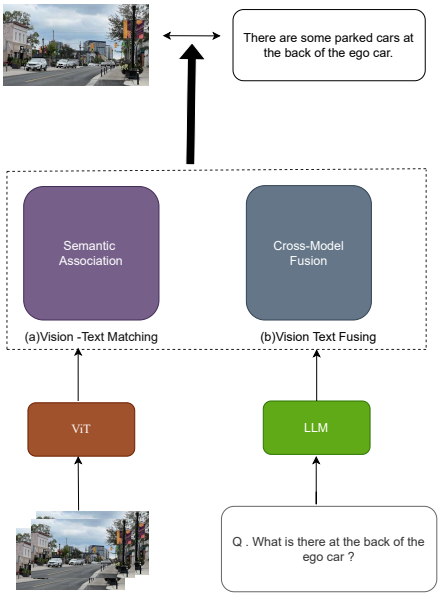

Figure 1: Two inter-modality connection approaches of Vision-Language Model in Autonomous Driving.(1)a Large Language Model (LLM) is used to process user questions that are sent in text format, while a Vision Transformer is used to interpret traffic photos.(2)The visual and text data are then incorporated early in the processing pipeline by early fusion, which combines these inputs.(3)Both the car's control signals and the user's inquiry responses are generated by the fused representation.

*Both authors contributed equally to this research.

Conference'24, August 2024, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/24/06
https://doi.org/XXXXXXX.XXXXXXX

cause of the traffic jam ahead?" and the car could analyze real-time data to provide an informed response.

But there are a lot of obstacles in the way of merging these two technologies. Conventional methods frequently handle the language and optical processing components as distinct entities, which can result in functionally fragmented outcomes where the car may detect barriers but be unable to comprehend spoken inquiries concerning them. Through the cooperative use of vision and language models, our study seeks to develop a more comprehensive system that will enable the smooth integration of language comprehension and visual data interpretation.

Our goal is to turn cars into extremely intelligent friends that can comprehend and provide precise answers to questions regarding driving. We are using cutting-edge technology like Large Language Models (LLMs) and Vision Transformers (ViT), as shown in figure 1, to accomplish this.

Integrating Vision Transformers with Large Language Models enhances autonomous vehicles with advanced vision and understanding capabilities. This combination makes cars safer and more proactive, turning them into intelligent and responsive partners on the road. Users can interact with the car in their own language, and the system reduces accidents by learning from historical data. Now, let's review the previous research that motivated this idea.

2 RELATED WORK

LLMs have the ability to make sensible and practical decisions in autonomous driving scenarios, as evidenced by recent research showing that they can function well in the majority of commonsense tasks [2][3][5].Furthermore, [15] shows that LLMs are able to learn from local regulations and accident reports, which helps them to effectively lower accident rates and increase the safety of autonomous driving.similarly,works like [14] and survey [4] shows further advancements in autonomous driving using LLM's.

2.1 Transformers in Multimodal Models

Transformers have fundamentally changed the field of machine learning since its invention by [13]. They enable more dynamic multimodal information interchange and a deeper contextual comprehension in the context of autonomous driving.Transfor-mers emerge as crucial orchestrators in [9] and [10], allowing for early fusion, and cross-modal attention. In a system where both image data and textual data are used, early fusion would involve concatenating the image features (e.g., from a Vision Transformer) with the text features (e.g., from an LLM) to create a unified input representation for the model and cross-modal attention would enable the model to focus on specific parts of an image that are relevant to the given text query. For instance, if the text asks about a "red car" in an image, the attention mechanism will help the model to focus on the regions of the image that likely contain the red car.

2.2 ViLT and ViLBERT: Pioneers of Joint Reasoning

ViLT and ViLBERT, introduced by [6] and [7], extend transformer capabilities by incorporating co-attention mechanisms. These models simultaneously focus on visual and textual stimuli, enhancing the alignment and understanding of verbal and visual data. This is crucial for autonomous driving, where interpreting dynamic visual scenes and relevant textual information (e.g., GPS data) is essential for safe operation.

However, despite their advanced capabilities, ViLT and ViLBERT face challenges in highly dynamic or unpredictable contexts. Their reliance on large pre-trained models and extensive datasets can lead to scalability and adaptation issues, especially in unique or rare scenarios not well represented in training data.

