Embodied Intelligence in Mining: Leveraging Multi-modal Large Language Model for Autonomous Driving in Mines

Luxi Li†, Yuchen Li†, Xiaotong Zhang, Yuhang He, Jianjian Yang, Bin Tian, Yunfeng Ai, Lingxi Li, Andreas Nüchter, Zhe Xuanyuan*

***Abstract*—With computer technology advancing in both software and hardware, the benefits of embodied intelligence are becoming increasingly evident. This robust interactive learning model enables artificial intelligence (AI) to be more flexibly deployed across diverse fields. In recent years, the development of multi-modal large language models (LLMs) has further accelerated the progress of AI, prompting extensive research on how to leverage these advancements to enhance the field of autonomous driving. This perspective believes that embodied intelligence can significantly enhance the application of LLMs, analyzing the new opportunities brought to the mining industry, and emphasizing the potential of their integration to revolutionize various aspects of the field. Meanwhile, This perspective also examines the challenges of deploying embodied agents in mining, while emphasizing their promising future and offering insights into potential research and development avenues.**

***Index Terms*—embodied intelligence, large language model, intelligent mining**

I. Introduction

Recently, the emergence of Large Language Model (LLM) has propelled artificial intelligence (AI) into unprecedented realms[1], [2], [3], [4], ushering in a new era of text generation capabilities. The advent of LLMs not only advances AI in the realm of textual comprehension and generation but also catalyzes developments in related fields. Their integration with robotics undeniably injects fresh vigor into embodied intelligence.

Luxi Li and Yuchen Li contributed equally to the article. Corresponding author: Zhe Xuanyuan. This work was supported in part by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College 2022B1212010006, in part by Guangdong Higher Education Upgrading Plan with UIC research grant R0400001-22 and R201902.

Luxi Li and Yuchen Li are with the Faculty of Science and Technology, BNU-HKBU United International College, Zhuhai, 519087, China, and the Faculty of Science, Hong Kong Baptist University, Kowloon, Hong Kong, 999077, China, and WAYTOUS Inc., Beijing, 100083, China. Yuchen Li is also with the Computer Science at the Department of Informatics of the Technische Universität München, Garching b. München, 85748, Germany. (e-mail: lucylee@whu.edu.cn).

Xiaotong Zhang, Yunfeng Ai and Bin Tian are with the Institution of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and WAYTOUS Inc., Beijing, 100083, China.

Yuhang He is with the University of Oxford, Oxford, UK.

Jianjian Yang is with the China University of Mining and Technology (Beijing), Beijing, China.

Lingxi Li is with Purdue University, West Lafayette, USA.

Andreas Nüchter is with Würzburg University, Würzburg, Germany.

Zhe Xuanyuan is with the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College. (e-mail: zhexuanyuan@uic.edu.cn)

Embodied intelligence refers to the ability of agents to understand and manipulate the objective world through interactions with their environment and learning [5]. Historically, intelligent algorithms have been constrained by their reliance on training data, limiting their flexibility when addressing real-world problems. This challenge has made dynamic interactive embodied learning a difficult feat. The advent of LLMs, however, has opened up the possibility for real-time interaction, allowing them to adeptly handle embodied question-and-answer tasks. Additionally, multimodal LLMs offer innovative solutions for embodied tasks such as visual exploration and navigation.

This perspective believes that the embodied AI with LLM is expected to play a crucial role in mining autonomous driving and offer new solutions for this field, better-achieving mining 5.0 [6], [7], [8]. Integrating embodied AI and LLM enhances the ability of mining autonomous vehicles to interact with and learn from the environment, thereby improving the overall performance of autonomous driving.[9] This perspective discusses embodied intelligence with LLM and its application in the field of autonomous driving, and analyzes the potential challenges and future direction.

II. Embodied intelligence with LLM

Built on large-scale neural networks, LLMs are trained on extensive textual data, enabling them to develop a deep understanding of language patterns, semantics, and syntax. They can generate coherent text, answer questions, summarize information, and engage in dialogue with humans. Additionally, multimodal LLMs integrate information from modalities beyond language, like sound and vision, enhancing their comprehension abilities. Embodied multimodal LLM integrates perception ability with embodied cognition, allowing them to comprehend and generate language while interacting with the environment, representing a significant advancement in AI.

