EmbodiedRDA: Connecting Foundation Models with the Physical World using Reconfigurable Drone Agents

Minghui Zhao[*][†], Kaiyuan Hou[*][†], Junxi Xia[+], Stephen Xia[+], Xiaofan Jiang[*]
[*]Columbia University, [+]Northwestern University

ABSTRACT

Foundation models excel in tasks such as content generation, zero-shot classifications, and reasoning. However, they struggle with sensing, interacting, and actuating in the physical world due to their dependence on limited sensors and actuators in providing timely contextual information or physical interactions. This reliance restricts the system's adaptability and coverage. To address these issues and create an embodied AI with foundation models (FMs), we introduce *Embodied Reconfigurable Drone Agent (EmbodiedRDA)*. EmbodiedRDA features a custom drone platform that can autonomously swap payloads to reconfigure itself with a diverse list of sensors and actuators. We designed FM agents to instruct the drone to equip itself with appropriate physical modules, analyze sensor data, make decisions, and control the drone's actions. This enables the system to perform a variety of tasks in dynamic physical environments, bridging the gap between the digital and physical worlds.

CCS CONCEPTS

• **Computing methodologies → Robotic planning**; • **Hardware → Sensor applications and deployments**; **Sensor devices and platforms**.

KEYWORDS

Embodied AI, Foundation Models, Drones

ACM Reference Format:
Minghui Zhao[*][†], Kaiyuan Hou[*][†], Junxi Xia[+], Stephen Xia[+], Xiaofan Jiang[*]. 2024. EmbodiedRDA: Connecting Foundation Models with the Physical World using Reconfigurable Drone Agents. In *The 30th Annual International Conference on Mobile Computing and Networking (ACM MobiCom '24), November 18–22, 2024, Washington D.C., DC, USA.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3636534.3698846

1 INTRODUCTION

Foundation models (FM), large language models (LLM) – a subset of FMs, and artificial general intelligence (AGI) promise significant technological and social impact through their adaptability to new applications and environments. While LLMs and FMs excel in human-like understanding and digital content generation, their capacity to react to real-world events and interact with the physical world remains relatively unexplored.

Recent works have incorporated FMs to control physical robotic systems via voice commands, utilizing vision-based sensors on the robots to perceive their surroundings. Others explore LLMs as human-like interfaces for internet-connected smart appliances, such as speakers, television, air conditioning, etc. However, these applications fall short of full autonomy in physical environments, unlike FMs' extensive capabilities in the digital domain. Several challenges persist: (1) limited adaptability to diverse tasks with fixed sensor/actuator configurations; (2) need for dense sensor deployments to cover events at arbitrary locations; (3) difficulty in processing varied sensor data formats and (4) dynamic nature of physical environments demanding real-time adaptability. These limitations highlight the gap between FMs' digital capabilities and their ability to fully operate in physical environments.

We propose EmbodiedRDA, an embodied AI system consisting of 1) a reconfigurable drone platform and 2) FM agents that leverage the drone platform to interact dynamically with the physical world. As shown in Figure 1a, the system operates as follows: upon receiving a user command, the FM instructs the drone to equip itself with appropriate modules. The drone then travels to the locations of interest given by FM to gather relevant information or perform actions. Subsequently, the FM agent generates and executes code to analyze the collected data, providing feedback to the user.

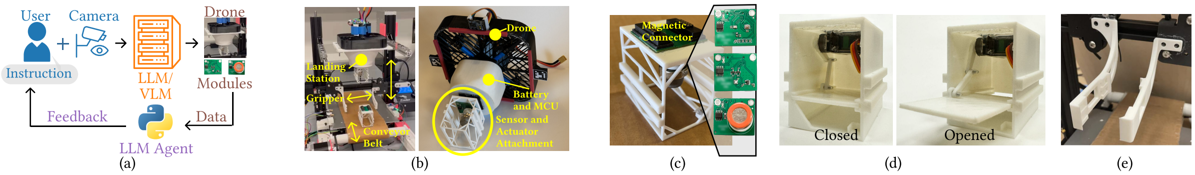
2 SYSTEM DESIGN

EmbodiedRDA implements the architecture proposed in [4], consisting of a customized drone platform capable of autonomously swapping payloads, an FM agent that empowers

[†]Both authors contributed equally to this research.

