Leveraging Foundation Models for One-Shot Learning via HRC and Aerial-Ground Multi-Robot Collaboration

Haokun Liu*, Moju Zhao, Member, IEEE

***Abstract*— Foundation models, including Large Language Models (LLMs) and Vision-Language Models (VLMs), offer new capabilities for improving robotic autonomy. This paper presents two independent approaches for applying foundation models to robotic task execution. The first approach employs LLM-driven task decomposition and teleoperation-based human-robot collaboration, enabling one-shot learning and rapid refinement of motion primitives for high-difficulty tasks. The second approach leverages an aerial robot for top-down perception, where VLMs process the captured images to support ground robots in manipulation and navigation. Experimental results demonstrate that LLM-driven task decomposition significantly improves robot adaptability to novel tasks while VLM-assisted multimodal perception enhances task-specified reasoning and scene understanding.**

I. Introduction

In recent years, foundation models, particularly Large Language Models (LLMs) and Vision-Language Models (VLMs), have significantly advanced the field of robotics [1], [2]. Their ability to process natural language [3] and visual information [4] enables robots to achieve higher levels of autonomy and intelligence. However, applying these models to complex robotic tasks remains a challenge due to the lack of structured task decomposition and the difficulty in integrating multimodal perception with low-level motion execution [5].

This paper explores foundation model-driven task execution by combining two independent research directions. The first study, illustrated in Fig. 1, introduces an LLM-based task decomposition approach that segments high-level tasks into motion primitives. This method incorporates teleoperation to refine and adapt motion primitives, enabling one-shot learning for high-difficulty tasks [6]. The second study, illustrated in Fig. 2, develops an aerial-ground robotic system where aerial drones capture top-view images, processed by VLMs for object detection, scene understanding, and semantic labeling. These visual data assist the ground robot in object manipulation, obstacle avoidance, and path planning. Meanwhile, LLMs decompose high-level instructions into structured sub-tasks and perform reasoning-based tasks such as constructing and updating a global semantic map, enabling efficient navigation and execution of complex tasks.

Although these studies address different aspects of robotic task execution, they share a common goal: leveraging foundation models to enhance robot adaptability and autonomy.

Haokun Liu, Moju Zhao are with DRAGON Lab at Department of Mechanical Engineering, The University of Tokyo, Tokyo, 113-8654, Japan (e-mail: {haokun-liu, chou}@dragon.t.u-tokyo.ac.jp.

Corresponding author: Haokun Liu

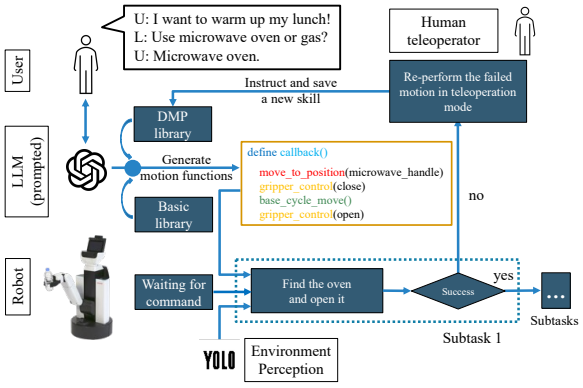


Fig. 1. An overview of an LLM-based Human-Robot Collaboration System, featuring user interaction, a basic library for pre-programmed motion functions, and a DMP library for adaptive motion function generation and storage to accomplish a complex real-world task (e.g., "warm up my lunch").

By presenting them together, we provide complementary perspectives on how LLMs and VLMs can improve both decision-making and perception in diverse robotic systems.

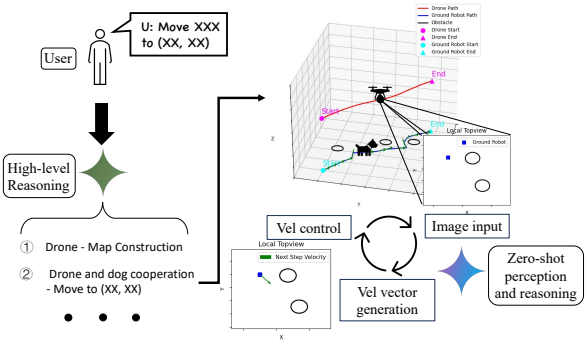


Fig. 2. Overview of the proposed aerial-ground robot cooperation system. For the task 'move to (XX, XX)', the system utilizes zero-shot perception and reasoning to enable the aerial robot to construct a semantic map according to the local top-view map of the environment. This semantic map is then used to generate motion planning for the ground robot.