Embodied LLMs have witnessed significant advancements across various models, each contributing uniquely to the field: The emergence of Flamingo [10] demonstrates the few-shot learning ability in the field of visual language modeling. By combining pre-trained language models with external knowledge, the system is adept at navigating open-domain generative question answering and other knowledge-intensive tasks. Google's PaLM-E[11] realizes the landing of a multi-modal LLM in robot scene training, which is an important

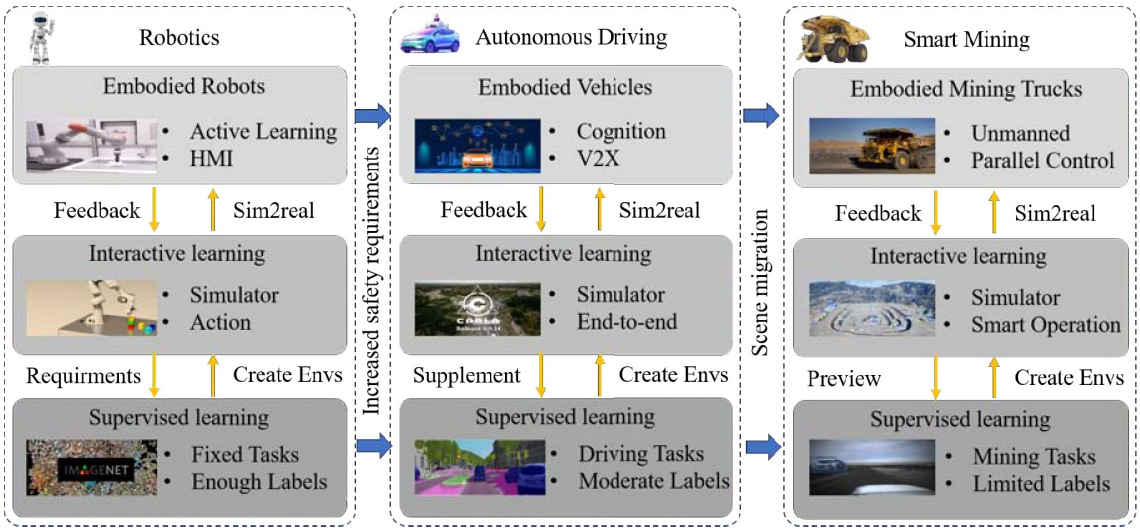

Fig. 1: The application of embodied AI in the field of robotics, autonomous driving and smart mining.

milestone because it demonstrates the feasibility of multi-modal language model in robot multi-data source multi-scene fusion. Microsoft published the paper[12], which proposed the basic idea of using ChatGPT to control robots, and provided a new perspective for the application of LLM in embodied intelligence. ShapeLLM[13] specializes as a 3D multi-modal large-scale language model designed for embodied interaction. Through the fusion of 3D point clouds and language, it ventures into the realm of 3D object understanding.

These advancements collectively underscore the rapid progress and immense potential of embodied intelligence with multimodal LLM in revolutionizing robotic cognition and interaction within the physical world.

At the same time, many LLMs for autonomous driving have emerged, which has promoted the development of autonomous driving. The application of LLM in autonomous driving is mainly focused on two aspects: providing environmental understanding and interaction, and generating driving behaviors and decisions. These models process and analyze a large amount of multi-modal driving data, including images, videos and language commands, so that autonomous driving systems can perform tasks more accurately and stably.

At present, there have been many datasets of LLM for autonomous driving [14], [15], [16], [17], [18], which contain a large number of video frames and corresponding text descriptions, which are suitable for perception and planning tasks in autonomous driving.

DriveLikeaHuman[19] explores the potential of using LLMs to understand the driving environment in a human-like way and analyze its ability to reason, interpret, and remember when faced with complex scenarios. DriveMLM[20] proposes a framework to integrate the world knowledge and reasoning capabilities of LLMs into an autonomous driving system, enabling closed-loop driving in real-world simulators. DriveVLM[21] integrates a unique combination of Chain-of-Thought (CoT) modules for scenario description, scenario analysis, and hierarchical planning. VELMA[22] designs LLM agents for navigation in street view, which greatly improves the performance of visual navigation tasks. LimSim++[23] proposes An autonomous driving closed-loop simulation platform designed for multi-modal LLMs, aiming to apply LLMs to autonomous driving, it addresses the need for a long-term closed-loop infrastructure and supports continuous learning and improved generalization capabilities in autonomous driving.

It can be noticed that the emergence of LLM has injected new vitality into the development of automatic driving. This perspective believes that embodied intelligence is expected to become a new carrier for multi-modal LLM to assist automatic driving to better assist automatic driving promoting the migration of autonomous driving from cities to other scenarios. The following chapters will explore the application of embodied intelligence migration to smart mines, and the implications and future directions.

III. Application of Embodied AI in Mining

Fig. 1 depicts the application of embodied intelligence in the field of robotics, autonomous driving and smart mining.

