Journal of Software Engineering and Applications, 2023, 16, 672-684
https://www.scirp.org/journal/jsea
ISSN Online: 1945-3124
ISSN Print: 1945-3116

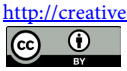
Object Detection Meets LLMs: Model Fusion for Safety and Security

Zeba Mohsin Wase[1], Vijay K. Madisetti[2], Arshdeep Bahga[3]

[1]School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India
[2]School of Cybersecurity and Privacy, Georgia Institute of Technology, Atlanta, Georgia, USA
[3]Cloudemy Technology Labs, Chandigarh, India
Email: zeba.wase@gmail.com, madisetti.vijay@gmail.com, arshdeep@cloudemy.io

How to cite this paper: Wase, Z.M., Madisetti, V.K. and Bahga, A. (2023) Object Detection Meets LLMs: Model Fusion for Safety and Security. *Journal of Software Engineering and Applications*, **16**, 672-684.
https://doi.org/10.4236/jsea.2023.1612034

Received: November 22, 2023
Accepted: December 24, 2023
Published: December 27, 2023



Abstract

This paper proposes a novel model fusion approach to enhance predictive capabilities of vision and language models by strategically integrating object detection and large language models. We have named this multimodal integration approach as VOLTRON (Vision Object Linguistic Translation for Responsive Observation and Narration). VOLTRON is aimed at improving responses for self-driving vehicles in detecting small objects crossing roads and identifying merged or narrower lanes. The models are fused using a single layer to provide LLaMA2 (Large Language Model Meta AI) with object detection probabilities from YoloV8-n (You Only Look Once) translated into sentences. Experiments using specialized datasets showed accuracy improvements up to 88.16%. We provide a comprehensive exploration of the theoretical aspects that inform our model fusion approach, detailing the fundamental principles upon which it is built. Moreover, we elucidate the intricacies of the methodologies employed for merging these two disparate models, shedding light on the techniques and strategies used.

Keywords

Computer Vision, Large Language Models, Self Driving Vehicles

1. Introduction

The fusion of vision and language models represents a crucial and swiftly advancing area within the realm of artificial intelligence. This convergence holds great importance as it enables machines to grasp and produce content that encompasses both visual and textual data, emulating the natural way in which humans perceive and communicate in reality. Vision-Language models focus on

the simple goal of (*image*, *text*) → *text*.

Often, the real-world data consists of inputs or multiple modalities, such as text and images. RCNNs [1] introduced the world to the field of enabling machines to understand and perceive images. Existing methods typically involve solving tasks independently using a single large vision model. This approach allows the model to have a holistic understanding of the data, leading to more coherent and contextually relevant responses. However, it also increases the complexity and execution time, making the outputs highly dependent on each other.

Our approach suggests a different strategy by combining two dedicated disjoint models to integrate information from both images and text into a single model pipeline. We have named this multimodal integration approach as VOLTRON (Vision Object Linguistic Translation for Responsive Observation and Narration). VOLTRON allows for a more seamless understanding and generation of content that incorporates both modalities effectively. The primary focus of this approach is on improving the way prompts are generated, harnessing the powerful approach to Large Language Model (LLM), ultimately enhancing the quality of the responses it generates.

While approaches like LLaVA (Large Language and Vision Assistant) [2] that use CLIP (Contrastive Language-Image Pre-training) as a visual encoder and LLaMA [3] as the language decoder, and later fine-tune on generated data, apply a more complex foundational architecture. CLIP encodes visual and textual information into a multimodal embedding space while maximizing the cosine similarity scores of the embeddings. However, it performs well only within the training dataset domain and weakly generalizes to out-of-training examples. In response to the specific use case where CLIP falls short, the approach takes into consideration the integration of an object detection model—YoloV8-n [4]. The probability of YoloV8-n's predictions, along with the simplicity of the architecture, becomes a crucial factor for providing suggestions through LLaMA2 [3]. This enhances the multimodal model's ability to handle situations where traditional methods may not work effectively.