The contributions of this paper are derived from two independent research directions, each exploring the role of foundation models in robotic task execution. Specifically:

- We demonstrate the effectiveness of LLM-driven task decomposition for complex robotic tasks by breaking them into motion primitives. Furthermore, we introduce a teleoperation-based Human-Robot Collaboration (HRC) mechanism that allows real-time refinement and substitution of motion primitives, enhancing task flexibility and adaptability.
- We propose a multimodal robot task execution framework that integrates LLM for high-level reasoning and

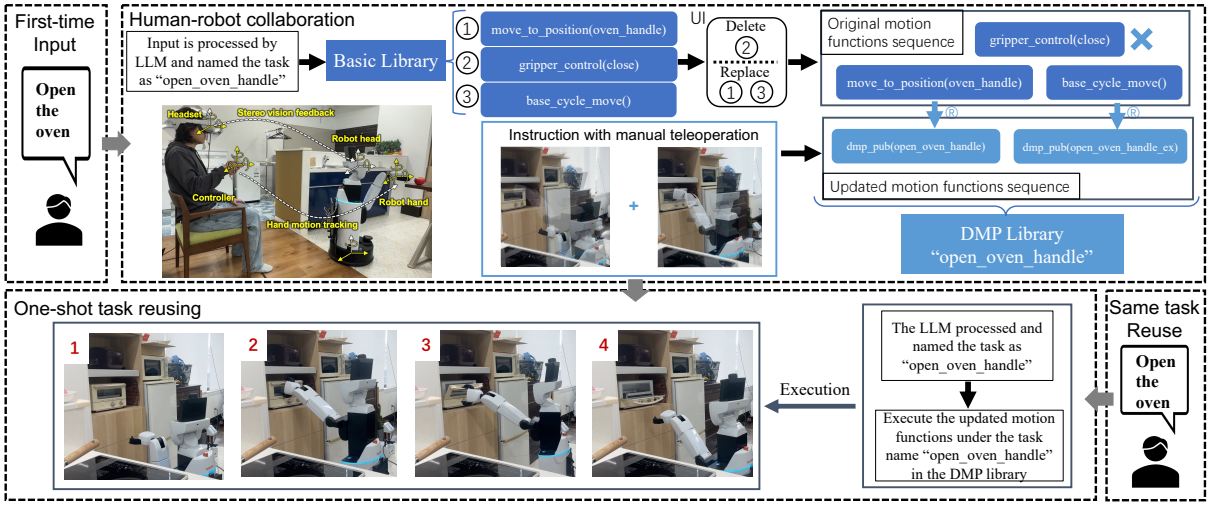


Fig. 3. An overview of the LLM-based autonomy with Human-Robot Collaboration (HRC) in sub-task (short-horizon task). The LLM processes user input to select motion functions from the basic Library. These selected motions are subsequently modified through the user interface with teleoperation. The updated motion functions are stored in the DMP Library with a specific name such as "open_oven_handle" (The LLM captures the action "open" and the target "oven_handle", then integrates them as "open_oven_handle") for future application (same task re-input or reusing in the long-horizon task), resulting in successful one-shot task execution.

VLM for perception. This framework enables zero-shot object detection, semantic-aware navigation, and precise manipulation, significantly improving autonomous task execution.

- We validate the advantages of foundation model-driven robotics by demonstrating both human-robot and multi-robot collaboration in complex task scenarios, highlighting the synergy between AI reasoning and physical robot execution.

II. Methodology

A. Task Decomposition and Motion Primitives Replacement with Human-Robot Collaboration

To enable the execution of high-difficulty robotic tasks, we employ an LLM-driven task decomposition approach. The LLM interprets high-level user commands and breaks them down into structured sub-tasks, aligning each step with the robot's available skills.

