STM JOURNALS

Journal of
Control & Instrumentation

https://journals.stmjournals.com/joci

ISSN: 2229-6972 (Online)
ISSN: 2347-7237 (Print)
Volume 16, Issue 1, 2025
January–April
DOI (Journal): 10.37591/JoCI

Review JoCI

Multimodal Generative AI for Vehicular Applications at Edge

Sundaresan Poovalingam[1,*], Bhoomi Shah[2],
Rani Malhotra[3], Nikhil Nandanwar[3]

Abstract

This paper explores how the rapid advancements in vehicular technology, including autonomous driving and intelligent transportation systems, have driven the need for real-time data processing and decision-making. Multimodal generative AI, when deployed at the edge, offers a powerful solution for vehicular applications by leveraging diverse data sources such as video, audio, sensor inputs, and environmental data. This paper explores the integration of multimodal generative AI with edge computing in vehicular networks, particularly for use cases like autonomous platooning, predictive maintenance, anomaly detection, and enhanced security. By processing data locally at the edge using AI chips, vehicles can respond instantly to critical events without relying on distant cloud servers, reducing latency and improving safety. Furthermore, vehicular ad hoc networks (VANETs) play a crucial role in supporting decentralized, low-latency communication between vehicles and infrastructure. This fusion of multimodal AI and edge computing unlocks the potential for more intelligent, efficient, and resilient vehicular systems, paving the way for next-generation transportation and smart city infrastructures.

Keywords: VANETs, AI chips, Platooning, multimodal AI, networks, traffic conditions

INTRODUCTION

As intelligent transportation systems continue to advance, integrating various communication categories is expected to significantly enhance the safety, efficiency, and convenience of modern transportation systems. By enabling vehicles to communicate with one another, as well as with infrastructure, networks, and individuals, we can develop smarter, more responsive urban environments that greatly improve mobility and quality of life [1–3]. This interconnectedness serves as a foundation for autonomous vehicles and smart city initiatives, paving the way for a sustainable and efficient transportation future.

Vehicle communication is essential for unlocking the full potential of intelligent transportation systems, as shown in Figure 1. As vehicular networks have become more sophisticated, multiple communication categories have emerged to support different objectives. Examples of these are as follows.

- *Vehicle-to-Vehicle (V2V):* communication between vehicles, sharing information, such as speed, position, and direction.
- *Vehicle-to-Infrastructure (V2I):* Communication between vehicles and road infrastructure, such as traffic signals, road signs, or toll gates.
- *Vehicle-to-Pedestrian (V2P):* Communicates with pedestrians and cyclists, typically via mobile devices or wearables.

***Author for Correspondence**
Sundaresan Poovalingam
E-mail: sundaresan_p@infosys.com

[1]Distinguished Technologist, Advanced Engineering Group, Infosys Limited, Bengaluru, Karnataka, India
[2]Constultant, Advanced Engineering Group, Infosys Limited, Bengaluru, Karnataka, India
[3]Lead Constultant, Advanced Engineering Group, Infosys Limited, Bengaluru, Karnataka, India

Receiving Date: December 09, 2024
Accepted Date: December 14, 2024
Published Date: December 27, 2024

Citation: Sundaresan Poovalingam, Bhoomi Shah, Rani Malhotra, Nikhil Nandanwar. Multimodal Generative AI for Vehicular Applications at Edge. Journal of Control & Instrumentation. 2025; 16(1): 27–34p.

- *Vehicle-to-Network (V2N):* Vehicles connect to cloud networks to access real-time data, such as traffic conditions, weather updates, or entertainment services.
- *Vehicle-to-Manufacturer (V2M):* Communication between vehicles and manufacturers for maintenance, software updates, and performance monitoring.
- *Vehicle-to-Vendor (V2Vdr):* Communication with service providers or vendors, such as fuel stations, repair shops, parking facilities, and retail stores.

