Proceedings of the 4th International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/2026.33498

Applications of Vision-Language Models in Autonomous Driving

Yunjia Li

International School, Beijing University of Posts and Telecommunications, Beijing, China
liyunjia@bupt.edu.cn

Abstract. This paper highlight how Vision-Language Models (VLMs) are used in self-driving technology. When it comes to self-driving system decision-making challenges, such as Long-Tail scenario generalization, causality, and explainability, modular and end-to-end reinforcement learning approaches have not been effective in representational learning and generalization. Vision-Language Models have the ability to develop and reason about images, which provides them with the ability to solve the problems that faced in the field. Thus, using VLMs in autonomous driving problems is an emerging field. The current research can be divided into three categories. The first category is VLMs modified into vision-language-action frameworks as the decision making core; the second category is the integration of VLMs with the classical autonomous driving perception and planning modules (collaborative hybrid framework) in order to address the limitations of VLMs; the last category is about employing VLMs as a "teacher" model using knowledge transfer to improve and/or steer a lightweight reinforcement learning strategy network. Both approaches, however, have numerous problems to solve. This paper hope that future research will improve the deployability and robustness in complicated situations for the driving models using techniques in knowledge distillation.

Keywords: vision-language models, reinforcement learning, autonomous driving.

1. Introduction

The fundamental challenges are the extreme complexity of and incompleteness of application scenarios in autonomous driving decision systems. Following the evolutionary path of the technology, the decision-making paradigm of autonomous driving has been considerably evolved. Initially, studies were predominantly based on expert systems and modular architectures. Classical modular approaches are based on the functional decomposition paradigm of perception, localization, planning and control [1,2]. Although the independent accuracy of individual sub-modules and completeness of intermediate representations are vital to the performance of this architecture, it exhibits logical clarity. This modularity leads to the system not having any context-awareness or cross-modularity to coordinate different modules and instead only a locally optimal solution will be obtained and not the global optimal decision that involves uncertainty perception, behavioral prediction and motion control [1,2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.33498

In order to put the end-to-end architecture into practice, the following study examined how to make decisions using the deep reinforcement learning. The approach will create one neural network approach to convert raw sensor signals to vehicle control instructions. It is often implemented by building a Markov decision process in a simulation. The agent maximizes cumulative rewards by designing reward functions to represent driving goals and safety constraints, and using optimization methods like policy gradient, through many trials and errors [2] of trial-and-error. The method however faces several practical difficulties: Firstly, sample efficiency is not optimal with agents finding it difficult to extract high-level semantic aspects of objects which leads to slow learning dynamics [2,3]. Second, decision-making process involves high-dimensional parameter spaces with nonlinear mappings that do not have interpretable intermediate representations. This results in a black box that is completely untraceable, and thus a major obstacle to credibility evaluation as well as determination of liability in safety realms [2].

Recent advancements in Vision-Language Models have offered new technical solutions in solving these problems [4]. VLMs are multimodal AI systems that combine both visual perception and natural language understanding. They are usually designed as two main components that are connected by an alignment module. The visual portion generally uses visual encoders pre-trained on large-scale image collections to extract features, such as Transformer-based Vision Transformer models or visual language models designed specifically to work with visual language. These models match images and text in a common semantic space by contrastive learning. Large language models in the language parts are mainly based on the Transformer decoder architecture, including the LLaMA series or GPT series models, which rely on self-attention mechanisms to process and understand text sequences. The technology is the multimodal alignment training that is normally carried out on millions or even billions of image-text pairs data. This allows the model to build associations between pixels and words thus gaining the ability to describe image content, answer complex image related questions and reason based on visual information [2-4]. This, in autonomous driving decision-making context, means that the system can directly derive scene understanding with common sense and physical logic based on multimodal perception information, beyond conventional object categories and positions. Its generative heart allows the system to provide natural language descriptions of scene reasoning of proposed actions, and so offers novel opportunities to the explainability of decision making [2,5,6]. Thus, as a direction of research, the integration of VLMs into the autonomous driving decision-making systems with the help of the advanced open-world knowledge and semantic reasoning has become one of the most promising research directions. This strategy improves the system generalization functionality on new grounds, safety regulatory limits, and creates a clear decision-making base.