Each sub-task is executed using motion primitives, which serve as parameterized low-level actions. We utilize Dynamic Movement Primitives (DMP) [7] to store and generate motion trajectories, ensuring adaptability and reusability. Initially, an operator provides teleoperation-based demonstrations for new tasks, allowing the system to refine and store motion primitives for future autonomous execution. Once a motion primitive has been recorded, it can be reused in subsequent executions of the same task without requiring further instruction. This enables one-shot learning for high-difficulty robotic tasks. The workflow is illustrated in Fig. 3.

B. Multimodal Perception and Multi-Agent Robots Collaboration

To enhance task execution and robot collaboration, we develop a multimodal aerial-ground robotic system where VLM and LLMs play complementary roles. The workflow is illustrated in Fig. 4.

The LLM is responsible for high-level task planning and decomposition. It processes user instructions, breaks down complex tasks into a sequence of motion primitives, and assigns execution steps to the aerial and ground robots accordingly. For global map construction, since the aerial drone relies on a global map for navigation, the first sub-task in the execution pipeline is to generate it. The LLM integrates information from VLM-provided local maps to infer the overall environment structure, enabling the aerial robot to navigate and assist the ground robot in task execution.

The VLM is responsible for zero-shot object detection and semantic scene understanding, enabling adaptive perception for real-world task execution. It identifies objects in the environment, assigns semantic labels (e.g., *target, obstacle, main actor*), and provides structured scene understanding to guide motion primitives. The VLM primarily operates within individual motion primitives, assisting with perception and environment-aware execution.

III. Experiments and Results

A. Evaluation of Task Decomposition and Motion Primitives Replacment with Human-Robot Collaboration

To evaluate the effectiveness of our task decomposition and motion primitive approach, we conducted experiments involving a variety of tasks, including both zero-shot and one-shot scenarios. We assessed the performance by measuring the task completion success rate. The tasks include 4 short-horizon task (Easy) - 'Put&Stack', 'Open microwave', 'Open oven (HRC)', 'Open cabinet (HRC)' and 3 long-horizon task - 'Clean table' (Medium), 'Warm up apple' (Hard), and 'Roast apple (HRC)'.

- **Zero-Shot Tasks:** We tested the robot's ability to perform tasks without any prior demonstration or fine-tuning. The results demonstrated the generalizability of our pre-defined motion primitives.
- **One-Shot Tasks (HRC):** We evaluated the robot's adaptability to novel tasks through teleoperation-based

