

# Compositional Coordination for Multi-Robot Teams with Large Language Models

Zhehui Huang<sup>1</sup>, Guangyao Shi<sup>1</sup>, Yuwei Wu<sup>2</sup>, Vijay Kumar<sup>2</sup>, Gaurav S. Sukhatme<sup>1</sup>

**Abstract**— Multi-robot coordination has traditionally relied on a mission-specific and expert-driven pipeline, where natural language mission descriptions are manually translated by domain experts into mathematical formulation, algorithm design, and executable code. This conventional process is labor-intensive, inaccessible to non-experts, and inflexible to changes in mission requirements. Here, we propose LAN2CB (Language to Collective Behavior), a novel framework that leverages large language models (LLMs) to streamline and generalize the multi-robot coordination pipeline. LAN2CB transforms natural language (NL) mission descriptions into executable Python code for multi-robot systems through two core modules: (1) Mission Analysis, which parses mission descriptions into behavior trees, and (2) Code Generation, which leverages the behavior tree and a structured knowledge base to generate robot control code. We further introduce a dataset of natural language mission descriptions to support development and benchmarking. Experiments in both simulation and real-world environments demonstrate that LAN2CB enables robust and flexible multi-robot coordination from natural language, significantly reducing manual engineering effort and supporting broad generalization across diverse mission types. Website: <https://sites.google.com/view/lan-cb>

## I. INTRODUCTION

**Problem:** Multi-robot coordination for complex collective behaviors, including coverage, formation, foraging, and exploration, has been extensively studied for decades [1]–[7]. However, most prior coordination methods target a specific and narrowly defined coordination task. This highlights the need for a general and unified framework that can address a wide range of coordination problems in a systematic way. The conventional multi-robot development pipeline starts with a natural language (NL) problem description, enabling both expert and non-expert users to specify objectives and constraints. An expert translates this description into a mathematical formulation, typically as an optimization problem, which inevitably requires additional assumptions. Following this, a domain expert designs algorithms to solve the formulated problem, which are subsequently implemented and validated on robotic hardware. However, this pipeline has three limitations: It is labor-intensive, inaccessible to non-experts, and inflexible. The first two limitations are apparent due to the significant use of skilled experts, despite the repetitive nature of problem formulation, algorithm design, and code implementation. The third becomes apparent when even

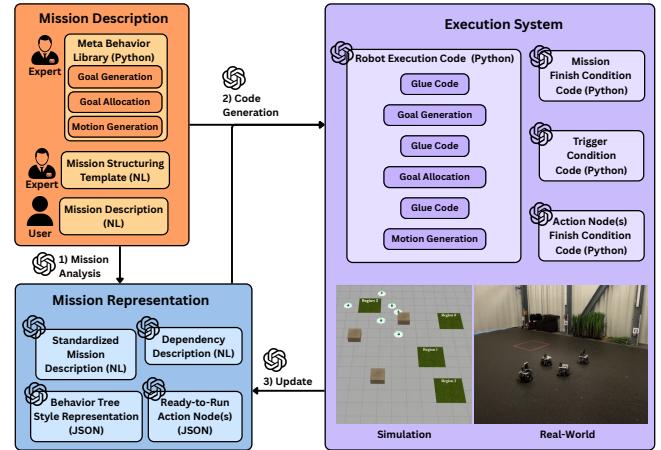


Fig. 1. LAN2CB is an LLM-assisted multi-robot framework that converts mission descriptions into structured representations (e.g., behavior trees), assigns roles and priorities to each robot, and automatically generates executable code to accomplish complex missions. A mission change does not require human intervention since new code is generated automatically. NL means natural language. █ user- or expert-provided content. █ LLM-generated mission representations. █ LLM-generated code.

minor changes to the problem statement or new requirements emerge, as the entire pipeline must be manually redesigned.

**Claim:** We posit that recent advances in large language models (LLMs) can be leveraged to alleviate these three limitations. Specifically, the use of such models can significantly reduce human expert effort and enhance the flexibility of multi-robot coordination systems [8]–[12].

**Background:** LLMs are trained on massive amounts of text data to understand and generate human-like language, allowing them to interpret natural language prompts, perform contextual reasoning, and produce coherent responses or actions in both visual and linguistic tasks [13]–[19]. In robotics, LLMs are increasingly being used to improve high-level decision-making, task planning, and human-robot interaction [20]–[28]. Applications include interpreting natural language instructions, generating structured plans, assisting in code generation for robot control, and enabling dialogue-based coordination among multi-robot systems. The key advantages of LLMs in robotics include their flexibility, ability to adapt to new tasks without extensive retraining, and capacity to bridge the gap between human intent and machine execution, thereby significantly reducing the need for manual programming and domain-specific engineering.