2.3 CNN-based Approaches: The Foundation

Before transformer-based models, autonomous vehicles relied heavily on Convolutional Neural Networks (CNNs) for visual processing. These CNN models [12], effectively extracted hierarchical features from visual inputs, aiding in object detection, lane maintenance, and environmental awareness.

While CNNs excel at identifying spatial hierarchies and patterns, they often struggle with capturing the broader context of scenes and can suffer from overfitting due to their high parameter count. To address these limitations, we employed Vision Transformers (ViT), which use fewer parameters and reduce complexity through Shallow Knowledge Infusion [11], infusing knowledge into the first transformer block.

Similarly, constraints mentioned in [8] motivated us to enhance the usability and efficiency of modern transformers. Next, we describe our approach, including our architecture, datasets, and metrics.

3 PROPOSED ARCHITECTURE

In this subsection explaining the working of the proposed model by defining different parts of the multimodal shown in figure 2.

- **Input:**The model takes an image as input. This could be in the form of pixel values representing the visual content.Simultaneously, the model receives a caption or textual input associated with the image. This could be a question or a query related to the content of the image.
- **Vision Encoder :** The Vision Encoder processes the image input. Utilizes convolutional neural network (CNN) layers to capture hierarchical and spatial features in the visual information.Produces a feature representation that encodes essential visual patterns.
- **Language Encoder:**The Language Encoder processes the caption text input.Utilizes recurrent neural networks (RNNs), transformers, or similar architectures to understand the sequential information in the textual input.Generates a feature representation that captures contextual information from the caption.
- **Cross-Modality Fusion:**Combines information from both the visual and textual modalities.The output feature representations from the Vision Encoder and Language Encoder are combined, creating a unified representation that integrates both visual and textual cues.A fused representation that enriches the model's overall understanding of the input.
- **Pooler:**Acts as a selection mechanism to extract essential information from the fused representation.Utilizes pooling

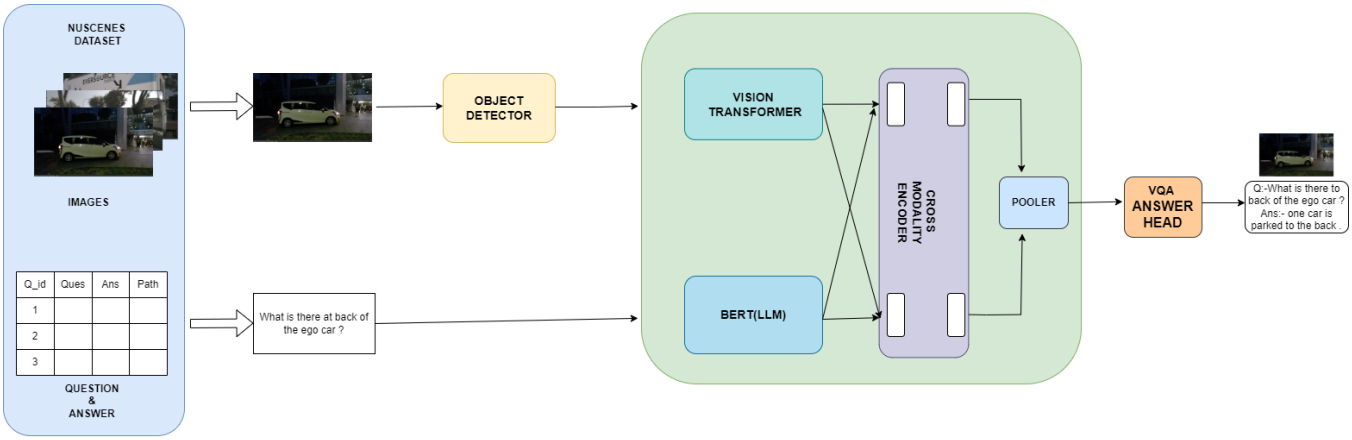

Figure 2: The model's operation is shown in the figure. text inquiry to language encoder; input image to object detector in vision transformer. fuse early, use a pooler to extract key features, and produce an image and a text response.

operations (e.g., max pooling, average pooling) to downsample and focus on relevant information in a computationally efficient manner.Further refined and focused representation.