PLATOONING USING V2V COMMUNICATION

An emerging area of V2X communication is platooning using V2V. Platooning, the concept of autonomously controlled vehicles traveling in close formations, aims to optimize fuel efficiency, enhance safety, and reduce traffic congestion. Platooning relies on tight coordination between vehicles, requiring real-time exchange of information related to speed, position, and the external environment, as shown in Figure 2. Current platooning systems often rely heavily on single-mode communication, such as radar or LiDAR, to detect the distance between vehicles. However, this approach can be limited to complex scenarios, such as

- Sudden weather changes (e.g., fog and rain) affect the visibility and sensor performance.
- Dynamic road conditions like traffic patterns, construction zones, or accidents.
- Mixed-mode traffic environments where platoons must adapt to human-driven vehicles.

By combining various data streams, multimodal generative AI [4] can overcome these limitations by creating a holistic understanding of the environment and providing enhanced decision-making for safe adaptive platooning, as shown in Figures 3 and 4.

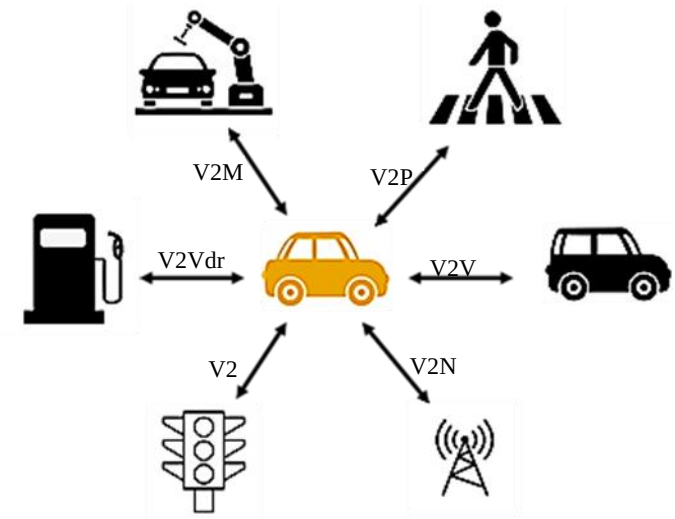

Figure 1. Vehicle-to-everything (V2X) models.

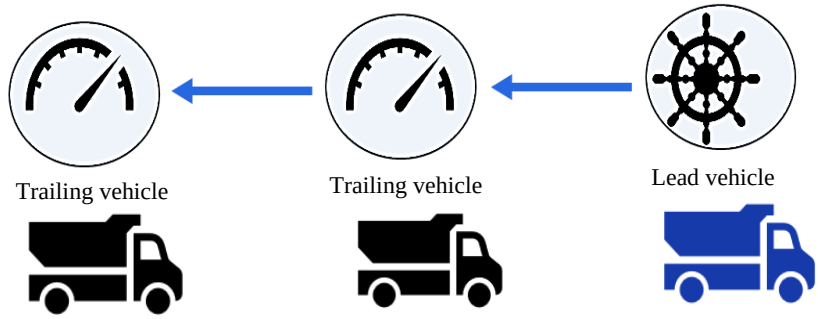

Figure 2. Truck platooning illustrations.