We present a novel and simplified model fusion approach without complex architectures for enhancing predictive capabilities in self-driving vehicles. The key research contributions and novel approaches of our work (VOLTRON) are as follows:

- Proposes a new model fusion approach to enhance predictive capabilities of vision and language models by integrating YoloV8-n and LLaMA2
- Utilizes YoloV8-n for object detection to identify small objects like cats crossing the road
- Translates YoloV8-n detection probabilities into sentences to provide contextual information to LLaMA2
- Employs LLaMA2 to generate natural language responses and recommendations
- Fuses models using a simple single layer rather than complex vision encoders

or transformers

- Demonstrates accuracy improvements up to 88.16% on specialized test datasets for self-driving vehicle use cases
- Adopts computational efficiency methods like LoRA, mixed precision training, quantization and batching
- Suggests future vehicular communication protocols for real-time road condition updates between vehicles

2. Related Work

There is growing interest in combining large language models with computer vision across a variety of settings to enhance multimodal understanding and generation. In this section, we provide a study of the related work, categorized into five different areas as follows:

1) Multimodal Instruction-Following: Current research in computer vision pursues two main methodologies by either developing an end-to-end trained models capable of autonomously handling vision-language tasks using instructions, or by creating systems that use multiple single models using frameworks like LangChain [5], as seen in Visual ChatGPT [6] and X-GPT [7]. Additionally, some approaches hybridize these two strategies, focusing on crafting end-to-end trained multimodal models to address various tasks concurrently, as exemplified by methodologies like LLaVA [2].

2) Instruction Tuning: In the realm of NLP, the capability of LLMs to understand natural language instructions and complete real-world tasks has proven to be a significant advancement. Implementations such as Flamingo [8] have borrowed ideas from NLP to Vision, resulting in remarkable performance in zero-shot task transfer and in-context learning. OpenFlamingo [9] and LLaMA-Adapter [10] are open-source implementations that allow LLaMA [3] to process image inputs, thus enabling the development of open-source multimodal LLMs. It is worth noting that these models demonstrate promising task transfer performance, even without explicit instruction tuning using vision-language instruction data.

3) Autoregressive Vision-Language models: Generative vision-language models generate texts based on an image-text sequence, which is a feature of numerous architectures such as BLIP2 [11] and LLaVa [2]. Unlike architectures that are limited to a single image in their context, autoregressive vision-language models embrace interleaved image-text sequences, facilitating contextual learning.

4) VisionLLM: Large Language Model is also an Open-Ended Decoder for Vision-Centric Tasks—It utilizes a large language model as an open-ended decoder for vision-centric tasks. In this context, the language model is not just used for generating text but is employed as a decoder for processing and interpreting visual information. The paper explores how this unique approach can be applied to various tasks in computer vision, making it a versatile tool for tasks like image

captioning, object recognition, and more. It likely discusses the potential benefits and applications of combining language models with visual data for improved performance in vision-related tasks. [12]

5) Drive Like a Human—Rethinking Autonomous Driving with Large Language Models: It explores a new perspective on autonomous driving by leveraging large language models. It suggests that incorporating these models can lead to more human-like and context-aware driving behaviors in autonomous vehicles. The paper likely discusses the integration of natural language understanding and generation to enhance communication between autonomous vehicles and human passengers or other road users. This approach aims to improve the safety and acceptability of autonomous driving technology by making it more understandable and predictable for humans. [13]

3. Statement of the Problem

After reviewing various existing methodologies, we observed that none of them employed an object detection model as a baseline. VOLTRON distinguishes itself by opting for a comprehensive approach without integrating transformers for model fusion, showcasing a unique strength in our strategy.

Existing solutions, such as Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (BLIP2) (Figure 1) and An Open-Source Framework for Training Large Autoregressive Vision-Language Models (OpenFlamingo), delve deeper into the intricacies of image description, striving to provide comprehensive insights. Whereas, OpenFlamingo focuses on creating multiple interleaved, web-scraped datasets, taking into consideration all

