Multimodal Agentic Model Predictive Control

Saptarashmi Bandyopadhyay
Department of Computer Science
University of Maryland
College Park, USA
saptab1@umd.edu

John (Jack) Cole
Department of Computer Science
University of Maryland
College Park, USA
jackcole@umd.edu

Tom Goldstein
Department of Computer Science
University of Maryland
College Park, USA
tomg@umd.edu

David Jacobs
Department of Computer Science
University of Maryland
College Park, USA
dwj@umd.edu

ABSTRACT

Control problems for autonomous AI agents, especially safety-critical applications such as autonomous vehicle control, require robust decision-making frameworks to ensure safe navigation in such complex and dynamic environments. This necessitates approaches such as Agentic Model Predictive Control (MPC), which can anticipate future problems and plan for them accordingly. Recently, Multimodal Vision Language Models (VLMs), have emerged as a way to give a semantic meaning to a scene that draws on extremely large amounts of information and contextual understanding of the world. These models vary in a wide range of sizes, trading off speed with performance as they scale further and further. This paper introduces a novel framework that integrates MPC with Multimodal VLMs in order to enhance the ability of autonomous vehicles to navigate and respond to real-world scenarios. Leveraging the open-source Waymax library released by Waymo, along with Waymo Open Motion, Berkeley DeepDrive and NuScenes Datasets, our method uses Multimodal VLMs to detect and draw bounding boxes around important parts of the scene, such as pedestrians or other vehicles. These models are helpful for querying specific attributes of identified objects, such as telling if a vehicle is accelerating or decelerating, or by recognizing if a newly detected obstacle is on a collision course with the vehicle. By incorporating these and other semantic insights into an MPC framework, an autonomous vehicle can make more informed and more context aware decisions to mitigate the risk of a collision and safely navigate its surroundings. We evaluate our approach in diverse simulated environments using VLMs of different scales, demonstrating improvements in safety metrics compared to traditional MPC methods. The integration of VLMs with MPC represents a significant advancement in autonomous decision-making, and especially in dynamic and uncertain situations. Our approach paves the way for future research in using Multimodal VLMs for more intelligent and adaptable autonomous agents.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

KEYWORDS

Model-Predictive Control, AI Agents, Multimodal (Vision-Language), Autonomous Cars, Safe Navigation

ACM Reference Format:
Saptarashmi Bandyopadhyay, John (Jack) Cole, Tom Goldstein, and David Jacobs. 2025. Multimodal Agentic Model Predictive Control. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 5 pages.

1 INTRODUCTION

Autonomous vehicle control is a long-standing and most grand challenge in robotic reasoning and planning. Driving, for example, is feasible for most people given sufficient training and instruction, but is an extraordinarily complex task for a computer to accomplish [5]. There are countless data to take into account when deciding which action to take, and a severe departure from a planned course of action may be required at any moment. For this reason, reliable autonomous vehicles have remained elusive despite significant advances in Artificial Intelligence (AI) research.

One method used to implement planning in autonomous vehicles is Model Predictive Control (MPC). In MPC, the system has an explicit model of its own evolution, allowing it to anticipate future states and plan accordingly, adjusting as needed to any new information it receives. It works even in a complex multivariate environment, such as autonomous driving, where route tracking and vehicle movement must be taken into account [1]. MPC is a powerful and efficient framework used not only in autonomous vehicle control but in a wide variety of safety-critical applications such as chemical synthesis and air traffic control.

Vision Language Models (VLMs) have emerged as a popular way to create AI systems capable of taking in different modalities of data similar to how a human would [8]. A VLM can ingest text, speech, images, video, and a host of other data to generate an output. Due to the vast amount of data used to train these models, VLMs are often used to give an AI a general world understanding, though they can also be fine-tuned for a particular task. The capabilities of a VLM scale considerably with the number of parameters, but an increase in size also comes with considerable latency, creating a tradeoff between performance and speed when it comes to model scale. For this reason, many VLMs often come packaged in a "family", leaving it to the user to select which one best fits their use case.