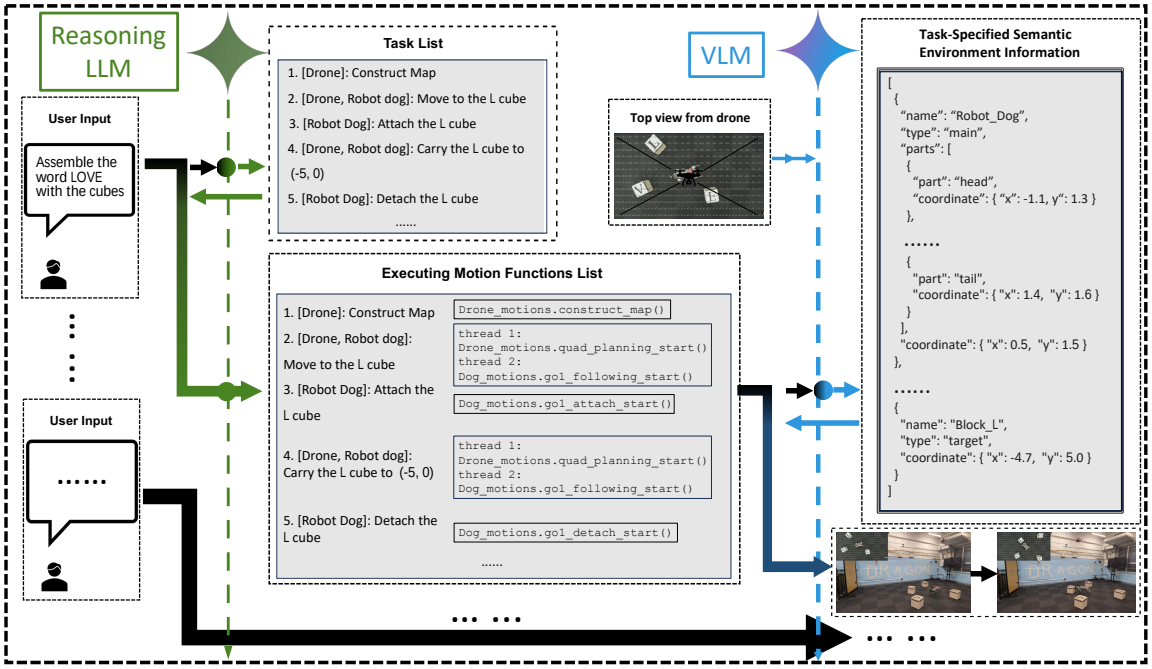


Fig. 4. An workflow of the foundation model-driven aerial-ground heterogeneous robotics system for a real-world long-horizon task. This process encompasses sub-task identification, motion function selection, environment perception integration, and robot motions generation.

motion primitive updates. The experiments showed a significant improvement in task success rate after a single demonstration.

The result in Table I indicates that LLM-based task decomposition with motion primitives replacement enables efficient adaptation in dynamic environments, making it suitable for real-world deployment in household automation.

TABLE I

EXECUTABILITY, FEASIBILITY, AND SUCCESS RATES OF LLM-BASED AUTONOMY AND HUMAN-ROBOT COLLABORATION

Tasks	Num of trials	Executability	Feasibility	Success rate
Put&Stack	23	100.0%	100.0%	91.3%
Open microwave	23	100.0%	100.0%	82.6%
Open oven (HRC)	23	100.0%	100.0%	91.3%
Open cabinet (HRC)	23	100.0%	100.0%	87.0%
Clean table	23	100.0%	95.7%	87.0%
Warm up apple	23	100.0%	100.0%	60.9%
Roast apple (HRC)	23	95.7%	87.0%	56.5%
Total	161	99.4%	97.5%	79.5%

B. Evaluation of Aerial-Ground Robots System

The experiments for the aerial-ground robots system were more extensive, covering various aspects of the system's performance.

1) Effectiveness of GridMask in VLM Fine-Tuning: To enhance the performance of VLM in describing environments and improving their integration within multimodal systems, we propose a GridMark-based fine-tuning method designed to improve 2D perception. The appearance of the GridMark is depicted in Fig. 9.

We investigated the impact of using GridMask during the fine-tuning of the VLM. The results indicated that GridMask significantly improved the robustness and accuracy of the

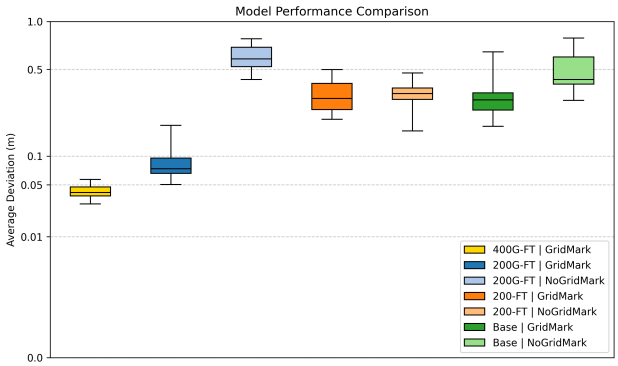


Fig. 5. Comparison of VLM accuracy with and without GridMask. The values show the average deviation of objects position. The number 200 or 400 means the datasets number, G-FT or -FT means if the GridMask is used for fine-tuning, GridMask or NoGridMask means if the GridMask is used for deployment.

VLM in parsing aerial images. The result is shown in Fig. 5. The adapter for Gemini's fine-tuning procedure is set to 4.

2) Accuracy of Target Localization by VLM: We evaluated the precision of the VLM in localizing target objects from aerial images. In our system, VLM provides local object coordinates, which are then transformed into global positions using the aerial robot's SLAM-estimated pose (orientation is kept to 0). To validate the accuracy of this approach, we compare the SLAM-based global localization with ground-truth positions obtained from a motion capture system. Fig. 6 presents the error analysis. In (a), we first assess the aerial robot's SLAM position errors compared to Mo-cap to understand the reliability of SLAM-based localization. In (b), we calculate the global position errors of the ground robot using both SLAM and Mo-cap reference positions. Finally,