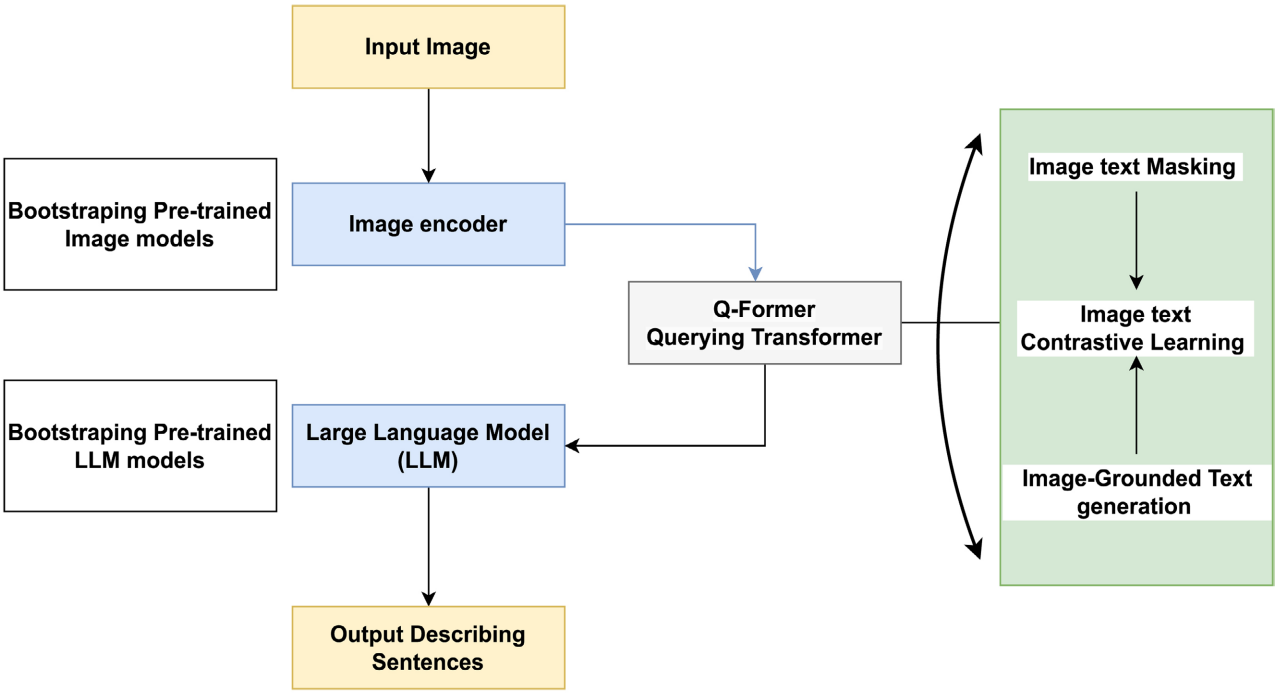


Figure 1. BLIP-2 architecture with model fusion [11].

the previous texts and the last preceding image to increase the performance and the complexity. On the other hand, BLIP2 takes a simpler approach, although it is not very robust with more complex tasks. In contrast, the latter, exemplified by LLaVA (Figure 2) and GPT4, prioritize generating outputs that are influenced by specific instructions and context (See Table 1).

VOLTRON (Figure 3) seeks to position itself at the intersection of these two approaches. We aim to take into consideration the significance of accommodating diverse needs and preferences when it comes to generating descriptions for images. Hence, we adopt a versatile approach that encompasses two fundamental

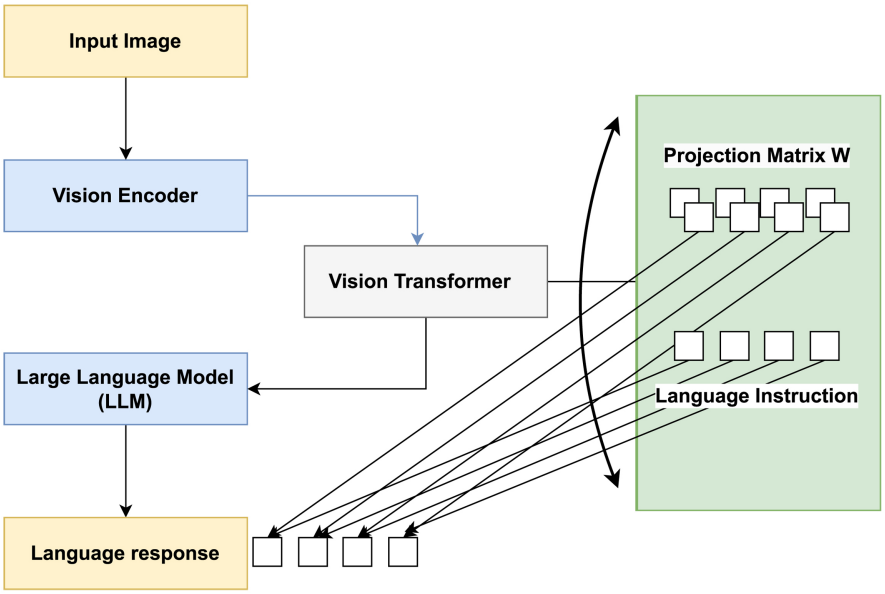


Figure 2. LLaVA architecture with model fusion [2].