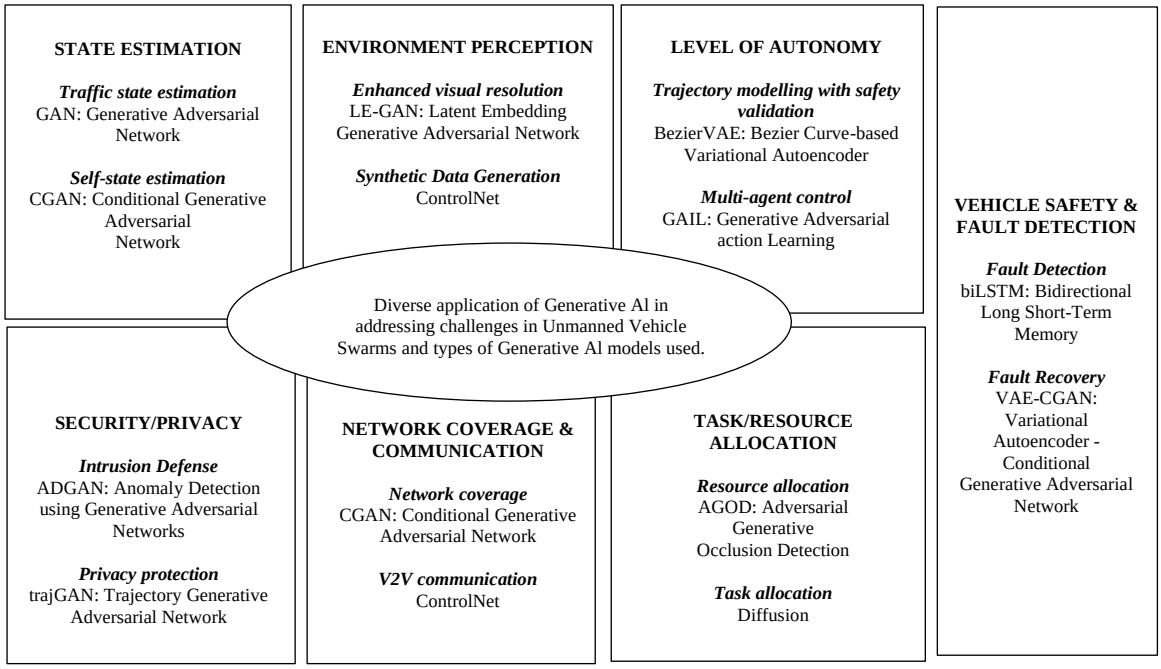

Figure 3. Types of generative AI models used for different applications.

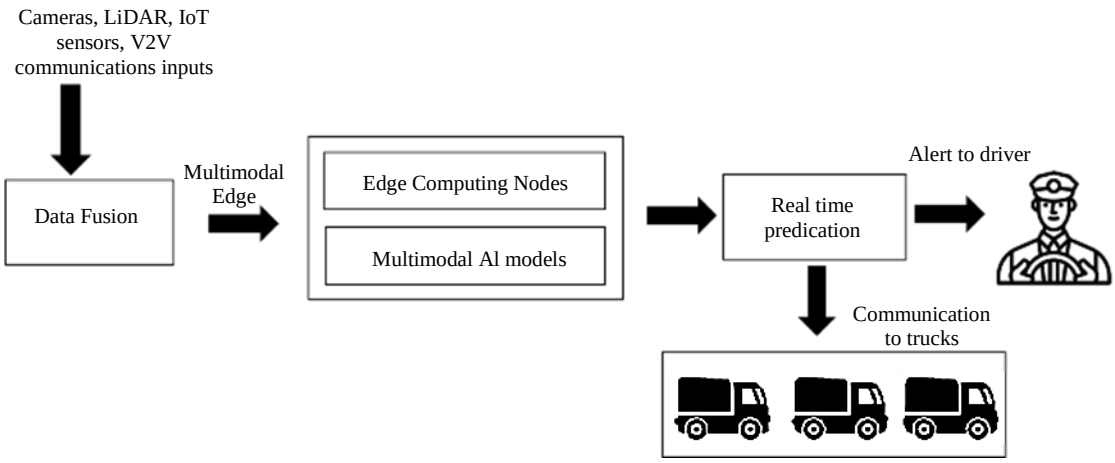

Figure 4. V2V communication process in a platoon.

COMPONENTS

Data Fusion and Processing

The data fusion layer of the system aggregates inputs from cameras, LiDAR, radar, vehicle-to-vehicle (V2V) communications, and IoT sensors. This multimodal data stream is fed into a deep neural network that generates real-time decisions for platoon control.