In the field of robotics, embodied AI refers to robots that can interact with the environment and engage in exploratory learning. Equipped with a Human Machine Interface (HMI), these robots can complete specific tasks through human-machine interaction commands, such as visual exploration, visual navigation, embodied question answering, and more. The development of simulators has facilitated embodied learning for robots. A high-quality simulator must construct not only realistic environments but also realistic interactions between agents and objects or between objects, modeling real-world physics properties. In such an environment, robots learn the best strategies to complete tasks through interaction with the physical environment. Whether in a simulator or the real world, the amount of data available to embodied robots is very sufficient, supporting the possibility of embodied robots learning in the real world in the future.

When it comes to the field of autonomous driving, the requirement for safety increases significantly. V2X technology

Fig. 2: The system for embodied intelligence in smart mining.

enables vehicles to have the physical foundation to interact and communicate with everything. Cognitive abilities allow them to better understand all interactive behaviors, laying the foundation for the implementation of embodied intelligent vehicles. Smart cars undergo end-to-end training in simulators, establishing a mapping between perception and control, thereby supporting the migration of models to actual vehicles to complete real driving tasks.

Embodied intelligence provides new ideas for autonomous driving, prompting the transition of the autonomous driving system from being a spectator in intelligent transportation to being an actual participant. The interactive learning mode has the potential to learn deeper abstract knowledge, enabling autonomous driving to break through the bottleneck of difficult scene migration. This also implies the huge potential of embodied intelligence in mining scenarios. The lack of data has always been a painful issue for traditional solutions to achieve autonomous driving in mines. By building a parallel mine simulator, vehicles can perform end-to-end learning in the simulator, alleviating the pressure of data collection and annotation. Alternatively, by utilizing small batches of data through migration, urban autonomous driving agents can adapt to mining scenes, creating favorable conditions for the progress of autonomous driving in mines.

The realization of an embodied intelligence system for mines requires the support of the following modules (shown in Fig.2):

Equipment: The driving operations of embodied vehicles should be uniformly controlled by a domain controller. The transportation equipment system interacts with vehicles and roadside equipment through the V2X module. Meanwhile, a remote control module supports drivers in remotely and synchronously operating vehicles. **Data**: Excellent mine-driving decisions rely heavily on data, including sensor data from cameras, lidars, and communication data such as scheduling and control information. **LLMs**: The implementation of embodied intelligence algorithms relies on the support of LLM. Multi-modal LLM integrates data from modalities such as vision and language to achieve embodied question answering. The introduction of CoT for reasoning and decision-making establishes an end-to-end mapping from perception to control while preserving the logic of planning and decision-making. This enhances the model's interpretability, improving its credibility and explainability, which is crucial for enhancing the safety of unmanned mining. **Platform**: The platform should meet the demands of embodied intelligence for computing resources, data storage, model training, simulation testing, and safety supervision. Therefore, storage devices, computing servers, simulators, test vehicles, and operating supervision modules are required. **Products**: Embodied intelligence should not only operate in simulators but also have the ability to migrate to reality. This requires specific products such as cars, trucks, wide-body trucks, and mining trucks, which together contribute to the construction of smart mines.

It is believed that embodied intelligence would contributes to the construction of smart mines, in the future, its realization should endow vehicles with the following capabilities:

Self-awareness: Intelligent algorithms are no longer mutually independent in a modular form but contribute to the embodied mining truck's understanding of itself. It should understand its size, shape, physical structure, and characteristics, as well as driving tasks and goals. It should possess the ability for mine visual exploration and navigation, enabling it to better complete tasks such as cargo loading, unloading, and

transportation in the mine.

Interactive cognition: This includes the inherent interaction between the sensory and control systems, and the external interaction between the individual and the environment. Inherent interaction requires vehicles to have a sensorimotor system that establishes a mapping relationship between sensation and control. External interaction requires vehicles to consider the reaction to themselves during interaction with the environment.

Risk awareness: Embodied mining trucks should be aware of the potential risks posed by the environment and control actions, such as vehicle bumps and slips caused by road conditions, understanding the dangers of colliding with obstacles to both parties, and the adverse effects of extreme weather on driving. Embodied mining trucks should have the ability to recognize and avoid risks.

IV. Implications and Future Directions

Although Embodied intelligence has shown a promotion role in smart mines, there are still some potential problems. This section discusses these potential issues and future developments.

High-quality data collection: Currently, most autonomous driving data is concentrated in urban areas, making the collection of high-quality mining data crucial.

System redundancy and fail-safe mechanisms: In order to cope with potential cyber-attacks, mine emergencies, or other failures, autonomous vehicle systems should be designed with redundancy and fail-safe mechanisms. These mechanisms can help mitigate the effects of cyberattacks and ensure that vehicles can still operate safely in case of a problem.