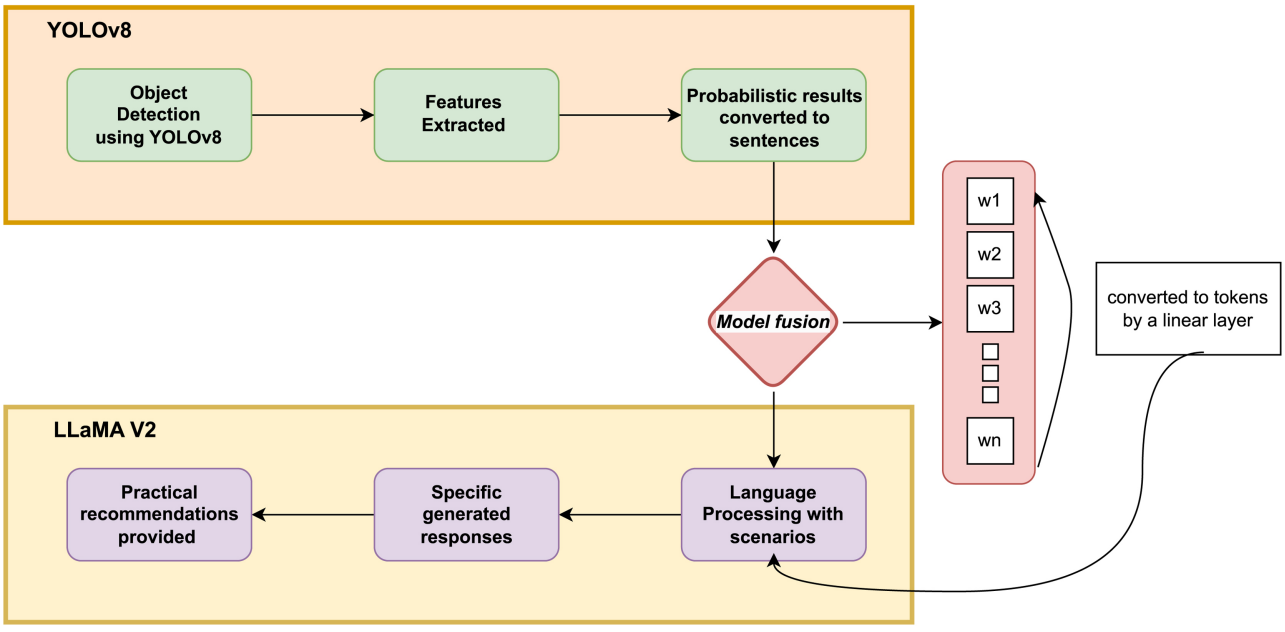


Figure 3. VOLTRON architecture with model fusion.

Table 1. Comparison of approaches.

VOLTRON	LLaVA	BLIP-2
1). VOLTRON combines the strengths of the Pretrained LLM model (LLaMA2) and Vision-object detection Model (YoloV8-n). 2). Integration of the two models into a unified pipeline involves converting object probabilities to sentences, passing them through a single/simple linear layer, and transforming them into embeddings using the LLaMA architecture, thus reducing complexity without the use of transformers.	1). LLaVA's architecture combines LLaMA for language tasks and CLIP visual encoder ViT-L/14 for visual understanding, enhancing multimodal interactions. It can fine-tune LLaMA using machine-generated instruction-following data. For visual content processing, LLaVA relies on the pre-trained CLIP visual encoder ViT-L/14, which excels in visual comprehension. The encoder connects visual features to language embeddings, bridging the gap between text and images. 2). Emphasis on generating instruction-oriented responses.	1). BLIP-2 effectively combines frozen pre-trained image models and language models for outstanding performance on various vision-language tasks. To bridge the modality gap, BLIP-2 employs a Q-Former model pre-trained in two stages: representation learning and generative learning. It extracts a fixed number of output features from the image encoder, regardless of input image resolution. 2). Focus on describing the image, prioritizing image understanding over specific user-generated instructions.

aspects. Firstly, we utilize a much simpler architecture and achieve results with fewer complexities for our use case. Secondly, we facilitate the generation of plain prompts that simply describe the scene depicted in the image. Additionally, we provide prompts to the system that not only describe the scene but also incorporate complex reasoning through in-depth analysis. This dual-pronged approach empowers users to tailor their interactions with our system according to their requirements. By offering this flexibility, we aim to ensure that our approach can be tested and utilized in a broad spectrum of scenarios. Whether one seeks concise scene descriptions or detailed explanations with intricate analysis, our approach can cater to their needs.