**Contributions:** 1. We propose, implement, and thoroughly evaluate an LLM-based framework, Language to Collective Behavior (**LAN2CB**), which addresses the limitations of the traditional pipeline and turns natural language mission descriptions into executable plans and Python code for robot

<sup>1</sup> Zhehui Huang, Guangyao Shi, and Gaurav S. Sukhatme are with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089, USA. Email: {zhehuihu, shig, gaurav}@usc.edu. <sup>2</sup> Yuwei Wu and Vijay Kumar are with the GRASP Lab, University of Pennsylvania, Philadelphia, PA 19104, USA. Email: {yuweiwu, kumar}@seas.upenn.edu.

teams. **LAN2CB** consists of two main modules (Fig. 1). The first module is MISSION ANALYSIS, which takes the mission specified in natural language as input, identifies the necessary tasks to be completed, and analyzes the dependency between tasks. The output of the first module, all ready-to-run nodes from a behavior tree, is given to the second module CODE GENERATION, together with a pre-constructed knowledge base, to generate executable code for robots. 2. We design a dataset of natural language mission specifications for multi-robot coordination. 3. Our experiments, both in simulation and hardware, demonstrate that **LAN2CB** produces effective multi-robot coordination from natural language. They also show that **LAN2CB** is flexible and extendable, obviating the need to design from scratch whenever the mission changes and providing a standard format to incorporate existing knowledge into its knowledge base.

## II. RELATED WORK

### A. Multi-Robot Coordination

Multi-robot systems can solve large tasks efficiently through cooperation, coordination, or collaboration [29]. Prior work has shown that these approaches enable robots to contribute effectively to overall team performance, either through global control strategies [30] or by relying on local observations and decentralized decision-making [31, 32]. As tasks become more complex, task-level coordination has received growing attention, where high-level goals are decomposed into interdependent sub-tasks that are allocated and executed collaboratively by the team [33]–[35]. For example, Messing et al. [36] introduced a four-layer structure covering task planning, allocation, scheduling, and motion planning to simplify decision-making. More recent work applied learning-based methods to multi-robot coordination by leveraging neural architectures such as graph attention networks for scheduling [37] and heterogeneous policy networks for communication and collaboration in diverse robot teams [38]. However, most existing approaches remain limited to specific tasks and lack the generalization to handle more dynamic objectives, understand natural language, or perform higher-level reasoning abilities.

### B. Integrating LLMs in Robotics

1) *Single-Robot Integration*: Recent work has explored integrating LLMs with single robots to enable more intuitive human-robot interaction and high-level task planning. These approaches leverage the language understanding and reasoning capabilities of LLMs to interpret natural language commands and translate them into executable actions. Because natural language is inherently tied to contexts and semantics, LLMs are particularly effective in understanding tasks such as navigation, exploration, and instruction following. In these scenarios, LLMs can help robots ground language inputs into spatial or semantic representations that guide behavior in complex, partially known environments [22]–[24]. In addition, real-time resilience [20] and anomaly detection [25] have been integrated into LLM-based robotic pipelines to improve robustness in dynamic and unforeseen situations.

2) *Team-Level Integration*: Extending LLMs to multi-robot systems introduces challenges in coordination, communication, and scalability, with research still in the early stages exploring their use for flexible, robust multi-robot coordination [39]. Some works use LLMs to translate natural language into formal representations [40]–[43], but they are limited to simple missions and cannot handle multi-robot teams or decomposable long-horizon tasks. Others propose prompt-based frameworks for multi-robot coordination using LLMs [27, 28]. For instance, Zhao et al. introduced RoCO [27], where robots coordinate via interactive dialogues. However, such frameworks face limitations with large robot teams (e.g.,  $\geq 10$ ) and complex language-based missions. Shyam et al. proposed SMART-LLM [28], an LLM-based task planner for multi-robot teams with task decomposition and allocation, but it lacks domain knowledge, limits generalization beyond in-context examples, and cannot handle motion-level language commands. Our work is closely related to GenSwarm [44], which translates natural language task descriptions into executable Python code. While both aim to enable language-driven multi-robot coordination, our framework differs in two key ways. First, our modular design supports a broader range of coordination problems, including multi-team tasks with interdependencies and trigger events, unlike GenSwarm’s focus on simple, single-team scenarios. Second, we incorporate a structured knowledge base to guide the LLM, enhancing extensibility, whereas GenSwarm relies solely on in-context learning, limiting generality.

