

Multimodal Machine Learning Models Using Zero-Shot Learning to Control Robots [†]

Vladimir Kotev ^{1,2}, Ivan Ivanov ^{3,*} , Kosuke Kakikoshi ⁴, Stanislav Georgiev ² and Ken'ichi Yano ⁴

¹ Mathematics Department, Applied Informatics and Statistics Faculty, University of National and World Economy, 1756 Sofia, Bulgaria; kotev@unwe.bg

² Mechatronics Department, Institute of Mechanics-Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria; stan@initiate.ai

³ Mathematical Analysis and Differential Equations—(MADE), Faculty of Applied Mathematics and Informatics, Technical University of Sofia, 1000 Sofia, Bulgaria

⁴ Intelligent Robotics Laboratory, Department of Mechanical Engineering, Faculty of Engineering, Mie University, Tsu 514-8507, Japan; kakikoshi.kosuke@robot.mach.mie-u.ac.jp (K.K.); yanolab@robot.mach.mie-u.ac.jp (K.Y.)

* Correspondence: ivan.s.ivanov@tu-sofia.bg; Tel.: +359-898476687

[†] Presented at the 15th International Scientific Conference TechSys 2026—Engineering, Technologies and Systems, Plovdiv, Bulgaria, 14–16 May 2026.

Abstract

Substantial growth has been observed in large language models (LLMs), which are increasingly applied across various fields. The integration of vision–language models trained on Internet-scale data into end-to-end robotic control systems to enhance generalization and enable emergent semantic reasoning is studied. A small-size mobile robot with a 6 DoF arm is designed and developed in order to study and test a control approach utilizing vision–language models trained on Internet-scale data. The current work studies whether an LLM (GPT-4) can directly predict sequences of commands for mobile robot control to execute given tasks. An algorithm for measuring distance among objects is developed because the robot has only one camera and there are no other sensors for distance measurement. The performance of a single task-agnostic prompt, devoid of in-context examples, motion primitives, or external trajectory optimizers, in executing various tasks is evaluated. Furthermore, a framework that leverages multimodal GPT-4 to enhance task planning by integrating natural language instructions with robot visual perceptions is proposed. Indoor experiments show that the robot could execute different tasks such as moving to various objects that surround us in rooms and offices. Users write/input commands on the PC and the robot executes them.

Keywords: multimodal machine learning; zero-shot learning; robot control; vision–language models; GPT-4 robotics



Academic Editors: Nikola G. Shakev,
Valyo N. Nikolov, Nikolay R.
Kakanakov and Sevil A.
Ahmed-Shieva

Published: 4 August 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Traditionally, control of robots refers to predefined models and extensive task-specific training, which limit their flexibility and adaptability in dynamic environments [1–4]. However, recent advancements in machine learning (ML) are driving the development of more sophisticated and versatile control systems for robots [2–5]. ML techniques, including reinforcement learning and deep learning, enhance robotic control by enabling robots to learn from data and improve their decision-making capabilities autonomously. These techniques empower robots to process sensory data, recognize patterns, and make real-time

adjustments, which are critical for complex tasks such as navigation, object recognition, and obstacle avoidance [2–5]. As a result, ML offers a promising path toward creating more adaptive and intelligent robotic systems.

The research in multimodal learning, especially in computer vision, has advanced rapidly over the past decade [3,5,6]. The potential of multimodal data streams and deep learning algorithms has led to the widespread use of deep multimodal learning, where models process and analyze various forms of real-world data, such as visual and textual content. Six key areas in deep multimodal learning are studied [6]: data representation, fusion techniques, multitask learning, alignment, transfer learning, and zero-shot learning. Additionally, it surveys current applications, presents benchmark datasets, and discusses the limitations and challenges of the field, offering insights for future research directions [6].