4. Methodology

4.1. Architecture

Our aim is to introduce a simplified approach by merging an object detection model and a Large Language Model for particular scenarios. This integration streamlines complex tasks, harnessing the strengths of visual data analysis and natural language understanding to enhance overall performance and usability in specific use cases such as:

1) The presence of small objects or animals crossing the road, potentially cats.
2) The occurrence of narrower or merging lanes ahead, along with the reasons

or events that could have led to in this road configuration, such as a natural disaster like a landslide.

In order to achieve the above, we propose an architecture that aims to enhance predictive performance. While we have tested VOLTRON with the exemplary use cases above, this approach can be utilized and also extended for a broad spectrum of scenarios.

4.1.1. Input Data

The input data involves capturing a video feed and transforming it into a sequence of individual images in a format that can be utilized by YoloV8-n.

4.1.2. Models

- **YoloV8-n** [4], a state-of-the-art object detection model, is utilized to process the visual data. It identifies and localizes objects of interest within the images or frames captured by the vehicle's cameras. Choosing YoloV8-n provides several key benefits over its predecessors. Notably, its anchor-free detection capability enhances flexibility and efficiency, eliminating the need for manual anchor box specifications. This is particularly advantageous when dealing with dynamic objects in varying scenes and perspectives, according to our use case. Additionally, YoloV8-n boasts improved accuracy and efficacy through its optimized network architecture, revised loss function, and modified anchor boxes, promising superior results compared to earlier versions.
- **LLaMA V2** [3] is a specialized language model integrated to handle language and textual data. With variable model sizes available to choose from and an increased amount of training data, as well as a double context length of 4096 tokens, we opted for LLaMA2 with 7 billion parameters. The integration of RLHF (Reinforcement Learning from Human Feedback) during its training phase has significantly improved our results and quality.

4.1.3. Model Fusion

The core of this architectural design centers around the model fusion stage, a crucial point where the outputs of YoloV8-n and LLaMA2 come together. This convergence signifies the beginning of multimodal integration, allowing the model to utilize the combined outputs of both models to make informed predictions and decisions in the realm of self-driving applications.

The YoloV8-n model, primarily focused on object detection, yields outputs in the form of probabilities. These probabilities are subsequently translated into descriptive sentences that comprehensively elucidate the identity of each individual object, along with its corresponding probability score. This transformation enhances the interpretability and contextual richness of the visual data analysis.

Subsequently, this transformed output, now presented in the form of coherent sentences, is transmitted to LLaMA2. In order to join the two models, we have utilized a single layer that provides LLaMA with the tokens instead of implementing Q-Transformers or Vision encoder-decoders. As a language model,

LLaMA processes this information to generate scenario-specific outputs and practical recommendations. By assimilating and interpreting the visual data within the realm of natural language understanding, LLaMA enhances the model's ability to provide relevant insights and real-time suggestions within self-driving scenarios. This integrated approach contributes to more sophisticated and well-informed decision-making processes.

4.2. Dataset and Training

4.2.1. Dataset Generation

We have taken a multifaceted approach for our dataset generation, creating a mixture of artificially generated images as well as publicly available images. Data collection is one of the most important starting points of this project. We made use of the Selenium library to scrape images from the internet. We ran a script to collect the necessary data and store it on our local system. Next, we annotated the data in order to create a better data source to begin with. We have taken into consideration the limiting case of variations in object orientations and lighting conditions with changing illumination intensity. For our experiments, we used a cluster of RTX 4000 Ada GPUs.

4.2.2. LLaMA2 Fine-Tuning

In our approach, we have utilized the capabilities of LLaMA V2. The primary focus of our efforts is dedicated to enhancing the quality of the responses generated by the model and providing valuable feedback to the end user. This endeavor has been thoughtfully segmented into three distinct sections, each with its own unique purpose and methodology.

The first of these sections is **Prompt Engineering**, which leverages the ReAct [14] framework. This approach is designed for rapid iteration and boasts the advantage of not requiring any specific training. It is instrumental in shaping and refining the thoughts, actions, and observations encapsulated in the model's responses.

The second section implements **RAG** (Retrieval Augmented Generation) [15] and serves as a critical component for expanding the external knowledge base. This module enhances the model's ability to query databases and execute fact-checking procedures, thereby increasing the depth and accuracy of the information it can provide.

Figure 4. Some images from our dataset that were used for the evaluation of the model.

The third and final section, referred to as **PEFT** (Parameter-Efficient Fine-Tuning) [16], adds a layer of complexity to our methodology. Unlike the other sections, PEFT requires training and access to a high-quality dataset. Its role is pivotal in enhancing the model's performance by imbuing it with a deeper understanding of nuanced and intricate aspects of the subject matter.