2. VLMs as end-to-end executor

VLMs function as end-to-end decision-makers and thus mostly within the VLA (Vision-Language-Action) paradigm [2]. Researchers were the first to explore the VLA paradigm in the area of robotic manipulation. It is fundamentally based on the fact that it combines the robust visual comprehension and language-teach-back abilities of the already trained VLMs with the inner logic of robots, which allow them to directly produce action sequences in response to natural language instructions [7]. This paradigm illustrates that VLMs have the potential to move across modalities and convert high-level semantic knowledge into actual action and thus offer a new way of solving intricate tasks in open-world contexts. The paradigm was rapidly expanded to the autonomous driving field, and the translation and implementation of the VLA paradigm to autonomous driving (VLA4AD) became a line of research [2].

DOI: 10.54254/2753-8818/2026.33498

The VLA does not explicitly refer to off-the-shelf VLMs, but rather does extensive deconstruction and recombination of them. It uses the strong visual encoder and language processor of Vlm to be the basis of perception and reasoning, but a critical change by introducing a new, trainable action decoder. Action decoder is the central execution component of the model VLA, which transforms the combined visual-linguistic high-level representations into particular driving commands [2]. It takes the cognition of the model about the scene and the results of its reasoning and translates them into physical actions to be physically implemented by the vehicle. Particularly, it has two main tasks: First, it produces control signals by either discrete prediction of continuous actions by autoregressive tokenization or noise reduction by diffusion models to come up with smooth trajectories [2]. Second, the output format, whether to provide simple steering, throttle, brake controls, or produce a sequence of future waypoints that are tracked later [2]. This provides a complete end-to-end unified architecture of vision, language and action. Training paradigm has changed image-text pair learning of VLM to a multi stage view incorporating imitation learning and reinforcement learning using image-instruction-action triplets, and eventually converting model outputs to the directly executable driving actions or trajectories [2]. OpenDriveVLA team proposes hierarchical vision-language alignment and hierarchical representations on the nuScenes, which has obtained high-quality planning performance and scene understanding [8]. FSDrive team is one step behind VLA by suggesting a visual spatio-temporal CoT (reasoning with future frames) on nuScenes which enhances accuracy of trajectories and minimizes collisions [5]. All these are indicators of the potential of VLA.

Although potentially promising, the existing VLA approaches have some basic theory-to-implementation problems. In theory their encapsulated design inherits and exacerbates intrinsic VLM weaknesses: overreliance on sparse, high-quality alignment data impedes acquisition of robust physical laws; generative Chain-of-Thought explanations can give an illusion of interpretation that is not linked to actual decision logic; and most importantly, VLM hallucinations are directly translated into unsafe driving behaviour [2,3,9]. Engineeringly, there are grave contradictions between billion-scale scales and real-time performance on resource-limited platforms, and the autoregressive generative paradigm is incompatible with milliseconds response times [2,9]. Moreover, it presents enormous computational pressures to process multimodal time-series data. As a result, current VLA systems still are more of research prototypes, and their journey to become production grade, automotive products requires radical innovations in base architecture, training models and hardware co-design [9].

3. VLMs as collaborative decision-makers

To cope with the extreme performance difference that occurs after simply applying visual language models to autonomous driving- where the high-level semantic understanding and common-sense reasoning abilities are hard to be mapped into high-frequency, high-precision, and verifiable underlying control signals- the study suggests a novel collaborative decision-making architecture [6]. This architecture is not about the conventional autonomous driving stack being replaced by Vlm, but rather about repositioning it as a slow-thinking collaborative decision-making unit that is a dynamically complementary dual-channel system with the conventional fast systems.

In particular, the system initially initiates the selective changes on the VLM so as to transform its output into structured hierarchical planning instructions, as opposed to direct control commands [6]. These top-level instructions are converted to constraints that can be understood by the traditional planners in real time using a carefully-designed interface layer, and are fed into the deterministic trajectory generation module using an optimization or rule-based approach. Its fundamental

DOI: 10.54254/2753-8818/2026.33498

innovations are reflected in various aspects: the knowledge-sharing paradigm enables the open-world knowledge of VLM to supply the decision-making priors in long-tail scenarios; the paradigm of asynchronous collaboration makes VLM work like a low-frequency, high-intelligence module in addition to the high-frequency modules; and the result caching and generalization technologies transform the VLM inference results into reusable sem The concept of autonomous driving is the dual-system called DriveVLM (Tian et al.), which has been deployed in production, a major breakthrough in the engineering deployment of autonomous driving [6].