- **Prediction Head:**The final layer responsible for generating ultimate answers based on the processed information.Involves fully connected layers and a binary classification loss function.Produces the final answer or prediction, often in the form of a binary decision (e.g., yes/no).
- **Output:**The final output is the model's answer to the question or query associated with the input image.If the question is binary (e.g., "Is there a cat in the image?"), the output could be "Yes" or "No" based on the model's prediction.

These were the sequential steps in the pipeline that our architecture took to create our model.Upon delving into the design, an inquiry emerges regarding the rationale behind the decision to utilize this model as a foundation for visual question answering in the autonomous vehicle domain.which we require to see the assessments and outcomes of these model training and testing procedures.

4 METHOD

This section explains the dataset collection ,performance metric and training testing pipeline through architecture of model.

4.1 Dataset & Metrics

We first tested on VQA (visual question answering)dataset to check the performance of our model because its small in size, image and question answer pair data is relatable to our model architecture .Later, we have used NuScenes dataset in our work, introduced by [1], offers multi-modal data for autonomous driving, including 360-degree camera, LiDAR, and RADAR data, plus IMU and GPS. It provides detailed 3D annotations, object attributes, and trajectory data from diverse urban driving scenes.

For evaluating the performance of the multi-moadal we used the WUP (Wu-Palmer) similarity score metric which measure the similarity between two words based on their depth in a hierarchical structure such as WordNet.

The WUP similarity score is calculated using the following formula:

$$\text{WUP Score} = \frac{2 \times \text{Depth of the Lowest Common Subsumer (LCS)}}{\text{Depth of Word 1} + \text{Depth of Word 2}}$$

Where:

- The "Depth of the Lowest Common Subsumer (LCS)" is the depth of the lowest common ancestor in the hierarchical structure.
- The "Depth of Word 1" and "Depth of Word 2" are the depths of the words in the hierarchy.

To train this multi-model we have used some arguments and hyper-parameters defines in next section.

4.2 Training Process

The training process involved defining the training arguments, initializing the model, and training it using the prepared dataset. The training was performed using the Hugging Face Trainer API, which facilitated efficient training and evaluation.

4.2.1 *Steps Involved.*

(1) **Define the training arguments:**
- Let θ be the set of model parameters.
- Define hyperparameters: learning rate α, batch size B, number of epochs E.
- Output directory D_{out}.
- Evaluation strategy (e.g., after each epoch).

(2) **Create the multimodal VQA collator and model:**
- Text model: $\text{BERT}_{\text{text}}$.
- Image model: $\text{ViT}_{\text{image}}$.
- Multimodal model M is defined as:

$$M(x_{\text{text}}, x_{\text{image}}; \theta) = f_{\text{combine}}(\text{BERT}_{\text{text}}(x_{\text{text}}), \text{ViT}_{\text{image}}(x_{\text{image}}))$$

where f_{combine} represents the fusion function of text and image features.

(3) **Initialize the Trainer:**
- Training dataset $\mathcal{D}_{\text{train}}$ and validation dataset $\mathcal{D}_{\text{val}}$.

- Loss function $\mathcal{L}$.
- Initialize trainer with model M, datasets $\mathcal{D}_{\text{train}}$, $\mathcal{D}_{\text{val}}$, and hyperparameters.

(4) **Train the model:**

- For each epoch e from 1 to E:

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\mathcal{D}_{\text{train}}; \theta)$$

- Evaluate model on validation dataset $\mathcal{D}_{\text{val}}$:

$$\text{Evaluate}(\mathcal{D}_{\text{val}})$$

where Evaluate computes validation metrics like accuracy and loss.

4.3 Testing Process

The testing process involves evaluating the model's performance on new and unseen data by splitting the dataset into a portion dedicated to testing.

4.3.1 Steps Involved.

(1) **Split Dataset:**

- Split dataset $\mathcal{D}$ into training set $\mathcal{D}_{\text{train}}$, validation set $\mathcal{D}_{\text{val}}$, and testing set $\mathcal{D}_{\text{test}}$.