Real-Time Decision-Making with AI

Once multimodal data are aggregated, generative AI models process this information to make decisions in real time. AI uses machine learning algorithms to predict potential road hazards, determine optimal speed adjustments, and determine the proper distance between vehicles. For example, if AI detects sudden braking in a leading vehicle or an obstacle on the road, it can instantaneously generate commands to the following vehicles to slow down, speed up, or change lanes. The data fusion layer ensures that the visual, sensor, and communication data are integrated to minimize blind spots and enhance situational awareness.

Edge Computing for Real-Time Execution

Edge computing, in which data processing occurs closer to the source of data, is a crucial component of platooning systems. By leveraging multimodal generative AI at the edge, vehicles can make real-time decisions and respond quickly to changing conditions, thereby enhancing their safety and efficiency. In addition, to minimize latency, edge computing nodes were installed in each vehicle. These edge nodes process multimodal data locally and execute immediate decisions, such as braking, acceleration, or lane changes, in near-real time.

Communication Between Vehicles

Vehicle-to-vehicle (V2V) communication plays a critical role in coordinating the actions of an entire platoon. Each vehicle continuously broadcasts its status, including speed, braking, and position, to the other vehicles in the formation.

The multimodal generative AI system uses this information to synchronize the movements of all the vehicles in the platoon. For example, if the lead vehicle accelerates or brakes, AI ensures that the entire platoon follows suit in a coordinated manner. V2X communication also connects the platoon to the smart infrastructure, enabling vehicles to receive alerts about upcoming traffic signals, road hazards, or optimal routes. An emerging technology that enables communication between vehicles is vehicular ad hoc networks (VANETs) [3], a type of mobile ad hoc network (MANET) that enables vehicles to communicate with each other (Vehicle-to-Vehicle or V2V) and with roadside infrastructure (V2I) in real time.

Such networks are crucial for enhancing road safety, improving traffic management, and supporting autonomous driving. VANETs are used in applications such as collision avoidance, traffic flow optimization, and providing infotainment services to drivers. By enabling rapid data exchange, they play a key role in connected vehicle technologies and intelligent transportation systems (ITS).

Human Interaction and Overrides

In semi-autonomous platooning systems, human drivers may still play a role in vehicle control. Multimodal generative AI ensures seamless interaction between an AI system and human drivers by integrating natural language processing (NLP) and monitoring driver behaviors [5].

The system can issue warnings or suggestions to human drivers, and drivers can provide commands or manually override the AI, if needed. This human-vehicle interaction is key to ensuring trust and safety during the transition to fully autonomous platoons.

Continuous Learning and Improvement

Multimodal generative AI systems use machine learning and reinforcement learning to continuously improve decision-making over time. By collecting data from multiple platooning scenarios (such as varying weather, road types, and traffic conditions), AI learns to handle increasingly complex driving environments.

The system was trained to optimize parameters such as fuel efficiency, safety, and traffic flow based on real-world conditions [6–9]. Periodic model updates from the cloud help ensure that AI always uses the latest algorithms, benefiting from global data across different platoons.

HOW GENERATIVE AI ON THE EDGE HELPS IN PLATOONING

In a platoon, vehicles must constantly communicate and respond to each other's movements, which demands low-latency and high-bandwidth data transfer to maintain safety and efficiency, as shown in Figure 5. Edge computing reduces dependence on distant cloud servers and enhances reliability by processing data locally, even in areas with limited network coverage, which is essential for the smooth functioning of autonomous or semi-autonomous vehicle platooning [2, 10].