## III. PROBLEM STATEMENT

Consider an environment with  $N$  robots. Given natural language mission descriptions from users, our goal is to develop a system that can automatically control robots to accomplish these missions without further human intervention. We focus on two fundamental categories of missions: 1) goal-approaching missions, such as geometric formation or region visitation, and 2) herding missions, where robots must guide dynamic objects into specific regions or achieve a specific coverage percentage of an area. Mission diversity is introduced by systematically varying key dimensions, resulting in a broad spectrum of possible tasks:

- 1) **Coordination Strategies**: Missions may require robots to coordinate, form a range of geometric patterns (e.g., lines, circles, characters), visit assigned regions, track dynamic targets, or herd dynamic objects.
- 2) **Motion Patterns**: Robots may be required to execute various motion trajectories, such as moving in straight lines, zigzagging, or spiral paths.
- 3) **Constraints**: Additional requirements can be imposed, such as entering a region from a specific direction or adhering to prescribed durations for actions.
- 4) **Trigger**: Tasks can include trigger conditions that allow actions to be terminated or modified in response to environmental events. For example, if a robot comes within a specified distance (e.g., 1 meter) of a forbidden region, it needs to switch to another task.

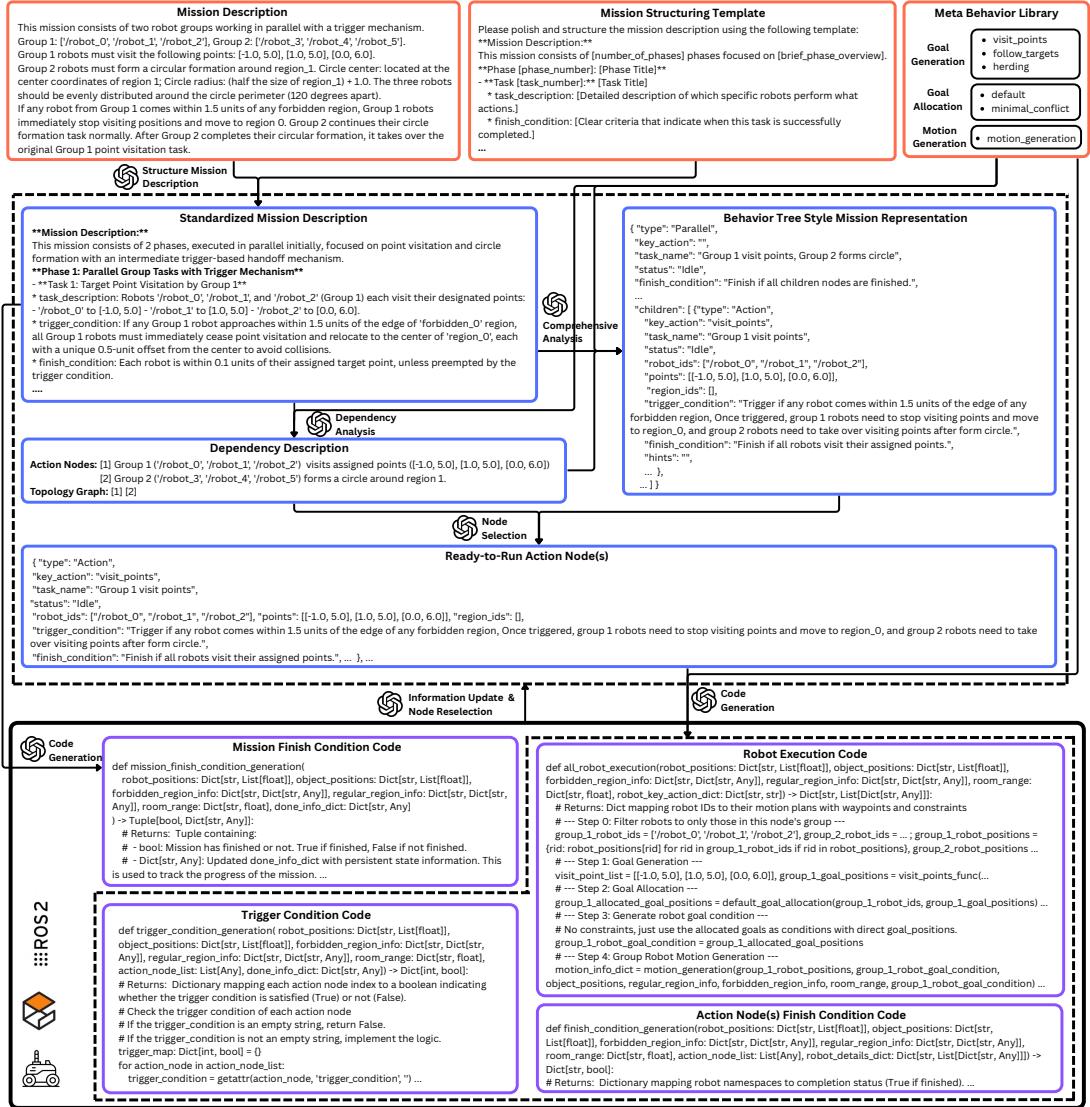


Fig. 2. A comprehensive demonstration of **LAN2CB**. █ human-provided content. █ LLM-generated mission descriptions. █ LLM-generated code.

- 5) **Task Finish Condition:** Each atomic task has a clearly defined completion criterion, such as reaching a designated position within a user-specified tolerance or herding objects within a given range of the target region.
- 6) **Mission Finish Condition:** Mission completion is defined more broadly, either by the successful completion of all constituent atomic tasks or by satisfying higher-level objectives (e.g., achieving at least 60% coverage of a specified region).

#### IV. METHOD

The **LAN2CB** workflow consists of the following stages, illustrated in Fig. 1:

- 1) **Mission Analysis:** Given natural language mission descriptions from users, the LLM autonomously decomposes missions into atomic task specifications and analyze inter-task dependencies. For each task, the LLM extracts robot assignments, relevant constraints, triggers, finish conditions, contextual hints, and other contextual

requirements. Tasks that are immediately executable are then identified and forwarded to the next stage.

- 2) **Code Generation:** Based on the mission description, the system first generates the mission finish condition, which is generated once for the entire mission cycle. For each selected task, considering its constraints, trigger conditions, finish conditions, and hints, the system separately generates code for: (a) robot execution, (b) trigger conditions, and (c) task finish conditions.