Most recently, research on the integration of large language models (LLMs) appears [7–13]. The application of LLMs in robotic control systems offers promising advancements in human–robot interaction and adaptability. Unlike traditional robots that rely on explicit programming, LLMs enable robots to understand and execute complex instructions in natural language, enhancing their ability to interact with humans intuitively [7–13]. This capability allows robots to dynamically adapt to changing environments, learn new tasks from context or demonstrations, and respond to user queries conversationally. In [7] is studied and considered how vision–language models trained on Internet-scale data can be incorporated directly into end-to-end robotic control to boost generalization and enable emergent semantic reasoning. A study on applications and limitations of four methods in several systems based on recent methods that integrate LLMs on robot has been done [8]. The potential of LLMs as high-level planners in robotics has recently been demonstrated in [9], but it is often assumed that they lack the capability for low-level trajectory planning. It is investigated whether GPT-4 can directly predict end-effector poses for manipulation tasks using only object detection and segmentation vision models. The performance of a single task-agnostic prompt was tested across 26 real-world tasks without using in-context examples or trajectory optimizers. It was found that LLMs do indeed possess an understanding of low-level robot control sufficient for a range of common tasks, and that they can additionally detect failures and then re-plan trajectories accordingly. Using an Internet-scale pre-trained vision and language foundation model as the backbone of a robot policy for producing personalized task plans to solve household multi-object rearrangement tasks is considered [10]. VLMs: using them for assistive human–robot collaboration in multi-object rearrangement tasks. Initial results on a one-step table setting task that shows proof-of-concept for this method are presented. A comprehensive overview of the integration of large language models (LLMs) and multimodal LLMs into various robotic tasks is provided in this study. Also, in [11], another novel framework is proposed, utilizing multimodal GPT-4 to enhance embodied task planning through the combination of natural language instructions with robot visual perceptions. Results derived from diverse datasets indicate that robot performance in embodied tasks is significantly improved by GPT-4. The understanding of LLM-centric embodied intelligence is enriched by this extensive survey and evaluation of LLMs and multimodal LLMs, offering forward-looking insights into improving Human-Robot-Environment interaction and thereby bridging existing gaps in communication and task execution.

The aim of the research is to study the application of multimodal machine learning models combined with zero-shot learning (ZSL) techniques for robotic control. By integrating multiple data modalities such as textual inputs, these models aim to enhance the adaptability and functionality of robots in diverse and dynamic environments. Zero-shot learning enables the model to recognize and respond to new, previously unseen commands or tasks without additional training. It is very important and useful that the robot plans its

trajectory and motion to the given object. This approach has the potential to significantly improve the efficiency and versatility of robotic systems.

2. An Approach to Controlling a Mobile Robot with Multimodal Machine Learning Models

A small-size mobile platform with a 6 DoF arm is designed and developed (Figure 1) in order to study the application of multimodal machine learning models combined with zero-shot learning (ZSL) techniques for robotic control. By integrating multiple data modalities such as textual inputs, these models aim to enhance the adaptability and functionality of robots in diverse and dynamic environments. Zero-shot learning enables the model to recognize and respond to new, previously unseen commands or tasks without additional training. This approach has the potential to significantly improve the efficiency and versatility of robotic systems.



Figure 1. The developed mobile platform with a 6 DoF arm.

2.1. Development of a Mobile Robotic Platform

A four-wheel mobile platform is designed and developed (Figure 1) to test a robot control approach based on vision–language models trained on Internet-scale data. It consists of four 12 V brushless DC motors, a driver for motors, micro controller boards Arduino and Jetson nano, Hama QHD camera, as well as a Braccio robot arm. Using a camera, Arduino, and Jetson Nano together in a small robot combines their strengths effectively. We should highlight that the robot only uses the camera sensor to measure distance to various objects.

The camera captures visual data for computer vision, the Arduino manages real-time control of hardware components, and the Jetson Nano processes the visual data and runs machine learning algorithms (Figure 2). This setup enables the robot to see its environment, analyze the data, and perform actions based on that analysis, making it ideal for testing ML control algorithms and computer vision applications.

Arduino Uno (Arduino, Boston, MA, USA) is a microcontroller board based on the ATmega328P (datasheet). It has 14 digital input/output pins (of which 6 can be used as PWM outputs), 6 analog inputs, a 16 MHz ceramic resonator (CSTCE16M0V53-R0), a USB connection, a power jack, an ICSP header and a reset button. It contains everything needed to support the microcontroller; simply connect it to a computer with a USB cable or power it with an AC-to-DC adapter or battery to get started [12]. The NVIDIA® Jetson Nano™ Developer Kit (NVIDIA, Santa Clara, CA, USA) is a compact, affordable AI computer designed for modern AI workloads. It features a 128-core Maxwell GPU, a quad-core

ARM Cortex-A57 CPU, 4 GB LPDDR4 RAM, and a microSD slot for storage. Connectivity options include Gigabit Ethernet, USB 3.0 ports, HDMI, and DisplayPort. It has a 40-pin GPIO header and MIPI CSI-2 camera connectors [13]. Running on NVIDIA's JetPack SDK, it supports AI frameworks like TensorFlow and PyTorch. Suitable for AI, robotics, IoT, and prototyping, it offers a balance of performance and power efficiency, making it ideal for developers and researchers. The TinkerKit Braccio is a fully operational robotic arm, controlled via Arduino [14]. It can be assembled in several ways for multiple tasks such as moving objects.