(2) **Assess Model on Testing Set:**

- Use the trained model M to make predictions on $\mathcal{D}_{\text{test}}$:

$$\hat{y} = M(x_{\text{text}}, x_{\text{image}}; \theta)$$

- Compare predictions $\hat{y}$ with ground truth y to compute metrics.

(3) **Evaluate Performance:**

- Calculate metrics WUP Score and other such as accuracy, precision, recall, and F1-score. for classification

(4) **Adjust and Fine-Tune:**

- Based on testing results, adjust hyperparameters α, B, θ:

$$\theta' \leftarrow \text{Optimize}(\theta, \text{metrics})$$

- Fine-tune the model using the adjusted parameters and retrain on $\mathcal{D}_{\text{train}}$.

5 EVALUATIONS

We switched from training and assessing our models on indoor datasets to the NuScenes dataset. Rich outdoor scene collections are available in the NuScenes dataset, which is especially useful for challenges involving autonomous driving and visual comprehension in real-world settings. Here, we showcase the outcomes of using the NuScenes dataset to train the BERT_ViT and RoBERTa_ViT models.

5.1 BERT_ViT Model Performance

The contrast between true and predicted labels for a subset of classes is shown in Figure 3. Four of the first seven classes demonstrate alignment between true and predicted labels, indicating correct classification by the models. This alignment underscores the models' ability to understand and interpret the data accurately.

However, there are discrepancies between the true and predicted labels for some classes, suggesting areas where the models require further training or improvement. These differences provide valuable insights into the models' efficiency and potential optimization areas.. The classification report for the BERT_ViT model is presented below:

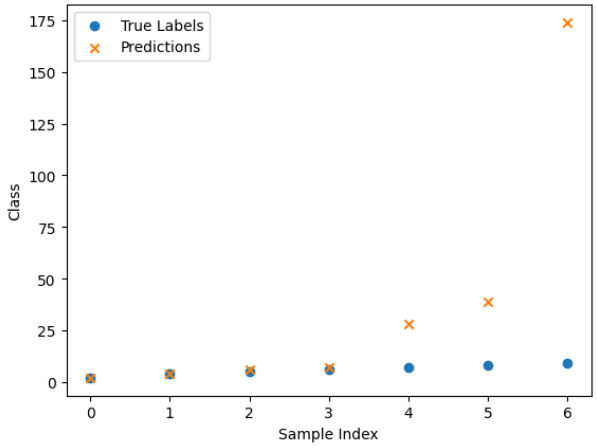

Figure 3: Graph showing correctness between samples of 7 different class by predicted and true_labels

Class	Precision	Recall	F1-Score	Support
2	1.00	1.00	1.00	1
4	1.00	1.00	1.00	1
5	0.00	0.00	0.00	1
6	0.00	0.00	0.00	1
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	1
9	0.00	0.00	0.00	1
28	0.00	0.00	0.00	0
39	0.00	0.00	0.00	0
174	0.00	0.00	0.00	0

Table 1: Classification Report for BERT_ViT Model

Table 1, presents the classification report for the BERT_ViT model. Classes 2 and 4 achieved perfect precision, recall, and F1-score, indicating correct classification with no false positives or negatives. Conversely, classes 5, 6, 7, 8, and 9, which include small or distant objects like potholes, traffic lights, and bicycles, all have zero precision, recall, and F1-score, showing the model's inability to identify these instances accurately. Additionally, classes 28, 39, and 174 have zero precision, recall, and F1-score due to insufficient training data or model limitations.

Overall, the classification report highlights the strengths and weaknesses of the BERT_ViT model across various classes. The model achieved an accuracy of 0.5619 and a label-versus-label agreement of 1.0. Further analysis with the RoBERTa_ViT model revealed interesting results in accuracy and WUPS scores, as discussed in the next subsection.