There are also a row of contradictions and challenges to the approach. First, it presents a complex cascaded system, in which there is a semantic alignment and quantization loss between the semantic output of VLM and the mathematical optimization of conventional planners. Any bias in understanding can result to the lack of certainty when planning. Secondly, the system has a limitation on its general performance due to constraints of traditional modules. The intelligence of VLM operates as a kind of co-pilot advisor, yet its deep understandings may not be achieved as fully as possible because of the capability limitation of the conventional planner- the executive driver, and this leads to performance trade-offs. Though the dual-system design increases performance of tasks substantially [1-3], in high complex continuous decision-making tasks, real-time performance bottlenecks remain, and can result in delays in decisions when frequent VLM activations occur. In addition, the verification of system reliability and safety becomes extremely complicated, and the joint verification of both the semantic robustness of neural networks and the stability of conventional control algorithms must be conducted simultaneously [1].

Finally, VLM as a collaborative decision-maker is a very engineering-smart and transitional solution. It was able to incorporate the cognitive abilities of large models into established trusted autonomous driving systems, with the first deployment of VLMs being used on production vehicles and showing the promise of integrating VLMs with conventional systems [6]. But it is exactly the success of compromise -the simplicity and possible performance ceiling of end-to-end principles sacrificed to modern deployability and security -that is its success. It can be said therefore to be a linking point between 'dedicated intelligence' and general driving intelligence as opposed to being the end result itself.

4. VLM as a knowledge provide

Another mainstream remedy to the performance gap due to VLM is concerned with knowledge transfer. Knowledge transfer is a machine learning strategy whose essence is to convey successful knowledge or patters acquired by a complex model (teacher model) to a simpler and more efficient model (student model) so that it can perform similarly or even better [10]. Its main idea is to learn or compress the VLM world experience (several semantic interpretations of traffic situations and common-sense knowledge of how objects and behaviors interact) into lightweight, specialized representations [3,11,12].

Nowadays, the prevailing method is knowledge distillation [3,11,12]. The main paradigms of knowledge distillation are the representation distillation and policy distillation. Representation distillation a prompt-inference pipeline is designed to guide the VLM through stepwise reasoning over historical observations, producing important semantic categories as supervision signals, allowing the visual encoder of the student model to learn task-relevant semantic representations. An example of this is the DGC framework, which translates the semantic interpretation capacity of VLM into the visual encoder of the reinforcement learning agents, thus improving the efficiency of samples [3]. The policy distillation approach applies VLM-based "teacher model to produce high-quality decisions or trajectories, which direct the training of student policy networks using either

DOI: 10.54254/2753-8818/2026.33498

reinforcement learning or multi-agent reinforcement learning methods. On the basis of this approach the LDPD (Liu et al.) incorporated semantic scene enhancement and a safety checker, which enables the students to ultimately outperform the teacher in the ramp merging cases [12].

Although innovatively new, these solutions bypass the performance bottlenecks of direct Vlm deployment and have a lot of potential, they are also associated with critical problems, especially those of reliability and security. To begin with, there are still problems of interpretability and security validation. The distillation/migration process is a black box and it is not easy to determine what information is learned by VLMs by student models in what form or whether the decisions arrived at by student models in extreme conditions are actually consistent with safety ethics [3,9,11]. There is a risk that student models can only learn the statistical correlations in VLM decision-making without learning the causal logic behind them, which creates hidden risks in real-world application. Second, the flaws, in turn, peculiar to the VLM, may be inherited as well as even aggravated during the migration process, which is a particularly acute issue. A VLM containing bias will inculcate this bias to the model [10] as knowledge to the students. In safety-critical areas like autonomous driving, the spread implications of such implicit bias can be disastrous. The issue is, can VLM, relying on preference judgments with the use of fixed images or brief sequences, fully and dynamically represent the optimality of long-term driving strategies?

Finally, establishing VLM as a knowledge provider is an extremely promising new trend in autonomous perception and decision-making. Through the transfer of knowledge it is an effective way of bridging the performance gap but by no means is a permanent solution.

5. Conclusion

This paper has provided a systematic literature review of the development of visual language models and their essential paradigms in the area of autonomous driving decision-making. To overcome the shortcomings of conventional modular methods and end-to-end reinforcement learning in long-tail scenario generalization, causal inference, and interpretability, VLMs leverage their strong cross-modal cognition and semantic reasoning capabilities to deliver a new breakthrough in the autonomous driving decision systems.