- 3) **Execution:** The generated code is deployed to the robots for execution. During run-time, if any trigger condition or task finish condition is met, the workflow transitions to stage 4. If the overall mission finish condition is satisfied, it proceeds directly to stage 5.
- 4) **Mission Progress Update:** Upon receiving execution results, the system uses the LLM to update the status of each task, then returns to stage 1 for further analysis and potential replanning if necessary.
- 5) **Mission Completion:** Once the mission finish condition is met, the mission is formally marked as complete.

TABLE I  
FIELD DESCRIPTIONS FOR BEHAVIORTREENODE.

Field (All Nodes)	Description
idx	Unique index of the node.
node.type	Parallel   Sequence   Action.
task.name	Concise task description of the node.
status	Idle   Running   Success   Failure.
constraints	Constraints such as max speed, time limit, regions to avoid, and etc.
trigger.condition	Condition(s) that trigger node termination and activate new nodes.
finish.condition	Condition(s) for determining task completion.
hints	Hints or suggestions.
children	List of child BehaviorTreeNode objects (empty for Action nodes).
Field (Action Nodes)	Description
action.type	visit.points   follow.targets   herd.
robot.ids	List of robot IDs this action node will control.
object.ids	List of object IDs related to the action node.
region.ids	List of region IDs related to the action node.
points	List of points related to the action node.

### A. Mission Analysis

Mission analysis consists of two key phases: mission understanding and ready-to-run action node selection. However, user-provided natural language mission descriptions often omit explicit definitions of atomic tasks and their execution dependencies. As a result, directly prompting LLMs for general mission understanding can yield incomplete or ambiguous task specifications and may overlook critical execution dependencies. To address these challenges, the mission understanding phase begins by standardizing the mission description using an expert-provided template. We then perform dependency analysis to explicitly extract task relationships, which serve as the basis for a comprehensive mission analysis. Afterward, the LLM selects ready-to-run action nodes based on the results of the dependency and comprehensive mission analyses.

**Dependency Analysis:** To initiate dependency analysis, we prompt the LLM to extract all atomic tasks from the mission description, initially disregarding trigger conditions and any tasks that may be activated post-trigger. The LLM then analyzes the dependency relationships among these atomic tasks. If triggers are activated during execution, we re-invoke the LLM to update the dependency analysis accordingly.