By breaking down our approach into these three sections, we can not only offer users more polished and contextually relevant responses, but also enhance the model's knowledge base and handle more intricate queries and interactions. This comprehensive approach highlights our dedication to pushing the limits of AI capabilities, while also guaranteeing responsiveness, dependability, and precision in user interactions.

To ensure computational efficiency and address the resource demands of inference, we have adopted several strategies. One key approach is LoRA (Low Rank Adaptation) [17], which reduces trainable parameters, saving memory without compromising results. Rank factorization makes LLaMA2 manageable on our constrained and specific setup and use case.

We have also explored mixed precision training to leverage lower-precision tensors for memory-efficient and rapid GPU processing. Additionally, we have implemented quantization to transition model weights from floats to low-bit integer representations. We have chosen PTQ (Post Training Quantization) [18] for cost-effectiveness, as it involves training to convergence and then converting weights without further training. This approach stands in contrast to QAT (Quantization-Aware Training) [19], which requires more computational resources and access to representative training data.

Furthermore, we have improved efficiency by adopting batching [20]. Instead of reloading model parameters for each input sequence, batching allows parameters to load once and serve multiple input sequences, streamlining the process.

5. Qualitative Evaluation

In the context of self-driving vehicles, a critical concern that we observed was the vehicle's inability to detect and respond rapidly to the presence of cats or similar objects rapidly crossing a road. This deficiency in object recognition and response capabilities has resulted in numerous accidents and, tragically, severe fatalities.

To address this issue, we carried out several experiments. Our initial experiment involved creating a specialized dataset comprising images featuring cats, with a particular focus on those exhibiting unique characteristics such as reflections and predominantly black fur. This dataset was then used with pretrained YoloV8-n weights, which is a state-of-the-art object detection model.

The YoloV8-n model was used to perform object detection, providing probabilistic results on the presence of cats within the images. Subsequently, we integrated the YoloV8-n model with LLaMA, which was equipped with a set of sophisticated instructions and scenarios. This integration aimed to not only im-

prove the quality of the response generated by the self-driving vehicle but also to provide valuable feedback that could enhance the vehicle's real-time decision-making process.

To assess our model's performance, we utilized a dataset of 300 images encompassing a diverse range of animals, orientations, colors, and slight variations in noise (See Figure 4 and Figure 5). This varied image set aimed to prevent any biases in the model's outputs. We experimented with different image combinations, calculating their accuracy scores and averaging the results. The accuracy takes into consideration the number of correct outputs generated by the model. For the initial 150 images, the accuracy scores were 88%, 89%, and 87% based on randomly sampled 100 images from this subset, resulting in a mean accuracy of 88%. In the subsequent 150 images, the accuracy scores were 90%, 88%, and 87%, yielding a mean accuracy of 88.33%. To ensure balance across the varied pools of images, we computed the mean again. This approach led to a final mean accuracy of 88% and 88.33%, resulting in an overall mean accuracy of 88.16%. This improvement signifies a significant stride forward in the mission to make self-driving vehicles safer and more proficient at responding to unforeseen road hazards, particularly in scenarios involving pedestrians (See Figure 6). We assess our system's performance using a set of metrics, with a primary focus on accuracy. In this evaluation, we consider the range of models currently available in our environment. This approach enables us to gain a comprehensive understanding of how our system compares to the existing models, which helps us in our pursuit of higher performance and accuracy.