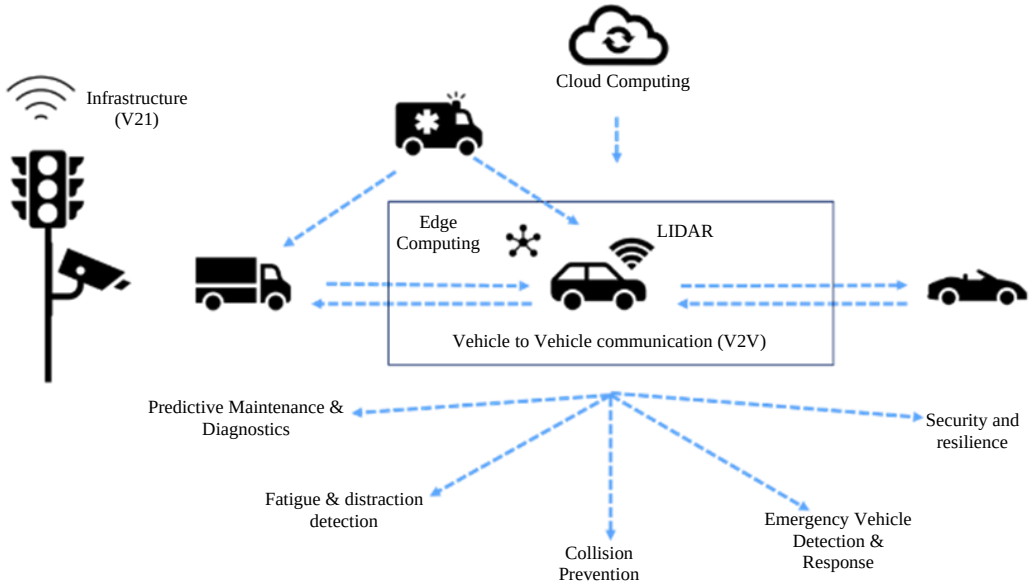

Figure 5. Generative AI on the edge applications in platooning.

Edge computing help with certain applications like predictive maintenance & diagnostics as well as on field repair instructions for when vehicles break down; collision prevention by processing data from multiple sensors and multimodal generative AI which fuses these data inputs to create real-time situational awareness, detecting obstacles and predicting potential collisions; fatigue and distraction detection where edge computing can instantly process the data and detect signs of drowsiness, inattention, or cognitive overload, and provide immediate alerts or assistance to the driver; driver assistance and training by utilizing multimodal generative AI at edge to analyze driver behavior and can then deliver personalized guidance and support, helping drivers enhance their skills and adapt to the unique dynamics of platooning; emergency vehicle detection and response utilizing generative AI and the vehicle sensors to detects the distinct sound or visual signals of an approaching emergency vehicle, such as sirens or flashing lights, then the generative AI algorithms can analyze the urgency and direction of the emergency vehicle, allowing the platoon to make real-time decisions, such as automatically creating space or adjusting speeds to facilitate a clear path for the ambulance; enhanced security and resilience where generative AI models are used to detect and mitigate security threats and edge has many more such applications in platooning. In this section, we discuss in detail how multimodal generative AI at the edge is used in platooning predictive maintenance and diagnostics, along with enhanced security and resilience [11].

Predictive Maintenance & Diagnostics

Edge computing in platooning enables predictive maintenance by processing data from multiple vehicle sensors, such as engine performance, brake wear, and tire conditions, in real-time using multimodal generative AI. By analyzing these data locally, vehicles can detect anomalies and predict potential failures before they occur, reducing downtime and preventing costly breakdowns. Edge-based diagnostics allows each vehicle to assess its health independently, share critical information with the platoon, and optimize maintenance schedules based on real-time operational data. In addition, the edge deployment of predictive maintenance models enables proactive scheduling of repairs based on real-world usage patterns of the platoon fleet for optimal uptime.

An offline Retrieval Augmented Generation (RAG) pipeline can be used to create a system wherein the diagnostic logs generated by the vehicle can be fully analyzed on the edge, and the recommended maintenance procedure and timelines indicated to the user. This can be achieved in two parts. First, we train an SLM such as Llama 3.1 on diagnostic logs and associated defects or issues with the vehicle. Second, we train another SLM on the maintenance procedure documents associated with these defects,

as shown in Figure 6. The output of SLM 1 is the input of SLM 2. If the proposed repair procedure(s) are identified as solvable by a non-expert, the SLM can generate step-by-step instructions that can be conveyed to the vehicle owner to perform the same.