In this paper, we propose combining MPC and VLMs to create autonomous vehicle agents that have a semantic understanding of the world. We believe that this will lead to agents that will:

(1) Have more context-aware decision-making and improved safety in complex and dynamic situations
(2) Be able address real-time decision making problems with powerful AI capabilities
(3) Leverage knowledge from pre-training that can handle unforeseen circumstances better than traditional Autonomous Vehicle Control
(4) Have superior interactions with humans by using language-based signals

Our position is that an approach such as this is not only possible but can be accomplished using open-source data such as the Waymo Open Motion Dataset [7] and the NuScenes Autonomous Driving dataset [4], as well as open-source simulators such as Waymax [9]. Our Multimodal Agentic Model Predictive Control gives us fine-grained control which can be useful for the AI Agent adapting to mistakes and navigating efficiently while training or finetuning self-driving vehicles.

2 BACKGROUND

AI Agents are systems where an Artificial Intelligence is endowed with the ability to observe an environment, choose an action, an action, and execute that action within the environment. The most important feature distinguishing AI Agents from traditional AI is that the AI Agent needs to be able to consider how its own behavior affects the world around it, rather than simply focusing on giving the "correct" output at any given moment. For this reason, we believe Autonomous Vehicles, and particularly Self-Driving Cars, fit neatly into the category of AI Agents.

Autonomous Vehicle Development has been an active area of research since as far back as the 1980s [6]. While rule-based approaches were originally the norm for these systems, the advent of practical deep learning led to the quick adoption of advanced learning-based approaches that could capture patterns in data that could not be explicitly modeled by a set of human-written rules [14]. Self-driving car companies such as Waymo, Kodiak, and Wayve have released blogposts with limited detail on using VLMs for Autonomous Vehicles, but to our knowledge they have not combined VLMs with classical control mechanisms like MPC as we are proposing.

MPC is a paradigm that has been adopted both in rule-based approaches to artificial intelligence as well as learning-based approaches. Due to its ability to anticipate future scenarios and plan in dynamic environments, it is well suited for an environment as sophisticated as that of a driver operating a vehicle [15]. Another advantage of MPC is its interpretability, being that they:

(1) Pptimize a cost function that has a physically-grounded meaning
(2) Make derivations entirely according to the constraints and objectives of the problem
(3) Have predictions that can be readily analyzed

Recent research has also proposed merging MPC with other methods of control such as Stanley-based control [1].

VLMs have increasingly been used to take actions, with Vision-Language-Action (VLA) models emerging to address this use-case [10, 18]. Specifically, VLMs and controllers have been used in conjunction with one another before [3, 16]. In fact, VLMs have been used to serve as controllers in and of themselves [13, 17]. Previous work has looked at using MPC and VLMs in combination with one another for Self-Driving Cars [12], but we seek to build a frameowrk that can serve as a basis for other Autonomous Vehicle paradigms as well, including Autonomous Drones and Autonomous Aquatic Vehicles. With our proposed framework we hope to further expand the horizon of what VLMs can be used for and improve the state-of-the-art for Autonomous Vehicle Control by introducing these new capabilities.

3 GENERAL APPROACH: MPC-VLMS FOR AUTONOMOUS VEHICLES

3.1 Conceptual Framework

Our proposed framework for an enhanced Autonomous Vehicle Control is centered on the potential for improving performance by leveraging the semantic understanding of the world embedded in a VLM with MPC that can safely anticipate and plan for future challenges faced by the autonomous vehicle using information percieved by the VLM. At each timestep, the autonomous vehicle would have information input to its sensors. This input would be fed into the context of the VLM, along with a sufficient history of prior information. The system would then query the VLM for information regarding the scene, such as whether there is a car in the scene or not, as well as changes in the scene, such as whether the car appears to be in the way of a pedestrian. We would have a VLM of a size that the Autonomous Vehicle is capable of safely using given resource and time constraints, as the vehicle must be able to quickly make decisions. A key advantage of this approach is that it gives humans a way to understand how the vehicle is planning for future obstacles, accomplishing our objective of a safe and trustworthy autonomous vehicle controller. An overview of the general framework we propose is illustrated in Figure 1.