5.2 RoBERTa_ViT Model Performance

According to the analysis shown in Table2, the model accurately classified classes 4, 5, 6, 7, and 185 with perfect precision, recall, and

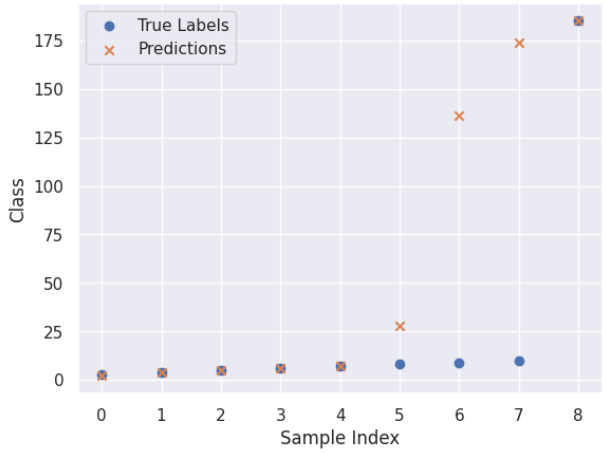

Figure 4: Graph showing correctness between samples of 9 to different class by predicted and true_labels

Class	Precision	Recall	F1-Score	Support
2	0.00	0.00	0.00	0
3	0.00	0.00	0.00	1
4	1.00	1.00	1.00	1
5	1.00	1.00	1.00	1
6	1.00	1.00	1.00	1
7	1.00	1.00	1.00	1
8	0.00	0.00	0.00	1
9	0.00	0.00	0.00	1
10	0.00	0.00	0.00	1
28	0.00	0.00	0.00	0
136	0.00	0.00	0.00	0
174	0.00	0.00	0.00	0
185	1.00	1.00	1.00	1

Table 2: Classification Report for RoBERTa_ViT Model

F1-score (all equal to 1), indicating no false positives or negatives. However, classes 2, 3, 8, 9, and 10 had F1-scores, precision, and recall values of 0, showing the model's inability to identify these classes accurately. Additionally, classes 28, 136, and 174 had no support, indicating insufficient training data or model limitations, with precision, recall, and F1-scores of 0.

Overall, the classification report highlights the model's strengths and weaknesses across different classes. The RoBERTa_ViT model achieved an accuracy of 0.6259 and perfect label versus label agreement of 1.0, outperforming the BERT_ViT model in classification tasks on the NuScenes dataset. However, precision, recall, and F1 scores varied across classes, with some having perfect scores and others zero due to lack of data.

Next, we examine the impact of changing the text model from BERT to RoBERTa.

5.3 Comparison

We observe that the RoBERTa_ViT model outperforms the BERT_ViT model in terms of accuracy and WUPS score. However, both models exhibit similar F1 scores. This indicates that the RoBERTa_ViT model provides better semantic understanding and partial match with the ground truth labels compared to the BERT_ViT model.

Model	WUPS Score	Accuracy (%)
BERT_ViT (VQA)	0.340	27
BERT_ViT (Nuscenes)	0.2857	56.19
RoBERTa_ViT (Nuscenes)	0.4444	62.59

Table 3: State-of-the-Art Comparison of BERT_ViT and RoBERTa_ViT Model on two different datasets VQA dataset and Nuscenes dataset

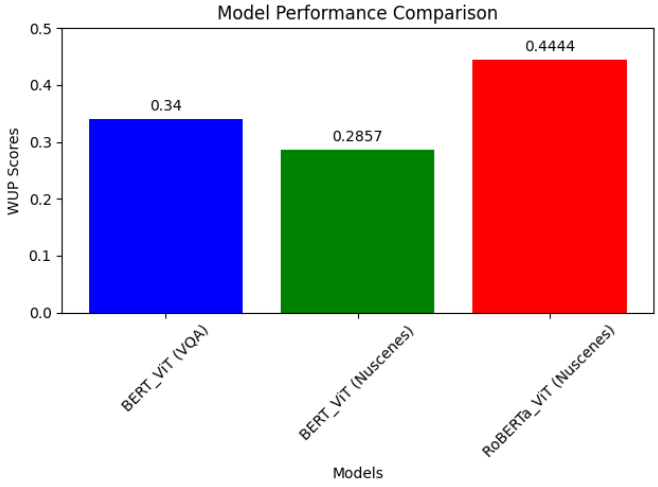

Figure 5: Plot showing Wup score performance for different models with different dataset

In the State-of-the-Art (SOTA) comparison presented in Table 3, we evaluate the performance of BERT_ViT and RoBERTa_ViT models on both indoor and outdoor datasets. For indoor datasets, BERT_ViT achieved a WUPS score of 0.340 with an accuracy of 27%, setting the benchmark for indoor scene understanding tasks. However, RoBERTa_ViT's performance on indoor datasets is not available. Transitioning to outdoor datasets, both models demonstrate improved performance. BERT_ViT achieves a WUPS score of 0.2857 and an accuracy of 56.19%, whereas RoBERTa_ViT surpasses its counterpart with a WUPS score of 0.4444 and an accuracy of 62.59%. These results indicate that RoBERTa_ViT outperforms BERT_ViT on outdoor datasets, suggesting its superiority in handling complex scene understanding tasks in real-world environments.