Retrieval augmented generation (RAG) is an AI framework that combines LLMs with external knowledge from specific sources. RAG allows AI models to gain industry-specific knowledge safely and efficiently, as shown in Figure 7. It embeds the knowledge of external data sources into "chunks" that allow the AI model to provide accurate information on a focused, specialized topic. Chunking embeds text chunks with additional information to link contexts.

ENHANCED SECURITY AND RESILIENCE

Innovative solutions have been deployed in the ever-evolving field of cybersecurity to counteract malicious attacks. One such method is the honeypot system [1], which serves as a decoy system designed to detect, deflect, or study potential cyberattacks that target vehicle networks. In the context of platooning, in which multiple autonomous or semi-autonomous vehicles communicate and coordinate closely, a honeypot can simulate vulnerabilities or weaknesses in vehicle-to-vehicle (V2V) or vehicle-to-infrastructure (V2I) communication systems. Cyber attackers attempting to infiltrate the platoon's network would interact with the honeypot, allowing security systems to monitor their behavior, gather intelligence, and develop defenses before they can compromise the actual platoon vehicles.

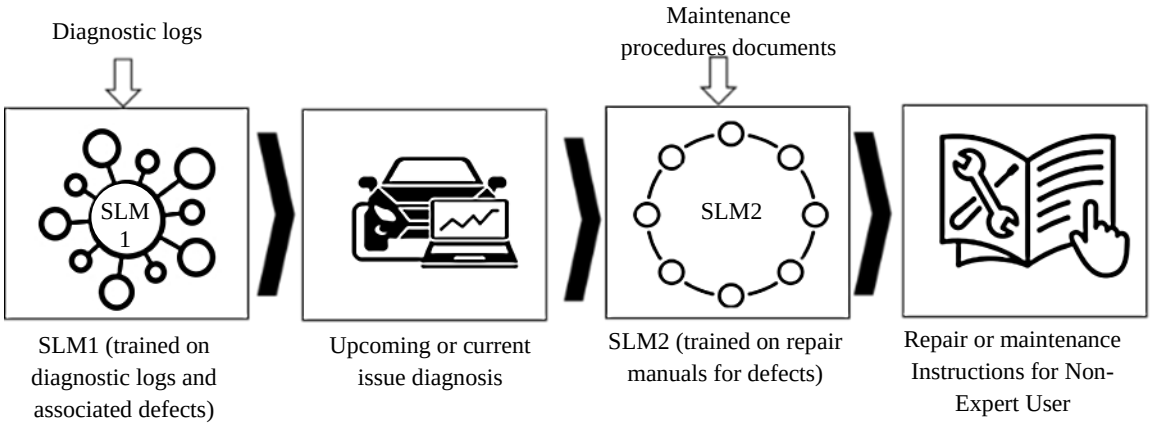

Figure 6. GenAI model pipeline for defect classification and prescription of repair/maintenance instructions.