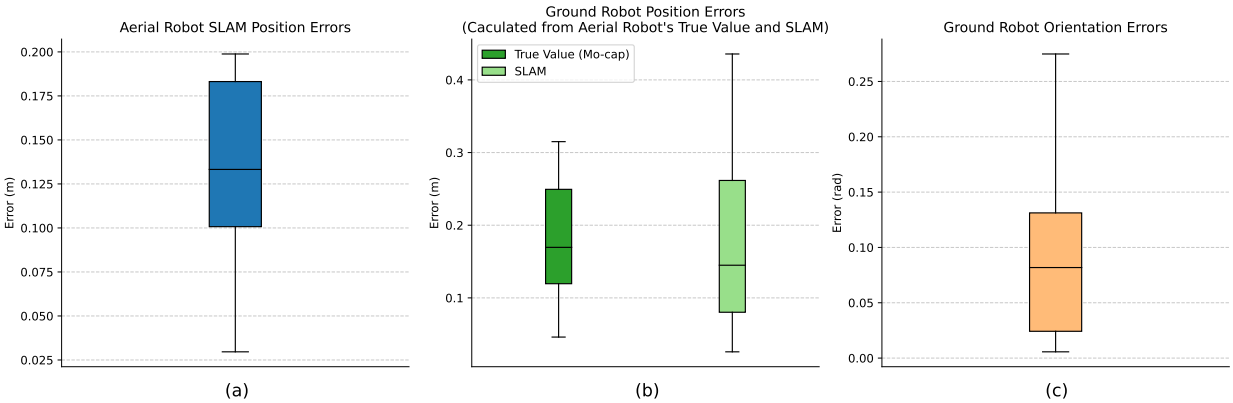


Fig. 6. Global localization error analysis of the ground robot based on aerial robot positioning data.

(c) illustrates the orientation estimation errors of the ground robot. The adapter for Gemini's fine-tuning procedure is set to 1.

3) Real-World Task Completion: To valid the system's practicability, we set high-intelligent word assembling task for the aerial-ground system, requiring the system to assemble letter blocks into specified words, such as "LOVE" and "OK," under different constraints and configurations. For instance, the system was tasked with aligning "LOVE" from right to left, arranging "BE" without repositioning the "B" block, among others.

We assessed the overall quality of task completion using multiple metrics, including:

- **Task Decomposition:** Human evaluators were involved in assessing whether the task decomposition adhered to the user's requirements, ensuring that the system correctly interpreted and planned tasks according to the specified goals.
- **Global Map Construction Accuracy:** This metric focused on the system's ability to generate precise global maps, which are crucial for effective aerial path planning. The accuracy of these maps directly impacts the aerial robot's navigation and task execution capabilities.
- **Recognition & Labeling:** This aspect assessed the model's ability to accurately recognize and position objects within the environment and apply correct semantic labels to different objects in the scene according to current task.
- **Collision Times:** To evaluate this, we recorded the average number of collisions occurring during one pick-transportation-placement task. This metric provides insight into the planner's capability to utilize semantic map information for collision-free motion planning.
- **Task Completion Success:** Finally, this metric examined whether the final arrangement of the letter blocks matched the user's specifications and constraints, indicating the overall effectiveness and accuracy of task execution.

The results shown in Table II demonstrated The aerial-ground system demonstrates how multimodal foundation models enable precise manipulation and adaptive task execution in heterogeneous robot collaboration. These results suggest potential applications in automated object handling, logistics, and assistive robotics.

TABLE II
PERFORMANCE EVALUATION OF THE AERIAL-GROUND ROBOT SYSTEM

Metric	Accuracy/Success Rate	Times
Task Decomposing	1.0	5
Global Map Construction	0.8	5
Recognition & Labeling	0.74	171
Collision Times	—	0.4
Task Completion Success	0.8	5

IV. CONCLUSION

In this paper, we presented two independent approaches for foundation model-driven robotic task execution: LLM-based task decomposition with motion primitives and a VLM-assisted aerial-ground robotic system. By examining these approaches together, we demonstrated how LLMs enable structured task decomposition and reasoning, while VLMs facilitate zero-shot perception and semantic navigation.

Experimental results validate the effectiveness of each approach in its respective domain. The first study highlights how LLM-driven task decomposition and motion primitives enable one-shot learning for high-difficulty tasks, improving robot adaptability with minimal human intervention. The second study showcases the role of foundation models in multi-robot collaboration, where LLMs decompose complex tasks and reasoning steps, while VLMs accurately parse aerial images to support ground robot navigation.