6 CONCLUSION & FUTURE WORK

In this study, we examined the effectiveness of two cutting-edge multi-modal models on both indoor and outdoor datasets: BERT_ViT

and RoBERTa_ViT. Our tests showed that RoBERTa_ViT performed better than BERT_ViT, especially on outdoor datasets, exhibiting higher accuracies and WUPS scores. This demonstrates its potential for use in robots and autonomous driving, among other applications.

Our classification reports provide thorough explanations of the models' performance in different classifications. While both models demonstrated good performance in a number of domains, we pinpointed some classes that require additional refinement. This highlights the requirement for multimodal models to be continuously improved in order to attain dependable performance in a variety of scenarios.

Our model architecture's capacity to operate with various datasets improves its resilience and adaptability, which is one of its main advantages. In order to further improve model performance and broaden application areas, future work will concentrate on integrating more sophisticated architectures and larger datasets, as well as exploring with additional modalities like audio and sensor data.

REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A Multimodal Dataset for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11621–11631.

[2] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Drive as you speak: Enabling human-like interaction with large language models in autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 902–909.

[3] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, and Ziran Wang. 2024. Receive, reason, and react: Drive as you say, with large language models in autonomous vehicles. *IEEE Intelligent Transportation Systems Magazine* (2024).

[4] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 958–979.

[5] Yaodong Cui, Shucheng Huang, Jiaming Zhong, Zhenan Liu, Yutong Wang, Chen Sun, Bai Li, Xiao Wang, and Amir Khajepour. 2023. DriveLLM: Charting the path toward full autonomous driving with large language models. *IEEE Transactions on Intelligent Vehicles* (2023).

[6] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. VILT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[7] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in Neural Information Processing Systems* 32 (2019).

[8] José Maurício, Inês Domingues, and Jorge Bernardino. 2023. Comparing vision transformers and convolutional neural networks for image classification: A literature review. *Applied Sciences* 13, 9 (2023), 5521.

[9] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems* 34 (2021), 14200–14213.

[10] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention Bottlenecks for Multimodal Fusion. *Advances in Neural Information Processing Systems* 34 (2021), 14200–14213. https://doi.org/10.48550/arXiv.2107.00135

[11] Kaushik Roy, Yuxin Zi, Vignesh Narayanan, Manas Gaur, and Amit Sheth. 2023. Knowledge-infused self attention transformers. *arXiv preprint arXiv:2306.13501* (2023).

[12] 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* 33, 4 (2020), 1364–1384.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017). https://doi.org/10.48550/arXiv.1706.03762

[14] Yi Yang, Qingwen Zhang, Ci Li, Daniel Simões Marta, Nazre Batool, and John Folkesson. 2024. Human-centric autonomous systems with llms for user command reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 988–994.

[15] Ou Zheng, Mohamed Abdel-Aty, Dongdong Wang, Chenzhu Wang, and Shengxuan Ding. 2023. Trafficsafetygpt: Tuning a pre-trained large language model to a domain-specific expert in transportation safety. *arXiv preprint arXiv:2307.15311* (2023).

(a) CAM_FRONT

(b) CAM_FRONT_LEFT

(c) CAM_FRONT_RIGHT

(d) CAM_BACK

(e) CAM_BACK_LEFT

(f) CAM_BACK_RIGHT

```
Question: What is the moving status of object <c3,CAM_FRONT,725.1,500.9>?
Answer: Going ahead. (Label: 4)
Predicted Answer:        Going ahead.
**********************************************************
```

(g) Output

Figure 6: Some visual results from Different Camera direction images and question answer pair related to images for that minute.