ACM MobiCom '24, November 18–22, 2024, Washington D.C., DC, USA

ACM ISBN 979-8-4007-0489-5/24/11…$15.00
https://doi.org/10.1145/3636534.3698846



(a) (b) (c) (d) (e)

Figure 1: (a) System workflow, (b) Customized drone, (c, d) Customized sensor and actuator modules, (e) Gripper for swapping sensor and actuation modules

tasks comprehension, drone command generation and data analysis, and a bird's eye view (BEV) camera.

Customized Drone. Figure 1b shows the custom-designed drone capable of autonomous payload reconfiguration. We used 3800 Kv brushless motors powered by a 4-cell 850mAh battery, tuned PID control parameters for stable flight while carrying a 50-gram payload. Additionally, we printed a protective guard to shield its propellers from objects and people. The drone can hover for 4 minutes, sufficient for executing most sensing and actuating tasks.

Sensor and Actuator Modules. The sensors (Figure 1c) and actuators (Figure 1d) are modularized with a unified enclosure. The sensor modules are designed following the modular architecture presented in [5]. The modules are connected to a Raspberry Pi on the drone through a standardized 24-pin magnetic connector, which provides power to the module and allows data communication to Raspberry Pi. Upon connection, Raspberry Pi loads sensor drivers automatically and stream the data to the FM agent.

Ground Station. We created the ground station by leveraging the chassis of the open-source Ender-3 3D printer [1], as it provides enough space for the drone to land, and precise mechanical control to transport sensors and actuator modules between a repository and the drone. Additionally, we designed the gripper (Figre 1e) to help the removal and attachment of modules on the drone. We replaced the build plate on the 3D printer with a conveyor belt as the repository of potential modules to be carried by the drone. To ensure the drone lands on the ground station accurately, we designed a funnel-shaped landing chute. The drone can land within the larger opening of the funnel and will automatically slide towards the narrower bottom, which aligns and latches onto the swappable module.

FM agent. When a user issues a voice command, this command is passed to a prompt template along with the BEV image into Large Language-and-Vision Assistant (LLaVA) visual-language model [2]. LLaVA outputs the names of potential objects in text in the scene that might be of interest, as well as the suitable sensor/actuation modules. The list of object names from LLaVA's output is then input into Grounding DINO [3], which outputs the bounding boxes of the specified objects that can be translated to locations in the top-down view image. After data is collected from the sensor module attached to the drone, an FM agent is used to generate code to process and analyze the collected data and provide a final conclusion to the task issued by the user.

3 DEMONSTRATION DESCRIPTION

In this demonstration, we will showcase EmbodiedRDA performing four tasks, each representing a different task category as defined in [4]:

Where is the warmest place to sit? In this task, we expect the drone to take the temperature sensor from ground station, with a succession of hovering over potential locations within the scene identified by the FM agent, and finally tell which place is warmest among the candidate locations.

Alert me if there is a chemical spill. In this task, the drone should take a gas sensor once there is liquid spill. EmbodiedRDA will determine if the liquid is chemical spill or not.

Where is my key? This task demonstrates EmbodiedRDA's capability in locating small objects using the drone camera for close-up inspection of potential areas identified by the FM Agent.

Bring the medicine to the table. This task illustrates the actuation capability of EmbodiedRDA as a drone-based personal assistant.

For safety during the demonstration, our drone features a fully enclosed protective cage to prevent propeller contact with the environment. Additionally, all flight operations will be strictly limited to a designated zone within the demo area.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under Grant Numbers CNS-1943396, and COGNISENSE, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Columbia University, Northwestern University, NSF, SRC, DARPA, or the U.S. Government or any of its agencies.

REFERENCES

[1] 2023. Creality3D Ender-3, a fully Open Source 3D printer perfect for new users on a budget. https://github.com/Creality3DPrinting/Ender-3.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485* (2023).

[3] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[4] Minghui Zhao, Junxi Xia, Kaiyuan Hou, Yanchen Liu, Stephen Xia, and Xiaofan Jiang. 2024. RASP: A Drone-based Reconfigurable Actuation and Sensing Platform for Engaging Physical Environments with Foundation Models. arXiv:2403.12853 [cs.RO] https://arxiv.org/abs/2403.12853

[5] Minghui Zhao, Stephen Xia, Jingping Nie, Kaiyuan Hou, Avik Dhupar, and Xiaofan Jiang. 2023. LegoSENSE: An Open and Modular Sensing Platform for Rapidly-Deployable IoT Applications. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*. 367–380.