These findings highlight the potential of foundation models in enhancing robotic autonomy across different levels of task execution. Future work will explore more dynamic and complex task scenarios, including real-time adaptation of LLM-generated motion primitives and enhanced robustness of VLM-based perception in unstructured environments. Additionally, we aim to extend 3D environment navigation capabilities in aerial-ground robotic systems, integrating foundation models for improved spatial awareness and decision-making in real-world applications.

V. Acknowledgment

This paper presents two independent studies on foundation model-driven robotic task execution. The first study, which focuses on LLM-driven task decomposition and one-shot learning through Human-Robot Collaboration, has been published in IEEE Robotics and Automation Letters (RA-L). The second study, exploring VLM-assisted multimodal perception and aerial-ground robot collaboration, is currently under preparation for submission.

References

[1] M. Ahn *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," 2022. [Online]. Available: https://arxiv.org/abs/2204.01691

[2] J. Liang *et al.*, "Code as policies: Language model programs for embodied control," 2023. [Online]. Available: https://arxiv.org/abs/2209.07753

[3] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[5] J. Wang *et al.*, "Large language models for robotics: Opportunities, challenges, and perspectives," *Journal of Automation and Intelligence*, 2024.

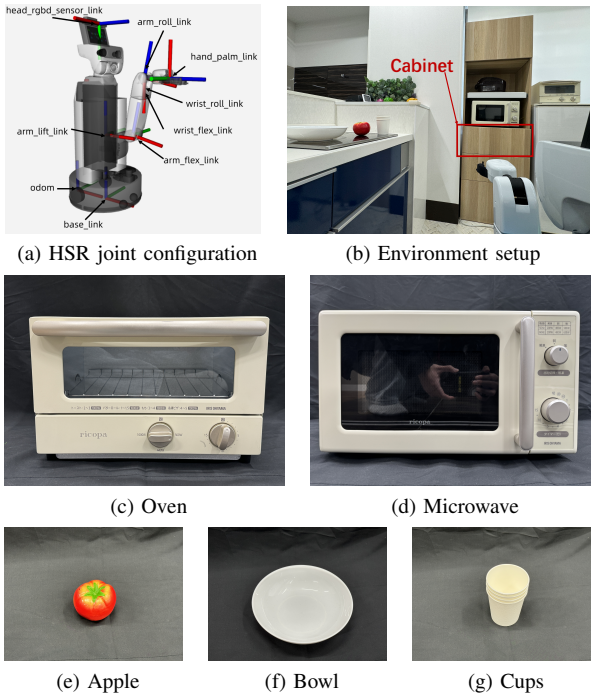
[6] H. Liu *et al.*, "Enhancing the llm-based robot manipulation through human-robot collaboration," *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 6904–6911, 2024.

[7] S. Schaal, "Dynamic movement primitives-a framework for motor control in humans and humanoid robotics," in *Adaptive motion of animals and machines*. Springer, 2006, pp. 261–280.

[8] Y. Li, M. Zhao, J. Sugihara, and T. Nishio, "Cooperative navigation system of agv and uav with autonomous and precise landing," in *2024 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, 2024, pp. 1477–1483.

Appendix

A. Experiment Devices



(a) HSR joint configuration (b) Environment setup

(c) Oven (d) Microwave

(e) Apple (f) Bowl (g) Cups

Fig. 7. HSR joint configuration, experiment environment setup, and objects utilized in the experiments

1) Task Decomposition and Motion Primitives: The system has undergone real-world testing on the Human Support Robot (HSR) from Toyota. Oculus's VR device is used for teleoperation. The experiment objects are shown in Fig. 7.

The experiments were set in a well-lit kitchen environment as shown in Fig. 7b, where the robot was tasked with performing a variety of domestic tasks initiated through natural language commands provided by users. The experiments aim to evaluate the performance of the system for routine zero-shot tasks and more intricate and specialized one-shot tasks.

