Runtime Autonomous Driving Context Analysis using Vision Language Model and Lingua Franca

Manuel Branco Nardi, Daniel Avalos, George Zhao, Martina Ibanez Peredo, and Chanhee Lee

University of Central Florida. {ma352154, daniel.avalos, ge615964, martina.ibanezperedo, chanhee.lee}@ucf.edu

I. Introduction

Introducing Vision Language Models (VLMs) to autonomous driving in urban environments has recently attracted significant interest from researchers and industry due to their revolutionary ability to recognize human-like context [1], [2]. Autonomous driving involves various CPS technologies, such as real-time image processing and sensor management, as well as runtime inference and control invocations to drive a vehicle properly. All these executions should be completed within the given time constraints to ensure safe, critical vehicle operations. Existing approaches [2] , however, mainly focus on simulation-based VLM integration without systematically addressing timing constraints, such as deadlines. Therefore, in this demonstration, we propose integrating a VLM and a Lingua Franca (LF) [3] to enable timely analysis of the autonomous driving context. The VLM serves as a driver, periodically analyzing camera images to understand the current driving situation and to decide on high-level actions based on context. LF orchestrates multiple image acquisitions, enforces deadlines, and transfers and invokes action commands.

II. Background

Lingua Franca. Lingua Franca is a coordination language designed to simplify the construction of reliable and predictable cyber-physical systems. Instead of relying on low-level concurrency mechanisms or middleware callbacks, LF structures applications around reactors, which are concurrent components that react to events in a deterministic order. Each reactor encapsulates its internal state, input and output ports, and a set of reactions. LF also supports explicit deadlines, enabling detection and handling of timing violations. These properties make LF particularly suitable for safety applications, where predictable timing and repeatable behavior are essential.

Vision Language Model. Vision–language models (VLMs) extend large language models to handle both images and text within a single unified framework, enabling joint understanding and generation across modalities. They are now a standard foundation for tasks such as image captioning, visual question answering, and multimodal reasoning. Modern vision LLMs usually combine three components: a vision encoder, a language model, and a multimodal fusion module. The vision encoder (often a Vision Transformer) converts an image into a sequence of visual tokens, which are then projected into the token space of a pretrained VLM and jointly processed with text tokens via attention mechanisms.

III. Autonomous Driving Context Analysis

A. Target Scenarios and Inputs

We choose the three images obtainable through camera sensors in vehicles from the NuScenes dataset [4] as shown in Fig. 1. The images depict representative driving scenarios and serve as inputs for autonomous driving context analysis.

Scenario 1: Pedestrian at crosswalk. Our vehicle approaches a crosswalk where a pedestrian is about to cross the road near the curb. This addresses the most critical cases in which inappropriate driving controls cause human injury or death.

Scenario 2: Vehicle crossing onto road. Our vehicle observes another car entering the central driving lane on an urban street. This scenario is one of the most common hazards from another vehicle: another vehicle interrupts our driving, creating a potential collision.

Scenario 3: Stationary vehicle. Our vehicle approaches a multi-lane urban roadway where a large truck is parked along the leftmost curb. People may be unloading particles from the truck. This scenario checks whether the VLM warns of sudden uncertainty and driving risks.

B. LLM Prompt

To obtain the appropriate control results from VLM for input images, we calibrate and use the following prompt.

```
Your task is to analyze the provided images of
road scene and generate a detailed description
that includes the following elements:

1. Describe the overall scene and road context
in detail.
2. Identify all safety-relevant actors
(e.g., vehicles, pedestrians, traffic signs).
3. Point out any potential hazards, conflicts,
or situations that would require caution from
a driving perspective.
4. If motion or intent is ambiguous,
clearly state the uncertainty and explain why.
5. List two possible actions a driver
could take in response to the scene
based on the description from 1.-4.,
along with the reasoning behind each action.

Do not assume intent or future actions beyond
what is visible in the images. Base your
reasoning strictly on visual evidence.
```

C. Lingua Franca Implementation

There are three reactors in our LF implementation of autonomous driving context analysis, as shown in Fig. 2.

(a) Scenario 1: Pedestrian at crosswalk.

(b) Scenario 2: Vehicle crossing onto road.