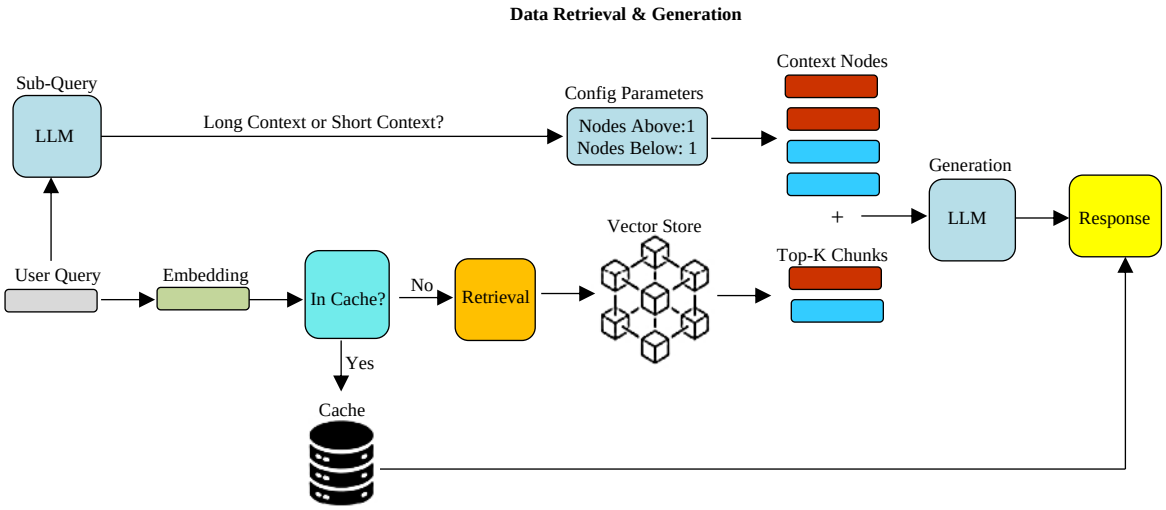

Figure 7. Generic RAG pipeline.

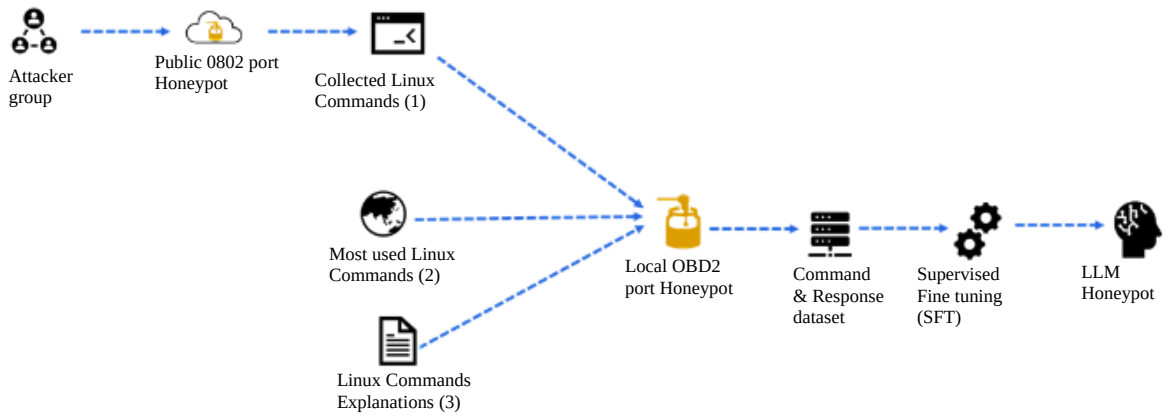

Figure 8. System of bait vehicle(s) incorporated with honeypots.

They can gather intelligence about insights into the types of attacks (e.g., DDoS, spoofing, man-in-the-middle) targeting the platoon's network, revealing techniques, tactics, and procedures (TTPs) used by adversaries; Intrusion Vectors of how attackers gain access to the network—whether by exploiting communication protocols (V2V, V2I), unsecured IoT devices, or other methods; identifying the origin of attacks, including IP addresses, regions, or countries from which the malicious activity is launched; how attackers interact with the network, including what data they are interested in, how they attempt lateral movement, and their end goals (disruption, data theft, etc.); information on when attacks occur and how frequently, helping to predict future attacks and adjust security defenses accordingly. This proactive approach strengthens the platoon's overall cybersecurity by identifying potential threats in real-time.

How: Multimodal generative AI can be used to detect and mitigate security threats such as spoofing or jamming attacks. Vehicular honeypots can be simulated using LLMs to generate synthetic vehicular data. When an attacker connects to the honeypot vehicle, say via an SSH (Secure Shell) connection to the On-Board Diagnostics (OBD) 2 port, the LLM responds with synthetic data mimicking the behavior of a real vehicle, as shown in Figure 8. The logs generated via these honeypot attacks can be utilized to identify the types of attacks that are possible, and accordingly, incorporate additional cybersecurity safeguards.

CONCLUSION

In conclusion, with new emerging technologies such as VANETS and AI accelerators such as Intel Automotive Software Defined Vehicle (SDV) architecture, which enhances edge data processing and multimodal generative AI capabilities, we will be able to see all these applications running on vehicles and platoons in the near future.