3.2 Scene Understanding using VLMs

While traditional computer vision techniques analyze a scene using statistical correlation, we believe VLMs are capable of a higher-level understanding of contextual relationships [19]. VLMs have previously been used for scene understanding in the context of autonomous driving, even when given unusual visual data such as LIDAR [11] which is a common modality of autonomous vehicle data. These approaches have not been integrated with Autonomous Vehicle controllers, but the potential for VLMs to enhance Autonomous Vehicles is clear. Relatively small VLMs, such as PaliGemma 3B [2] are capable of taking input frames from a video and detecting a particular kind of object, such as a car. They are also capable of segmenting the environment into its most important components, which makes them suitable for extracting a large amount of information from their output given a suitable query.

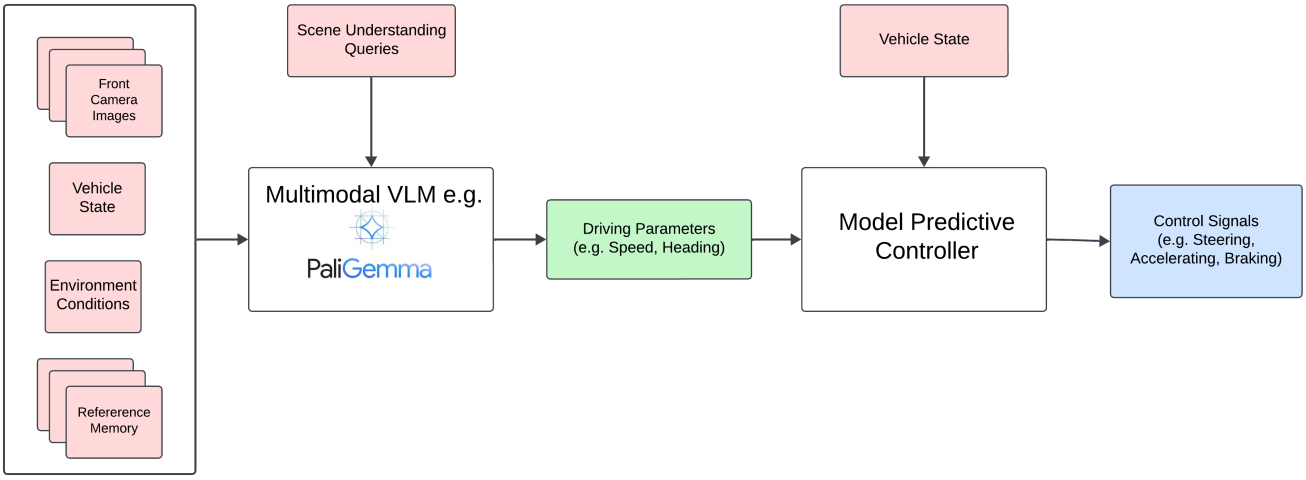

Figure 1: Overview of the decision-making process for our proposed VLM-MPC system for Autonomous Vehicle Control.

Figure 2: PaliGemma 3B can accurately identify detect moving cars in a video.

3.3 Semantic Querying for Object Attributes

In addition to a general understanding of the world that can be provided to us through VLMs, we can also gain additional information from making queries to the VLM about important objects detected in the video. If the VLM detects a car for example, we could ask the VLM whether the car appears to be heading towards us or away from us, whether our vehicle is on a collision course with a traffic cone or whether we are safely out of the way, along with many other pieces of information we could extract from the scene. This step leverages the semantic knowledge of the VLM in order to learn more about each object in the scene.

3.4 Integration of Semantic Insights into MPC

Once we have semantic insights on the scene from the VLM, we need to find a way to integrate them into the MPC that makes full use of the information available while also ensuring that information is usable by the MPC. We propose that the output of the VLM be formatted in terms of driving parameters such as the Speed and Heading of objects of note such as other vehicles. This information, combined with the known state of the Vehicle, can better inform the MPC and lead to better performance due to the additional information about the environment.