Real-time and computational efficiency: Autonomous vehicles require real-time data processing and reaction capabilities, presenting a computational challenge for Embodied LLM. Any sluggishness in processing speed can lead to delays that ultimately impact the vehicle's responsiveness.

Interpretability and transparency: The decision-making process of autonomous driving systems needs to be transparent to the outside world so that users and regulators can understand how they work and the basis for their decisions.

Ethical and ethical Decision Framework: Autonomous driving systems need a clear framework to guide their behavior when faced with ethical decisions. Autonomous vehicles can conduct decision training under this framework to expect the system to make ethical decisions in complex situations.

V. Conclusion

This perspective takes a deep dive into the application of embodied intelligence in mining, detailing the setup of an embodied AI system and prospects the characteristics of mining vehicles with embodied intelligence. By integrating language processing capabilities, embodied intelligence offers the potential to revolutionize the mining field, despite potential challenges, it is believed that embodied intelligence holds immense promise for the advancement of intelligent mining. As the field continues to evolve, further research and development efforts are warranted to fully realize the transformative potential in the mining industry.

References

[1] M. Zhang and J. Li, "A commentary of gpt-3 in mit technology review 2021," *Fundamental Research*, vol. 1, no. 6, pp. 831–833, 2021.

[2] O. (2023), "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] N. Roy, I. Posner, T. Barfoot, P. Beaudoin, Y. Bengio, J. Bohg, O. Brock, I. Depatie, D. Fox, D. Koditschek *et al.*, "From machine learning to robotics: Challenges and opportunities for embodied intelligence," *arXiv preprint arXiv:2110.15245*, 2021.

[6] S. Teng, X. Li, Y. Li, L. Li, Z. Xuanyuan, Y. Ai, and L. Chen, "Scenario engineering for autonomous transportation: A new stage in open-pit mines," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 3, pp. 4394–4404, 2024.

[7] L. Chen, J. Xie, X. Zhang, J. Deng, S. Ge, and F.-Y. Wang, "Mining 5.0: Concept and framework for intelligent mining systems in cpss," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3533–3536, 2023.

[8] Y. Li, S. Teng, L. Li, Z. Xuanyuan, and L. Chen, "Foundation models for mining 5.0: Challenges, frameworks, and opportunities," in *2023 IEEE 3rd International Conference on Digital Twins and Parallel Intelligence (DTPI)*, 2023, pp. 1–6.

[9] S. Teng, L. Li, Y. Li, X. Hu, L. Li, Y. Ai, and L. Chen, "Fusionplanner: A multi-task motion planner for mining trucks via multi-sensor fusion," *Mechanical Systems and Signal Processing*, vol. 208, p. 111051, 2024.

[10] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[11] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[12] S. H. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "Chatgpt for robotics: Design principles and model abilities," *IEEE Access*, 2024.

[13] Z. Qi, R. Dong, S. Zhang, H. Geng, C. Han, Z. Ge, L. Yi, and K. Ma, "Shapellm: Universal 3d object understanding for embodied interaction," *arXiv preprint arXiv:2402.17766*, 2024.

[14] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 563–578.

[15] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, "Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 4542–4550.

[16] T. Deruyttere, S. Vandenhende, D. Grujicic, L. Van Gool, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," *arXiv preprint arXiv:1909.10838*, 2019.

[17] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, and J. Li, "Drama: Joint risk localization and captioning in driving," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 1043–1052.

[18] M. Nie, R. Peng, C. Wang, X. Cai, J. Han, H. Xu, and L. Zhang, "Reason2drive: Towards interpretable and chain-based reasoning for autonomous driving," *arXiv preprint arXiv:2312.03661*, 2023.

[19] D. Fu, X. Li, L. Wen, M. Dou, P. Cai, B. Shi, and Y. Qiao, "Drive like a human: Rethinking autonomous driving with large language models," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 910–919.

[20] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li *et al.*, "Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving," *arXiv preprint arXiv:2312.09245*, 2023.

[21] X. Tian, J. Gu, B. Li, Y. Liu, C. Hu, Y. Wang, K. Zhan, P. Jia, X. Lang, and H. Zhao, "Drivevlm: The convergence of autonomous driving and large vision-language models," *arXiv preprint arXiv:2402.12289*, 2024.

[22] R. Schumann, W. Zhu, W. Feng, T.-J. Fu, S. Riezler, and W. Y. Wang, "Velma: Verbalization embodiment of llm agents for vision and language navigation in street view," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 18 924–18 933.

[23] D. Fu, W. Lei, L. Wen, P. Cai, S. Mao, M. Dou, B. Shi, and Y. Qiao, "Limsim++: A closed-loop platform for deploying multimodal llms in autonomous driving," *arXiv preprint arXiv:2402.01246*, 2024.