Mainstream research methods currently revolve mainly around three paradigms: First, converting VLMs into visual-language-action models to be used as interpretable end-to-end decision-making cores. Second, make VLMs appear as slow-thinking decision units that interact with conventional perception planning stacks, building hybrid intelligence architectures to trade off between open-world knowledge and strong real-time control. Third, applying VLMs as teacher models to pass domain knowledge to lightweight specialized policy networks, either with techniques of knowledge distillation or with techniques of reward design. These strategies use the semantic background and logic of VLMs to different extents, seeking to improve the flexibility of the system in unforeseen conditions as well as the transparency of the decision-making process.

Nevertheless, there are still a row of dramatic challenges with this technological approach. The safety-critical scenario model hallucination problems can be amplified by error in content, to operational risks; its autoregressive generation paradigm is fundamentally incompatible with high real-time requirements; decision-making logic of this black-box model is poorly verifiable; and its computational requirements are high enough that deployment in automotive systems is not feasible. All these issues are the fundamental obstacles of the VLMs to go beyond research prototypes to industry-grade products in autonomous driving.

In the future, studies on this topic should aim to find breakthroughs at various levels. On the algorithmic level, it is crucial to consider more productive and stable model structures and training

DOI: 10.54254/2753-8818/2026.33498

paradigms to essentially address hallucinations, improve interpretability, and decrease computational complexity. At the engineering level, more efficient model compression, distillation, and hardware co-design technologies are needed to be developed. Meanwhile, it is essential to define simulation and evaluation standards that can be used to properly reflect driving physics and social rules to validate and develop related research. Finally, with ongoing technological innovation and optimization of the system, VLMs will be easily incorporated into the closed-loop driving decision-making system offering fundamental capabilities to facilitate safer, more adaptable and reliable autonomous driving systems.

References

[1] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. (2023). Planning-oriented Autonomous Driving. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17853–17862.

[2] Sicong Jiang, Zilin Huang, Kangan Qian, Ziang Luo, Tianze Zhu, Yang Zhong, Yihong Tang, Menglin Kong, Yunlong Wang, Siwen Jiao, et al. (2025). A Survey on Vision-Language-Action Models for Autonomous Driving. IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 4524–4536.

[3] Haoran Xu, Peixi Peng, Guang Tan, Yiqian Chang, Luntong Li, and Yonghong Tian (2025). VLMs-Guided Representation Distillation for Efficient Vision-Based Reinforcement Learning. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 29534–29544.

[4] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat (2023). Gpt-4 technical report. arXiv preprint arXiv: 2303.08774.

[5] Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., ... & Guo, N. (2025). Futuresightdrive: Thinking visually with spatio-temporal cot for autonomous driving. arXiv preprint arXiv: 2505.17685.

[6] Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, Hang Zhao, et al (2024). DRIVEVLM: The Convergence of Autonomous Driving and Large Vision-Language Models. arXiv preprint arXiv: 2402.12289.

[7] Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., & Zhu, Y. (2025). Vision-language-action models for robotics: A review towards real-world applications. IEEE Access.

[8] Zhou, Xingcheng, Xuyuan Han, Feng Yang, Yunpu Ma, Volker Tresp, and Alois Knoll (2026). OpenDriveVLA: Towards End-to-end Autonomous Driving with Large Vision Language Action Model. AAAI Conference on Artificial Intelligence, 13782-13790.

[9] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, Liang Pan. (2025). Are VLMs Ready for Autonomous Driving? An Empirical Study from the Reliability, Data, and Metric Perspectives. IEEE/CVF International Conference on Computer Vision (ICCV), 6585-6597.

[10] Gholizade, M., Soltanizadeh, H., Rahmanimanesh, M., & Sana, S. S. (2025). A review of recent advances and strategies in transfer learning. International Journal of System Assurance Engineering and Management, 16(3), 1123-1162.

[11] Hegde, D., Yasarla, R., Cai, H., Han, S., Bhattacharyya, A., Mahajan, S., ... & Porikli, F. (2025). Distilling multi-modal large language models for autonomous driving. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 27575-27585).

[12] Liu, J., Xu, C., Hang, P., Sun, J., Ding, M., Zhan, W., & Tomizuka, M. (2025). Language-driven policy distillation for cooperative driving in multi-agent reinforcement learning. IEEE Robotics and Automation LettersAuthor, A. A., Author, B. B., & Author, C. C. (Year). Title of article. Title of Periodical, volume number(issue number), pages.