3.5 Real-time Processing and Decision Making

Due to the flexibility of our proposed framework, we can apply VLMs of various scales and sizes to see which VLMs allow for the best balance of robust information and speed. Autonomous vehicles are absolutely required to make real-time decisions, as a life-or-death situation could come at any moment. With this in mind, larger models may be infeasible, while ultra-small models such as those designed for mobile phones may not fit into the right semantic capabilities required for the MPC to make good decisions. We can explore these different scales of VLM to get a model size that fits our requirements for both robust decision making and real-time decision making. Processing the information will also need to be done quickly in order for the VLM output to be realized in time in the first place, meaning the other inputs to the VLM such as the vehicle state, environment conditions, adn reference memory must also be represented in a compact format that can be processed quickly.

3.6 Safety-Centric Approach

As the main goal of integrating VLMs with a MPC is to provide control signals for an autonomous vehicle, safety must be of paramount importance. This is where the advantages of using an MPC really shine through. In order to reach safe control signals, we can impose additional safety constraints on the MPC. We can prevent the MPC from accelerating too much or from turning too sharply. This adds an additional layer of robustness to the controller, as output from the VLM still has the potential to be off due to hallucinations and other inaccuracies. There is a robust body of literature on imposing these sorts of constraints, but specifically engineering constraints

for information sourced from the VLM can be another contribution of implementing this framework.

3.7 Adaptability and Generalization

With a VLM-backed MPC system, there is a significant ability for the autonomous vehicle to be able to adapt to unseen scenarios and other information that it sees repeatedly. The world knowledge stored in the VLM as a result of pretraining on a vast amount of data means that decisions can be made with information that has a very slim but nonzero chance of showing up in an environment such as the road. The massive corpus of data that is used to pretrain the VLM will have many different pieces of data that an ordinary dataset for training an autonomous vehicle may not necessarily have exposure to. This advantage could give our framework a significant advantage over other autonomous vehicle frameworks.

3.8 Human-AI Collaboration

Current methods still fall short of true autonomous vehicles. It is necessary for humans to be able to see how the VLM-MPC system for autonomous vehicles is creating and using its control signals. For this purpose we can potentially integrate explainable AI (XAI) techniques to gain more insight into how our VLM is coming up with the parameters fed into the MPC. Thankfully, MPC controllers are more naturally interpretable and so XAI techniques may not necessarily even be needed in order to understand that component, so long as the concepts behind the output parameters from the VLM make intuitive sense to a human observer.

4 CONCLUSION AND CALL TO ACTION

The integration of Multimodal Vision Language Models (VLMs) with Model Predictive Control (MPC) represents a significant leap forward in autonomous vehicle technology. This novel framework has the potential to revolutionize how autonomous vehicles perceive, understand, and navigate complex real-world environments.

By leveraging the semantic understanding capabilities of VLMs and combining them with the predictive and adaptive strengths of MPC, we can create autonomous systems that are more context-aware, safer, and more efficient. This approach addresses critical challenges in autonomous vehicle control, such as real-time decision-making in dynamic environments, handling of unforeseen scenarios, and improved interaction with human drivers and pedestrians.

However, realizing this vision requires concerted effort from the research community. We call upon researchers, engineers, and practitioners in the fields of artificial intelligence, robotics, and autonomous systems to:

(1) Explore VLM-MPC Integration: Investigate novel ways to effectively combine the strengths of VLMs and MPC, focusing on real-time performance and safety guarantees.
(2) Advance VLM Capabilities: Develop more efficient and task-specific VLMs that can operate within the computational constraints of autonomous vehicles.
(3) Enhance Safety Frameworks: Create robust safety validation methods that can rigorously test and verify VLM-enhanced autonomous vehicle control systems.
(4) Address Ethical Considerations: Engage in interdisciplinary research to tackle the ethical implications of using AI for critical decision-making in autonomous vehicles.
(5) Standardize Evaluation Metrics: Develop comprehensive benchmarks and evaluation frameworks that can assess the performance of integrated VLM-MPC systems across diverse scenarios.
(6) Foster Collaboration: Encourage partnerships between academia, industry, and regulatory bodies to accelerate the development and deployment of these advanced autonomous systems.