Predictive maintenance and enhanced security/resilience are two key applications that can benefit greatly from the capabilities of multimodal generative AI and edge computing for vehicle platooning. Overall, the edge and generative AI's unified processing of live vehicular data streams in a disaggregated manner endows predictive and protective abilities beyond any single system.

Their advanced pattern recognition and coordination skills across a platoon provide inherent redundancy against cyberattacks and hardware failure. Being tightly integrated with fleets, this autonomous infrastructure can self-update to evolving threats. Leveraging such multilayered cognition and computing resources at the network edge is thus pivotal for introducing predictive quality assurance, as well as active safeguards against disruptions in connected and automated commercial vehicle operations such as platooning. This ultimately helps to maximize safety, utility, and commercial viability on public roads.

REFERENCES

1. Liu G, Van Huynh N, Du H, Hoang DT, Niyato D, Zhu K, et al. Generative AI for unmanned vehicle swarms: Challenges, applications and opportunities. [Preprint]. arXiv. 2024 Feb 28. Available from: https://doi.org/10.48550/arXiv.2402.18062.
2. Xie G, Xie R, Zhang X, Nie J, Tang Q, Lim WYB, et al. GIoV: Achieving generative AI services in Internet of Vehicles via collaborative edge intelligence. 2024 IEEE Wireless Communications and Networking Conference (WCNC), Dubai, United Arab Emirates. 2024. p. 1–6. doi:10.1109/WCNC57260.2024.10571334.
3. Ameur AI, Lakas A, Yagoubi MB, Oubbati OS. Peer-to-peer overlay techniques for vehicular ad hoc networks: Survey and challenges. Veh Commun. 2022;34:100455. doi:10.1016/j.vehcom.2022.100455.
4. Marks N. (2024). In It for the Long Haul: Waabi Pioneers Generative AI to Unleash Fully Driverless Autonomous Trucking. [online] NVIDIA Blog. Available from: https://blogs.nvidia.com/blog/waabi-autonomous-trucking/.
5. Intel. (2024). Software-defined vehicle transformation starts with Intel. [Online]. Available from: https://download.intel.com/newsroom/2024/automotive/Intel-SDV-Demo-Fact-Sheet.pdf
6. Stappen L, Dillmann J, Striegel S, Vögel HJ, Flores-Herr N, Schuller BW. Integrating generative artificial intelligence in intelligent vehicle systems. 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain. 2023. p. 5790–7. doi:10.1109/ITSC57777.2023.10422003.
7. Xu M, Niyato D, Chen J, Zhang H, Kang J, Xiong Z, et al. Generative AI-empowered simulation for autonomous driving in vehicular mixed reality metaverses. IEEE J Sel Top Signal Process. 2023;17:1064–79. doi:10.1109/JSTSP.2023.3293650.
8. Xu M, Niyato D, Zhang H, Kang J, Xiong Z, Mao S, et al. Joint foundation model caching and inference of generative AI services for edge intelligence. GLOBECOM 2023 - 2023 IEEE Global Communications Conference, Kuala Lumpur, Malaysia. 2023. p. 3548–53. doi:10.1109/GLOBECOM54140.2023.10436771.
9. Desai B, Patil K. Secure and scalable multi-modal vehicle systems: A cloud-based framework for real-time LLM-driven interactions. Innov Comput Sci J. 2023;9:1–1.
10. World Economic Forum. Autonomous Trucks: An Opportunity to Make Road Freight Safer, Cleaner and More Efficient. Cologny/Geneva: World Economic Forum; 2021. Available from: https://www.weforum.org/publications/autonomous-trucks-an-opportunity-to-make-road-freight-safer-cleaner-and-more-efficient/
11. Ivanovic B, Leung K, Schmerling E, Pavone M. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. IEEE Robot Autom Lett. 2021;6:295–302. doi:10.1109/LRA.2020.3043163.