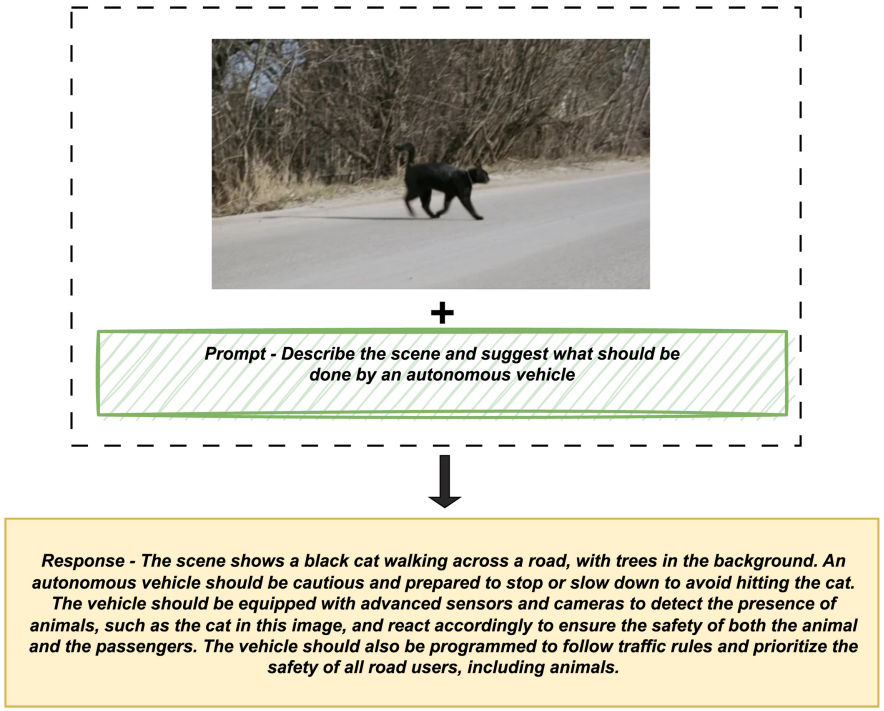


Figure 5. Output generated when a cat is running across the road which is blurred and not very clearly visible (image is open source).

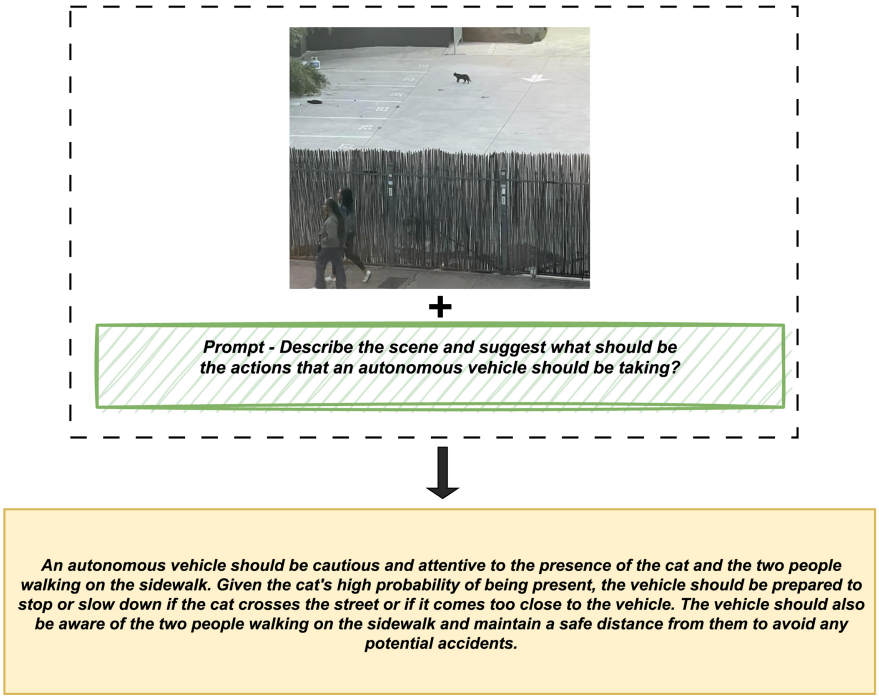


Figure 6. Output generated when the cat is at distance but there are people also walking (image clicked by author).

6. Conclusion

In conclusion, we introduced a unique model fusion method, merging YoloV8-n and LLaMA2, to amplify both vision and language models' predictive prowess. YoloV8-n was used for precise object detection and LLaMAV2 which offered a novel paradigm to enhance object detection and language-based contextual understanding. VOLTRON seamlessly integrates vision-based insights without relying on complex encoders. Instead, it adopts a simplified, single-layer fusion approach, showcasing its efficacy by merging visual and linguistic understanding. Empirical evaluations underscore significant accuracy enhancements in results for self-driving vehicle applications.

7. Future Works

There are several promising ways for extension. First, we can delve into the development of a robust protocol that facilitates seamless communication among a network of vehicles. This communication infrastructure would empower LLMs to engage in meaningful exchanges, enabling them to collectively harness their linguistic capabilities. A fleet of vehicles can collaborate by sharing valuable information among themselves, becoming nodes in a network that operates cohesively to enhance overall safety and efficiency.