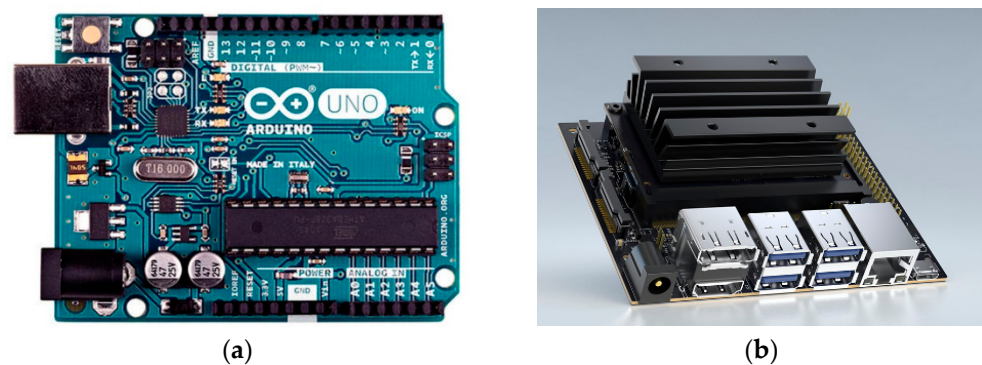


Figure 2. The main electronic hardware of the mobile platform with a 6 DoF arm: (a) Arduino Uno; (b) Jetson Nano.

2.2. Architecture of the Mobile Robotic Platform

In our study, the developed mobile robot utilizes large language models (LLMs) to process natural language instructions and generate actionable plans in the form of detailed step-by-step procedures or executable code. The architecture of the system consists of three interconnected modules (Figure 3).

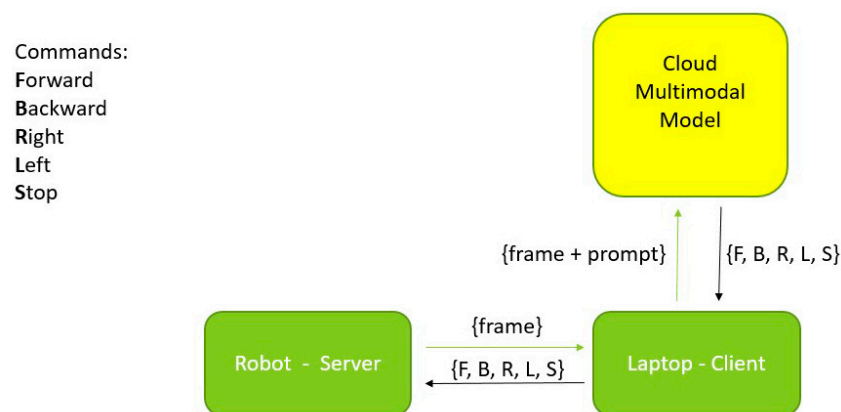


Figure 3. The architecture of the developed mobile platform with a 6 DoF arm.

The first module is the client, which is represented by the laptop. Users input their tasks through this interface, specifying what they want the robot to do.

The third module is the cloud, which houses the multimodal model, such as ChatGPT-4 in our research. When the user provides input, this information is sent to the cloud. The multimodal model processes the input to generate a description of the desired tasks and returns low-level commands to the laptop. These commands are translated from the task descriptions into specific instructions for the robot, which include: 'F' for forward, 'B' for backward, 'L' for left, 'R' for right, and 'S' for stop. The robot then executes these commands to carry out the tasks specified by the user.

3. Robot Software

The control software for the robot is developed using Python 3.12.0, as depicted in Figure 4. The software operates on two distinct levels: low-level control and high-level task management. The low-level control involves direct manipulation of the robot's movement, including commands to move forward, reverse, stop, turn left, or turn right. This is achieved by adjusting the wheel movements accordingly.

