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

Internship Presentation - Fall 2025

Table of Contents

Table of Contents

Motivation

- The rapid advancement of AI and machine learning has led to development of vision language models (VLMs)
- VLMs show remarkable capabilities in natural language processing tasks, including code generation
- Automated code generation has significant implications for software development in robotics
- Opportunity to explore VLMs for robot arm applications at VietDynamic JSC

Project Objectives

Primary Objective

Explore the capabilities of vision language models (VLMs) in generating code for robot arm applications

Secondary Objectives

- Automate the coding process for robotics
- Improve efficiency and reduce development time
- Integrate visual understanding with code generation
- Validate generated code in simulation environments

Internship Details

Duration: September 2025 - December 2025
Location: VietDynamic JSC, Ho Chi Minh City
Team Members:

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

GitHub Repository: `https://github.com/Minhtrna/Code-gen-for-robot-arm-OJT-FALL-2025-FPT`

Table of Contents

State of the Art

Key Research Areas

- RoboCodeX [1]: LLMs for robotic task code generation
- Robotic Programmer [2]: Video-instructed policy code generation
- MobileVLM [3]: Multimodal vision-language models

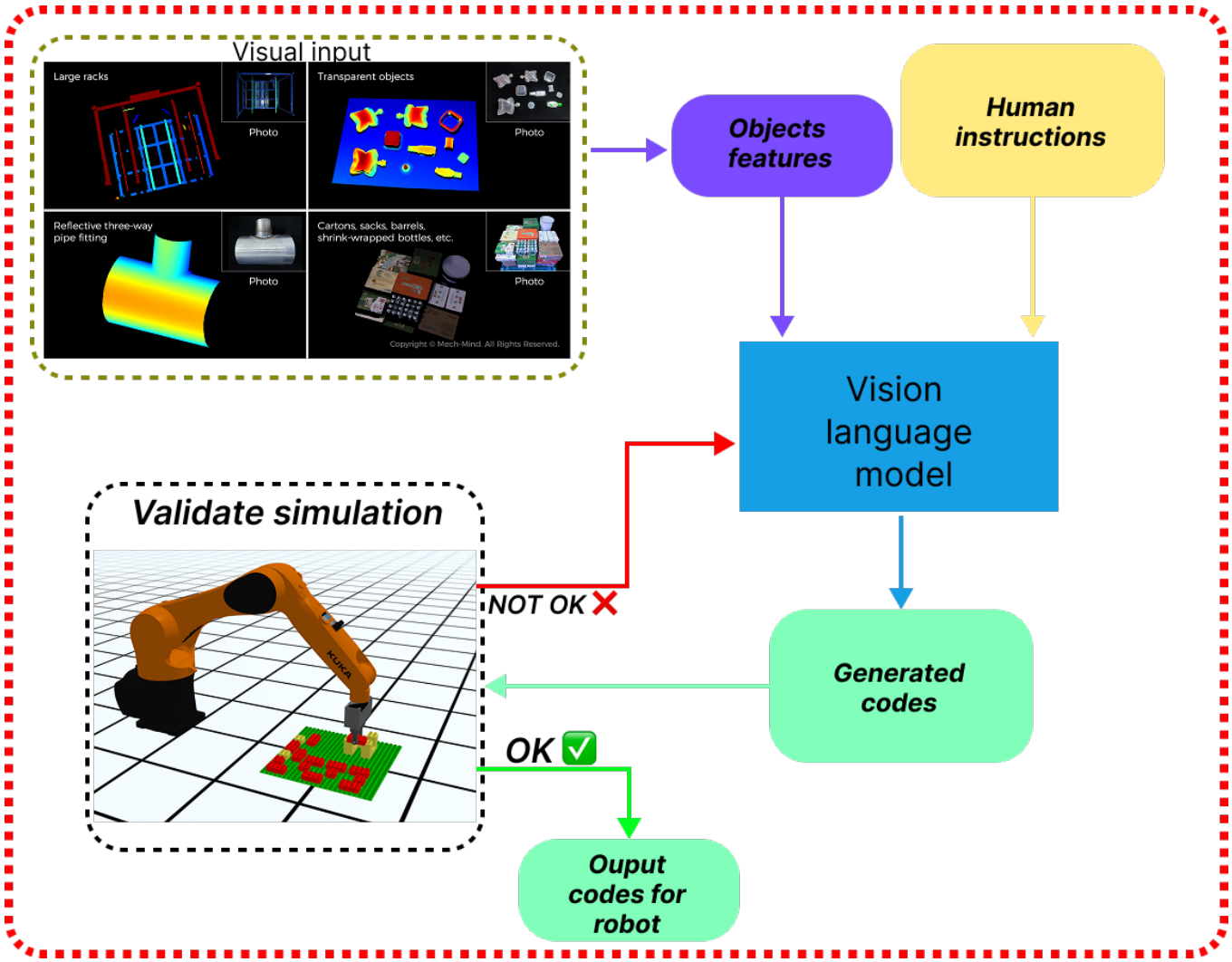
Research Gap

Integration of 3D visual sensors with VLMs for industrial robot arm applications remains underexplored