By pursuing these research directions, we can unlock the full potential of VLM-enhanced autonomous vehicles, paving the way for safer, more intelligent, and more adaptable transportation systems. The fusion of VLMs with MPC not only promises to advance the field of autonomous vehicle control but also opens up new possibilities for intelligent autonomous agents across various domains.

REFERENCES

[1] Mustafa Hamid Al-Jumaili and Yasa Ekşioğlu Özok. 2024. New control model for autonomous vehicles using integration of Model Predictive and Stanley based controllers. *Scientific Reports* 14, 1 (2024), 19872.
[2] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. 2024. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726* (2024).
[3] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. 2024. pi0: A Vision-Language-Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164* (2024).
[4] 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.
[5] Tianhui Cai, Yifan Liu, Zewei Zhou, Haoxuan Ma, Seth Z Zhao, Zhiwen Wu, and Jiaqi Ma. 2024. Driving with Regulation: Interpretable Decision-Making for Autonomous Vehicles with Retrieval-Augmented Reasoning via LLM. *arXiv preprint arXiv:2410.04759* (2024).
[6] Ernst D Dickmanns and Alfred Zapp. 1987. Autonomous high speed road vehicle guidance by computer vision. *IFAC Proceedings Volumes* 20, 5 (1987), 221–226.
[7] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9710–9719.
[8] Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214* (2024).
[9] Cole Gulino, Justin Fu, Wenjie Luo, George Tucker, Eli Bronstein, Yiren Lu, Jean Harb, Xinlei Pan, Yan Wang, Xiangyu Chen, et al. 2024. Waymax: An accelerated, data-driven simulator for large-scale autonomous driving research. *Advances in Neural Information Processing Systems* 36 (2024).
[10] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. arXiv:2406.09246 [cs.RO] https://arxiv.org/abs/2406.09246
[11] Guibiao Liao, Jiankun Li, and Xiaoqing Ye. 2024. VLM2Scene: Self-Supervised Image-Text-LiDAR Learning with Foundation Models for Autonomous Driving Scene Understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 3351–3359.
[12] Keke Long, Haotian Shi, Jiaxi Liu, and Xiaopeng Li. 2024. VLM-MPC: Vision Language Foundation Model (VLM)-Guided Model Predictive Controller (MPC) for Autonomous Driving. arXiv:2408.04821 [cs.RO] https://arxiv.org/abs/2408.04821
[13] Runliang Niu, Jindong Li, Shiqi Wang, Yali Fu, Xiyu Hu, Xueyuan Leng, He Kong, Yi Chang, and Qi Wang. 2024. Screenagent: A vision language model-driven computer control agent. *arXiv preprint arXiv:2402.07945* (2024).

[14] Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. 2018. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems* 1, 1 (2018), 187–210.
[15] TM Vu, R Moezzi, J Cyrus, and J Hlava. 2021. Model Predictive Control for Autonomous Driving Vehicles. Electronics 2021, 10, 2593.
[16] Xiaohan Zhang, Yan Ding, Saeid Amiri, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang. 2023. Grounding classical task planners via vision-language models. *arXiv preprint arXiv:2304.08587* (2023).
[17] Wentao Zhao, Jiaming Chen, Ziyu Meng, Donghui Mao, Ran Song, and Wei Zhang. 2024. Vlmpc: Vision-language model predictive control for robotic manipulation. *arXiv preprint arXiv:2407.09829* (2024).
[18] Wei Zhao, Pengxiang Ding, Min Zhang, Zhefei Gong, Shuanghao Bai, Han Zhao, and Donglin Wang. 2025. VLAS: Vision-Language-Action Model With Speech Instructions For Customized Robot Manipulation. *arXiv preprint arXiv:2502.13508* (2025).
[19] Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. 2024. Can Large Language Models Understand Context? *arXiv preprint arXiv:2402.00858* (2024).