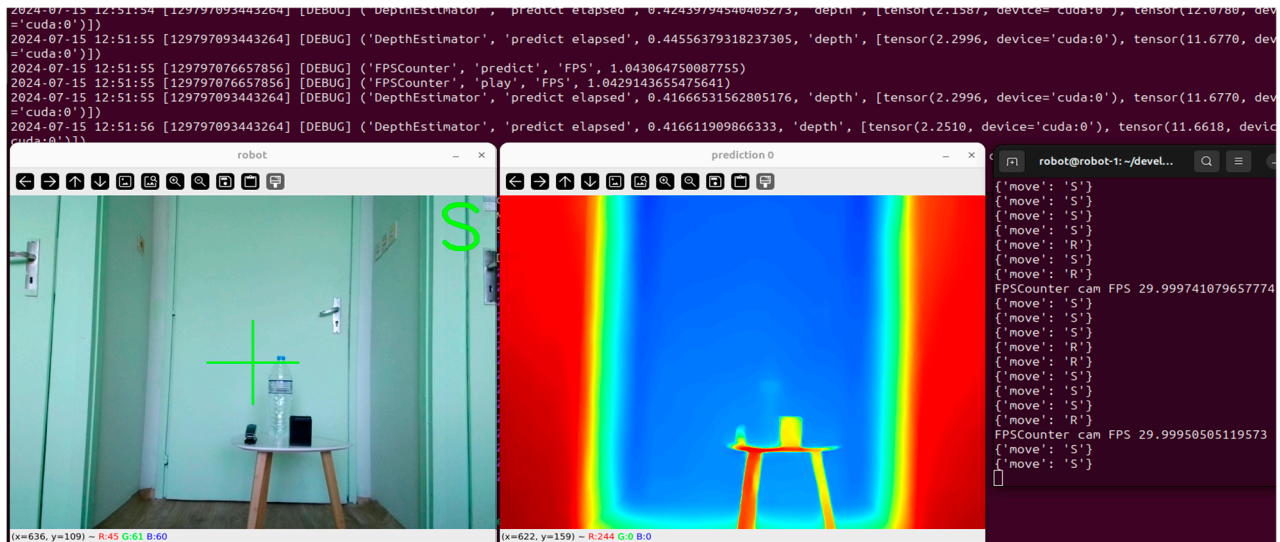


Figure 4. Robot software. The left picture shows a view from the robot's camera and models prediction. The next picture in the middle shows depth estimation.

At the high level, the software interacts with a multimodal model through API calls. The user's input is processed in two parts. The first part is task-agnostic and remains consistent; it instructs the multimodal model that it will receive camera frames and must respond with a command corresponding to one of the following actions: forward ('F'), backward ('B'), stop ('S'), turn left ('L'), or turn right ('R'). The second part is task-specific and is provided in natural language, describing the specific task to be performed.

These two prompts are concatenated and, along with the latest frame from the robot's camera, are sent to the cloud-based multimodal model. The model then returns a single-letter command, which the laptop translates into actionable instructions for the robot.

In our indoor experiments, the robot's task was to identify, move towards, and stop near a specified object within an office environment. The left side of Figure 4 illustrates the view from the robot's camera, where the next command is shown in green capital letters. For example, the letter 'S' indicates that the robot will stop. An algorithm for distance measurement among objects is employed due to the robot being equipped with only a single camera, as shown in the middle of Figure 4. Objects in close proximity to the robot are highlighted in red and yellow.

Our experimental results demonstrate that the robot is able to identify, approach, and stop near the target object with relatively high precision. However, the overall speed of operation is limited due to delays in each iteration of the process.

4. Conclusions

This study reveals the potential of combining multimodal machine learning models with zero-shot learning techniques for advanced robotic control. The resulting systems exhibit enhanced flexibility and intelligence, paving the way for more versatile robotic applications in diverse fields. The key insight is that multimodal foundational models enable robots to perform complex tasks using zero-shot learning with minimal prior

training. A mobile platform with a 6 DoF robot arm has been designed and developed to test robot control using language and vision models. It is equipped with only one camera sensor for distance measurement, and there are no other sensors for this purpose.

Different prompt inputs are written for executing various tasks. Experiments show that the robot successfully performs tasks such as locating objects specified in the prompt, moving to these objects, and stopping near them. The robot is able to recognize objects in rooms and reach the objects specified in the prompt. The precision is relatively good, but the speed is slow. There is some delay in each step of the robot's motion because each frame from the robot's camera is transmitted to the cloud, where the multimodal model sends the next command.

