

# Internship Report - Code Generation with Vision Language Models for Robot arms application.

Tran Quang Minh , Luu Trong Hieu , Nguyen Cong Khanh , Nguyen Quang Trung 

*Department of Artificial Intelligence*

*FPT University — VietDynamic JSC*

*Ho Chi Minh City, Vietnam*

Emails: quantran102005@gmail.com, Luutronghieu0709@gmail.com, congkhanhtruongthi@gmail.com, trungnqse183108@fpt.edu.vn

**Abstract**—This report presents the internship experience of our team, who worked on a project titled "Code Generation with Vision Language Models for Robot arms application." The internship took place at VietDynamic JSC from September 2025 to December 2025. The primary objective of the project was to explore the capabilities of vision language models (VLMs) in generating code for robot arm applications. The report details the tasks undertaken, challenges faced, and the skills acquired during the internship. It also highlights the significance of VLMs in automating code generation and their potential impact on the robotics industry. We express our gratitude to VietDynamic JSC for providing this valuable learning opportunity. Code is available at: [GitHub/Code-gen-for-robot-arm-OJT-FALL-2025-FPT](#)

## I. INTRODUCTION

The rapid advancement of artificial intelligence (AI) and machine learning has led to the development of vision language models (VLMs) that can understand and generate human-like text. These models have shown remarkable capabilities in various natural language processing tasks, including code generation. The ability to generate code automatically has significant implications for the software development industry, particularly in specialized fields such as robotics. During our internship at VietDynamic JSC, we had the opportunity to work on a project focused on leveraging VLMs for code generation in robot arm applications. The project aimed to explore how VLMs can assist in automating the coding process, thereby improving efficiency and reducing the time required for software development in robotics. This report provides a comprehensive overview of our internship experience, including the tasks we undertook, the challenges we encountered, and the skills we developed. We also discuss the potential applications of VLMs in the robotics industry and their impact on future developments.

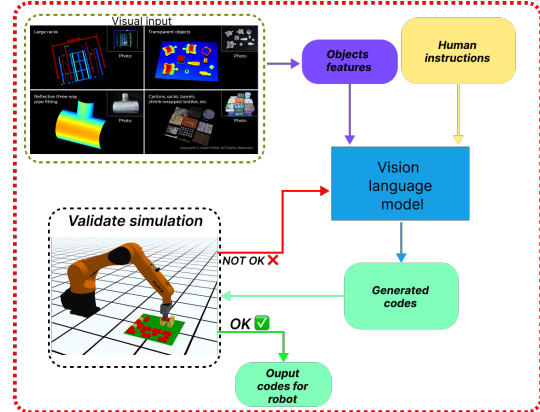
## II. RELATED WORK

Recent advancements in large language models (LLMs) have demonstrated their potential in various applications, including code generation for robotic systems. Notable works in this domain include Mu et al.'s RoboCodeX[1], which explores the use of LLMs to generate code for robotic tasks, showcasing the ability of these models to understand and execute complex instructions. Another significant contribution is the Robotic Programmer by Xie et al.[2], which focuses on

video-instructed policy code generation for robotic manipulation, highlighting the integration of visual inputs with LLMs to enhance robotic capabilities. These studies underscore the transformative potential of LLMs in automating and optimizing code generation for robotics, paving the way for more efficient and intelligent robotic systems. Other relevant works include the development of vision-language models (VLMs) like MobileVLM[3], which are designed to handle multimodal inputs, making them suitable for applications that require both visual and textual understanding. The integration of VLMs in robotics can significantly enhance the interaction between robots and their environments, enabling more sophisticated and context-aware behaviors.

## III. METHODOLOGY

### A. Overview



We leverage the capabilities of vision language models (VLMs) to generate code for robot arm applications. Combined with visual information from dedicated sensors like Mech-EYE we could enhance the understanding of the environment and improve the accuracy of the generated code. The overall pipeline consists of several key components: Mech-EYE 3D industrial camera, VLMs, and a simulator for validating the generated code without the need of physical hardware. The Mech-EYE camera captures high-resolution images and 3D point clouds of the robot's surroundings, providing essential visual context for the VLMs. The VLMs, such as MobileVLM or specialized models like RoboCodeX[1], are then employed to generate code based on the visual data and

specific task requirements. More comprehensive tasks such as video instructions for robotic manipulation are also considered [2]. Finally, the generated code is executed in a controlled environment built with ROS2 and GAZEBO to validate its functionality and performance.

#### B. Data Collection

### IV. RESULTS AND DISCUSSION

#### V. CONCLUSION

#### APPENDIX A

#### DETAILS OF THE CODE GENERATION PIPELINE

#### DUMP TEXT

#### APPENDIX B

#### ADDITIONAL FIGURES

#### DUMP TEXT

#### REFERENCES

- [1] Y. Mu, J. Chen, Q. Zhang, S. Chen, Q. Yu, C. Ge, R. Chen, Z. Liang, M. Hu, C. Tao *et al.*, “Robocodex: Multimodal code generation for robotic behavior synthesis,” *arXiv preprint arXiv:2402.16117*, 2024.
- [2] S. Xie, H. Wang, Z. Xiao, R. Wang, and X. Chen, “Robotic programmer: Video instructed policy code generation for robotic manipulation,” *arXiv preprint arXiv:2501.04268*, 2025.
- [3] X. Chu, L. Qiao, X. Lin, S. Xu, Y. Yang, Y. Hu, F. Wei, X. Zhang, B. Zhang, X. Wei *et al.*, “Mobilevlm: A fast, strong and open vision language assistant for mobile devices,” *arXiv preprint arXiv:2312.16886*, 2023.