A notable aspect of this endeavor is the handling of incidents like natural disasters that may remain concealed from incoming vehicles. To address this, we propose a mechanism where vehicles that have previously encountered such incidents can act as informants. By sharing their experiences, these vehicles ensure

Figure 7. How we perceive the future work to be.

that the arriving vehicle is promptly made aware of the situation. This awareness is conveyed to the driver via a pop-up text message, significantly enhancing the driver's responsiveness. This newfound awareness offers the driver two practical courses of action: the option to reroute and circumvent the incident or to proceed with heightened caution.

In this dynamic system, data remains in constant flow, with updates occurring at regular 15-minute intervals. These updates are not arbitrary; instead, they serve to continuously optimize the performance, ensuring their adaptability and effectiveness in a constantly evolving environment. This comprehensive approach to vehicular communication and information exchange promises to enhance safety, efficiency, and overall road network performance.

Implementing this method could pave the way for a more effective and promising realization of the extended possibilities and future prospects of our proposed approach (See Figure 7).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Xie, X.X., Cheng, G., Wang, J.B., Yao, X.W. and Han, J.W. (2021) Oriented r-cnn for Object Detection.

[2] Liu, H.T., Li, C.Y., Wu, Q.Y. and Lee, Y.J. (2023) Visual Instruction Tuning.

[3] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., *et al.* (2023)

Llama 2: Open Foundation and Fine-Tuned Chat Models.

[4] Jocher, *et al*. (2023) YoloV8. https://github.com/ultralytics/ultralytics

[5] Langchain. https://www.langchain.com/

[6] Wu, C.F., Yin, S.M., Qi, W.Z., Wang, X.D., Tang, Z.C. and Duan, N. (2023) Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models.

[7] Xia, Q.L., Huang, H.Y., Duan, N., Zhang, D.D., Ji, L., Sui, Z.F., Cui, E., Bharti, T., Liu, X. and Zhou, M. (2020) Xgpt: Cross-Modal Generative Pre-Training for Image Captioning.

[8] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., *et al*. (2022) Flamingo: A Visual Language Model for Few-Shot Learning.

[9] Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W.R., Marathe, K., *et al*. (2023) Openflamingo: An Open-Source Framework for Training Large Autoregressive Vision Language Models.

[10] Zhang, R.R., Han, J.M., Liu, C., Gao, P., Zhou, A.J., Hu, X.F., Yan, S.L., Lu, P., Li, H.S. and Qiao, Y. (2023) LLaMA-Adapter: Efficient Fine-Tuning of Language Models with Zero-Init Attention.

[11] Li, J.N., Li, D.X., Savarese, S. and Hoi, S. (2023) Blip-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models.

[12] Wang, W.H., Chen, Z., Chen, X.K., Wu, J.N., Zhu, X.Z., Zeng, G., *et al*. (2023) Visionllm: Large Language Model Is Also an Open-Ended Decoder for Vision-Centric Tasks.

[13] Fu, D.C., Li, X., Wen, L.C., Dou, M., Cai, P.L., Shi, B.T. and Qiao, Y. (2023) Drive like a Human: Rethinking Autonomous Driving with Large Language Models.

[14] Yao, S.Y., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. and Cao, Y. (2023) React: Synergizing Reasoning and Acting in Language Models.

[15] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., *et al*. (2021) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

[16] Liu, H.K., Tam, D., Muqeeth, M., Mohta, J., Huang, T.H., Bansal, M. and Raffel, C. (2022) Few-Shot Parameter-Efficient Fine-Tuning Is Better and Cheaper than In-Context Learning.

[17] Hu, E.J., Shen, Y.L., Wallis, P., Allen-Zhu, Z., Li, Y.Z., Wang, S., Wang, L. and Chen, W.Z. (2021) Lora: Low-Rank Adaptation of Large Language Models.

[18] Zhang, J.J., Zhou, Y.X. and Saab, R. (2023) Post-Training Quantization for Neural Networks with Provable Guarantees. *SIAM Journal on Mathematics of Data Science*, **5**, 373-399. https://doi.org/10.1137/22M1511709

[19] Liu, Z.C., Oguz, B., Zhao, C.S., Chang, E., Stock, P., Mehdad, Y., *et al*. (2023) Llm-qat: Data-Free Quantization Aware Training for Large Language Models.

[20] Cheng, Z.J., Kasai, J. and Yu, T. (2023) Batch Prompting: Efficient Inference with Large Language Model APIs. *Proceedings of the* 2023 *Conference on Empirical Methods in Natural Language Processing*: *Industry Track*, Singapore, December 2023, 792-810. https://doi.org/10.18653/v1/2023.emnlp-industry.74