Zero-shot learning facilitates rapid deployment and adaptation in dynamic environments without extensive retraining. Moreover, the ability to translate high-level objectives into precise low-level commands allows for accurate task execution based on visual and textual data. Multimodal foundational models significantly advance robot control by enabling zero-shot learning and precise execution of low-level commands, thereby enhancing their utility in various applications.

Future research will focus on three main areas. The first will involve controlling the robot arm to perform different tasks by utilizing advanced algorithms to enhance precision. The second will include conducting experiments with various prompts to improve the performance and responsiveness of multimodal models in different scenarios. The third will be on fine-tuning local foundational models to increase their efficiency and adaptability for specific tasks and environments.

Author Contributions: Conceptualization, I.I. and S.G.; methodology, K.Y. and V.K.; software, S.G., I.I. and K.K.; validation, I.I., V.K., S.G. and K.Y.; resources, V.K., K.K. and I.I.; writing—original draft preparation, V.K.; writing—review and editing, I.I.; supervision, K.Y. and V.K. These authors contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research and Development Sector at the Technical University of Sofia.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: The authors would like to thank the Research and Development Sector at the Technical University of Sofia for the financial support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Siciliano, B.; Sciavicco, L.; Villani, L.; Oriolo, G. *Robotics: Modelling, Planning and Control*; Springer: Berlin/Heidelberg, Germany, 2008.
2. Sturmfels, J. *Approaches to Probabilistic Model Learning for Mobile Manipulation Robots*; Springer: Berlin/Heidelberg, Germany, 2007.
3. Corke, P. *Robotics, Vision and Control: Fundamental Algorithms in MATLAB*, 3rd ed.; Springer: Cham, Switzerland, 2017.
4. Liu, Z.; Liu, Q.; Xu, W.; Wang, L.; Zhou, Z. Robot learning towards smart robotic manufacturing: A review. *Robot. Comput.-Integr. Manuf.* **2022**, *77*, 102360. [[CrossRef](#)]
5. Alatabani, L.E.; Ali, E.S.; Mokhtar, R.A.; Khalifa, O.O.; Saeed, R.A. Robotics architectures based machine learning and deep learning approaches. In Proceedings of the 8th International Conference on Mechatronics Engineering (ICOM 2022), Kuala Lumpur, Malaysia, 9–10 August 2022; pp. 107–113. [[CrossRef](#)]
6. Bayouddh, K.; Knani, R.; Hamdaoui, F. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *Vis. Comput.* **2022**, *38*, 2939–2970. [[CrossRef](#)] [[PubMed](#)]

7. Zitkovich, B.; Yu, T. Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of the 7th Conference on Robot Learning*; PMLR: New York, NY, USA, 2023; pp. 2165–2183. Available online: <https://proceedings.mlr.press/v229/zitkovich23a.html> (accessed on 24 March 2026).
8. Google DeepMind. Demonstrating Large Language Models on Robots. In *Proceedings of the Robotics: Science and Systems 2023*, Daegu, Republic of Korea, 10–14 July 2023.
9. Kwon, T.; Palo, N.D.; Johns, E. Language Models as Zero-Shot Trajectory Generators. *IEEE Robot. Autom. Lett.* **2024**, *9*, 6728–6735. [CrossRef]
10. Newman, B.A.; Gupta, P.; Bisk, Y.; Kitani, K.; Admoni, H.; Paxton, C. Leveraging Vision and Language Models for Zero-Shot Personalization of Household Multi-Object Rearrangement Tasks. In *Proceedings of the Human-Large Language Model Interaction Workshop (HRI'24)*, Boulder, CO, USA, 11 March 2024. Available online: https://fomo-hri.github.io/workshop/hri24/papers/hllmi24_paper_9.pdf (accessed on 13 July 2026).
11. Wang, J. Large language models for robotics: Opportunities, challenges, and perspectives. *arXiv* **2024**, arXiv:2401.04334.
12. Arduino. Arduino Uno Rev3. Available online: <https://store.arduino.cc/products/arduino-uno-rev3> (accessed on 24 March 2024).
13. NVIDIA. Jetson Nano Developer Kit. Available online: <https://developer.nvidia.com/embedded/learn/get-started-jetson-nano-devkit#intro> (accessed on 24 March 2024).
14. Arduino. TinkerKit Braccio Robot. Available online: <https://store.arduino.cc/products/tinkerkit-braccio-robot> (accessed on 24 March 2024).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.