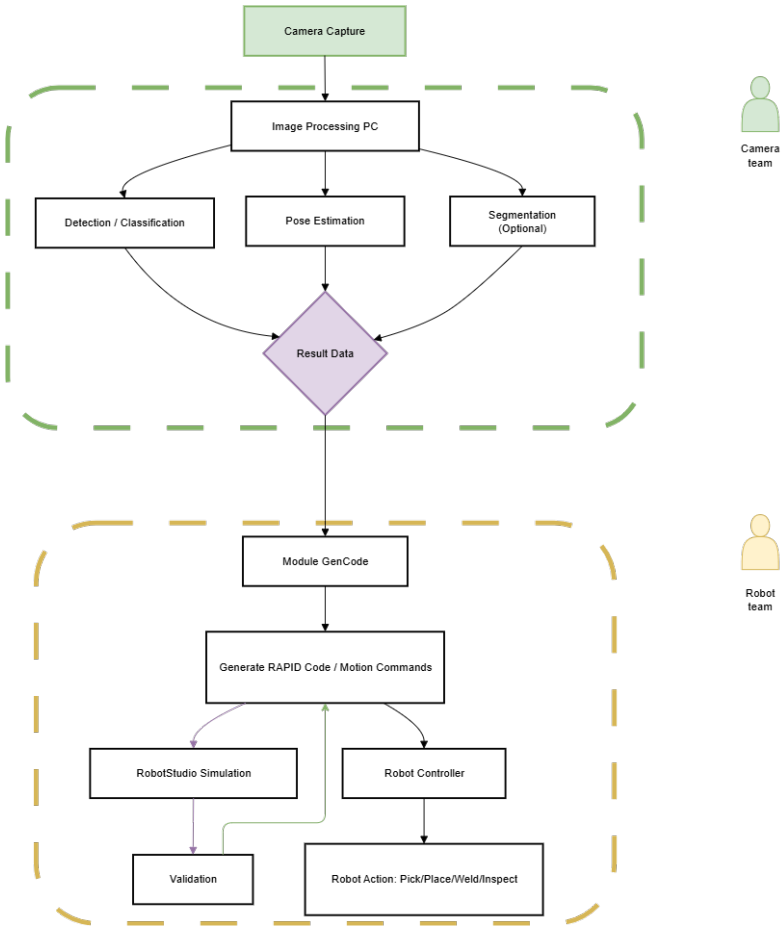
⇒ **Our work addresses this gap by combining Mech-EYE 3D cameras with VLMs.**

Table of Contents

System Overview



Pipeline Components



- **Mech-EYE 3D Camera:** Captures high-resolution images and 3D point clouds
- **Vision Language Models:** Process visual data and generate robot code
- **ROS2 + Gazebo:** Simulation environment for code validation
- **Integration:** Seamless pipeline from perception to execution

Technical Stack

Hardware Components

- Mech-EYE 3D Industrial Camera - High-precision visual sensing
- Robot arm simulation platform

Software Framework

- ROS2 - Robot Operating System for communication
- Gazebo - Physics simulation environment
- Vision Language Models - MobileVLM, RoboCodeX variants

Table of Contents

Current Progress

GitHub Repository: `https://github.com/Minhtrna/Code-gen-for-robot-arm-OJT-FALL-2025-FPT`

Key Achievements:

- Successfully integrated Mech-EYE 3D camera with ROS2
- Developed VLM-based code generation pipeline
- Created simulation environment for validation
- Established end-to-end workflow

3D Point Cloud Visualization

Figure: Mech-EYE camera output

Gazebo Simulation

Figure: Robot arm simulation

Technical Contributions

Data Collection Pipeline

- Automated capture of 3D visual data
- Integration with robot workspace mapping
- Real-time processing capabilities

Code Generation Framework

- VLM adaptation for robotics domain
- Context-aware code generation
- Safety constraint integration

Table of Contents

Challenges and Limitations

- Model Accuracy: VLMs require fine-tuning for robotics-specific tasks
- Real-time Performance: Balancing accuracy with processing speed
- Safety Constraints: Ensuring generated code follows safety protocols
- Hardware Integration: Synchronizing 3D vision with motion planning

$\Rightarrow$ **Future work will focus on addressing these technical challenges.**

Impact and Applications

Industrial Applications

- Automated assembly line programming
- Pick-and-place operation optimization
- Quality control integration

Research Contributions

- Novel integration of 3D vision with VLMs
- Open-source framework for community use
- Validation methodology for generated code

Table of Contents

Key Takeaways

Technical Achievements

Successfully demonstrated VLM-based code generation for robot arms with 3D visual input

Learning Outcomes

- Deep understanding of vision-language model integration
- Practical experience with ROS2 and Gazebo simulation
- Industry collaboration skills at VietDynamic JSC

Future Directions

Real-world deployment and performance optimization

Acknowledgments

Special Thanks

- VietDynamic JSC for providing this valuable learning opportunity
- FPT University for academic support and guidance
- Internship supervisors and mentors
- Team members for collaborative effort

Contact:

- quantran102005@gmail.com
- Luutronghieu0709@gmail.com
- congkhanhtruongthi@gmail.com
- trungnqse183108@fpt.edu.vn

Thank You!

Questions & Discussion

References I

[1] Yao Mu et al. "Robocodex: Multimodal code generation for robotic behavior synthesis". In: *arXiv preprint arXiv:2402.16117* (2024).

[2] Senwei Xie et al. "Robotic programmer: Video instructed policy code generation for robotic manipulation". In: *arXiv preprint arXiv:2501.04268* (2025).

[3] Xiangxiang Chu et al. "Mobilevlm: A fast, strong and open vision language assistant for mobile devices". In: *arXiv preprint arXiv:2312.16886* (2023).