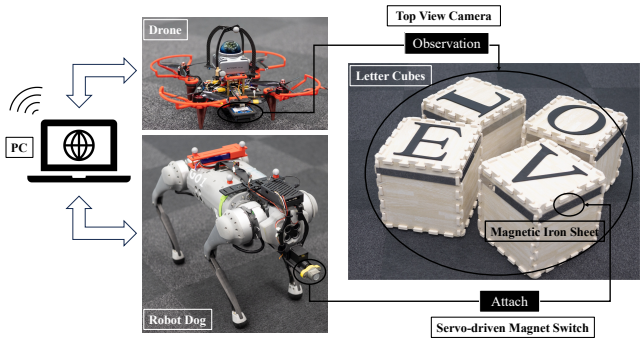


Fig. 8. The illustration of how the aerial-ground robots system interact with the objects which need pick-transport-place.

2) Aerial-Ground Robot System: The multimodal robots framework is implement on an aerial-ground robot system which including one Go1 - quadruped robot from Unitree

and one self-designed quadrotor aircraft which equipped with SLAM system [8].

Besides, serval letter blocks are provided for Pick-Transportation-Placement, the ground robot need to assemble them to words according to the command from user.

The appearance of the aerial-ground system and the letter blocks are depicted in Fig. 8.

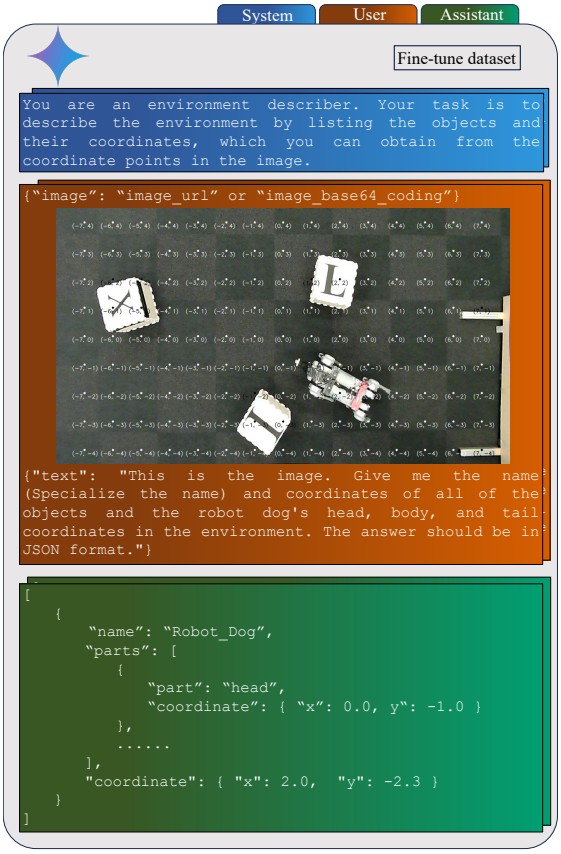


Fig. 9. A illustration of the fine-tuning dataset, including 'system' for prompting, 'user' for task requiring and GridMask-based images input, and 'assistant' for ideal answer designing.

B. Prompt Engineering

1) LLM - Task assignment: These prompts are designed to decompose the command into sub-tasks and assign the sub-tasks to according to the robot characteristics.

```
# Duty Clarify
You are a task decomposer for a heterogeneous
multirobot system. Decompose complex tasks into
sub-tasks for each robot according to their
abilities. Assume the robots already know the
positions of all objects. Each sub-task should
align with a specific ability of a robot.
```

```
# Robots List
- Drone Abilities:
1. Construct the map. (Must be the first step).
- Robot Dog Abilities:
1. Attaching objects.
2. Detaching objects.
- Drone and Robot Dog Cooperation Ability:
1. Move or carry something to somewhere.
```

2) LLM - Motion Functions Selection: These prompts are used to give the available motion functions of each robot to the LLM.

```
# Drone
You are responsible for choosing the motion
function to finish the task.
Motion Functions Library:
1. construct_map()
- Controls the drone to construct a global map.
# Robot Dog
You are responsible for choosing the motion
function to finish the task.
Motion Function Library:
1. go1_attach_start('name_of_target_object')
2. go1_detach_start('name_of_target_object')
- 'name_of_target_object'. Name of the object
need to be attached or detached. Example:
'A_letter_cube', 'green_cube'
```

C. Fine-tuning

This section illustrate appearance of GridMask and dataset structure in Fig. 9. Including 'system' for prompting, 'user' for task requiring and GridMask-equipped images input, and 'assistant' for ideal answer designing.