(c) Scenario 3: Stationary vehicle

Fig. 1: Input images from NuScenes datasets for autonomous driving context analysis.

Fig. 2: The LF diagram of the proposed autonomous driving context analysis.

CameraSensor Reactor. This reactor periodically captures image frames and forwards them to the next reactor, which runs a VLM with the prompt described in Section III-B.

VLMInference Reactor. This reactor receives image frames, performs VLM inference, and returns results to the `ActionDisplay` reactor. We set the specified deadline based on the target VLM performance (latency) to ensure VLM inferences are completed within the time constraint.

ActionDisplay Reactor. This reactor receives VLM responses and displays recommended actions to drivers. Moreover, it parses actions and invokes relevant physical operations on vehicle hardware in emergent situations, e.g., sending signals to operate the brake pedals when no driver action is detected and a possible collision with the front vehicle is imminent.

IV. Experimental Results

Setup. We run our experiments on a desktop with Intel Core i5-9300H, 32GB of RAM, and an NVIDIA GeForce GTX 1650 with 4GB of VRAM as the CPU, DRAM, and GPU. We use two VLMs, `vikhyatk/moondream2` and `Qwen/Qwen2.5-VL-7B-Instruct-Q4`, from Hugging Face , with model sizes of 3.5GB/1.8B and 6GB/7B.

Context Analysis Results. In all target scenarios, VLMs produce similar and appropriate descriptions for each input image context and recommended actions (RA) to operate the current vehicle as follows:

RA of Scenario 1: Stay Ahead. The driver should maintain a safe distance from other vehicles and pedestrians to avoid obstructing traffic flow or causing accidents.

RA of Scenario 2: Continue straight. There is a risk of a collision from an incoming car, so the driver should be cautious about continuing straight without noticing any unusual activity or vehicles.

RA of Scenario 3: Continue on. The road continues ahead, and drivers should maintain a safe distance from other vehicles and traffic signals. The presence of construction materials and barriers suggests ongoing work or traffic management.

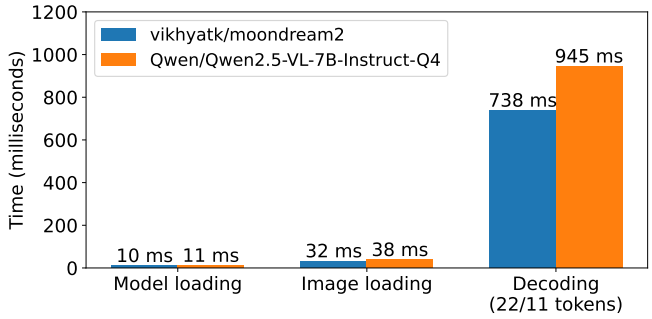

Fig. 3: The latencies of VLM loading, image loading, and decoding in two VLMs.

Latencies. We measure the runtime performance of our analysis using the latencies of three key operations: image loading, VLM prefill, and VLM decoding. Note that VLM inference consists of one prefill stage followed by repeated decoding stages. The VLM loading times are 14.8 and 28.7 seconds, respectively. The results of latencies in the two VLMs are shown in Fig. 3. The latency of the Qwen2.5 model is 19-143% higher than that of the moondream2 model. This is because the Qwen2.5 model size is 1.7 times larger.

V. Conclusion and Future Work

We integrate VLMs with the LF model to ensure timely inference for autonomous driving context analysis. The results show that our design performs acceptably. As future work, we will apply a wider range of scenarios to various VLMs. In addition, we extend the current LF model to include reactors that control vehicle hardware, such as lights and brakes.

References

[1] R. Zhao, Q. Yuan, J. Li, Z. Wang, Y. Li, Z. Gao, H. Hu, and F. Gao, "Vlm-driver: Human-like autonomous driving decision-making via vision language model," *IEEE Transactions on Vehicular Technology*, 2025.

[2] H. Tian, K. Reddy, Y. Feng, M. Quddus, Y. Demiris, and P. Angeloudis, "Large (vision) language models for autonomous vehicles: Current trends and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 27, no. 1, pp. 187–210, 2025.

[3] Lingua Franca Project, "Lingua franca documentation," https://www.lf-lang.org/docs/.

[4] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 621–11 631.