**Comprehensive Mission Analysis:** Building on the results of dependency analysis, we combine these atomic tasks and their dependencies with the standardized mission description to prompt the LLM for comprehensive mission analysis. The mission is represented in a behavior tree structure, with the overall mission as the root and atomic tasks as action nodes. Non-leaf nodes of type Parallel or Sequence capture task execution topologies according to dependencies. For each node, we instruct the LLM to generate a unique index, identify the node type, task name, status, constraints, trigger conditions, finish conditions, hints, and children. For action nodes, the LLM also specifies the action type as well as the associated robots, objects, regions, and points. See Tab. I for detailed field descriptions of the BehaviorTreeNode.

**Ready-to-Run Action Node(s) Selection:** After constructing the complete behavior tree and integrating the dependency analysis results, we ask the LLM to identify action nodes that are immediately executable. Specifically, these are

nodes for which all preconditions are met. The ready-to-run action nodes are then selected and forwarded for downstream processing, code generation. This approach ensures that robot actions are initiated only when all prerequisite tasks and constraints have been resolved, thereby supporting safe and efficient mission execution. After selecting action nodes that are ready-to-run, we instruct the LLM to update the status of all nodes in the behavior tree, marking the selected nodes and their parent nodes as *Running*.

### B. Code Generation

After mission analysis and the identification of ready-to-run action nodes, we prompt the LLM to generate Python code to control all robots in the environment. The code generation process consists of four key components, executed synchronously with the robots' control loop (Fig. 1 and 2): 1) Robot execution code generation; 2) Trigger condition code generation; 3) Action node finish condition code generation; 4) Mission finish condition code generation.

**Robot Execution Code Generation:** Recent works have shown that LLMs can effectively assist code generation [45]–[48]. However, to improve success rates and mitigate hallucinations, it is more effective to provide core functions in advance and expose them as APIs, rather than asking the LLM to generate all code from scratch [49]. To this end, we construct a meta behavior library for robot execution. To ensure flexibility and the capability to process multiple action nodes simultaneously, the meta-behavior library comprises three sub-libraries:

- *Goal Generation:* Provides primitives for three fundamental behaviors: (a) visiting points, (b) following targets, and (c) herding.
- *Goal Allocation:* Offers two strategies for mapping generated goals to robots: (a) a default mapping that assigns goal positions based on robot IDs, and (b) a minimal conflict strategy that assigns goals to minimize trajectory intersections, assuming straight-line movement from start to goal.
- *Motion Generation:* Generates feasible paths from start to goal, considering environmental constraints. For each waypoint, the function not only outputs the position but also constraints such as maximum speed for approach.

This modular design allows the LLM to compose robust execution code for all selected action nodes, promoting reusability, flexibility, and scalability in multi-robot systems.

**Trigger Condition Code Generation:** For each action node, LLMs are prompted to implement the corresponding trigger condition. If no trigger is specified, the function always returns `False`. Otherwise, the LLM encodes the trigger logic using available environmental information, including robot and object states and region properties.

**Action Node Finish Condition Code Generation:** For each selected action node, the LLM generates code to evaluate whether the node's specified finish condition has been met. For example, if the task is to visit three points, the action node is considered complete when all three points have been visited.

**Mission Finish Condition Code Generation:** Mission finish condition generation is similar in nature to composite task finish condition generation. However, the LLM only needs to generate the mission finish condition code once for the entire mission.

### C. Execution

Once all code scripts are generated, the system initiates or resumes execution. During execution, if a trigger condition is satisfied, we prompt the LLM to update the dependency analysis by removing the current task that caused the trigger and adding any new tasks associated with the trigger. The updated dependency analysis, together with the mission description, is then used to prompt the LLM to generate a new behavior tree. The LLM subsequently selects new ready-to-run action nodes and generates code for them. If the finish condition of any action node is satisfied, the LLM updates the behavior tree. This may also include updating parent nodes if their completion depends solely on the completion of their children. The LLM then selects new ready-to-run action nodes and generates the corresponding code. If the finish condition of any composite node is satisfied, the LLM updates the behavior tree by marking all of its child nodes that have not started or are still running as `Failure`, and marking the composite node as `Success`. Execution proceeds iteratively until the mission finish condition is satisfied.

## V. EXPERIMENTS

The experimental study is designed to rigorously evaluate **LAN2CB** under a variety of multi-step, long-horizon missions. We focus on three questions: 1) How effectively does **LAN2CB** perform across diverse missions? 2) How does the use of a standardized template for mission descriptions impact the performance of **LAN2CB** compared to raw natural language input? 3) Can **LAN2CB** be reliably deployed on real-world robotic platforms?

### A. Performance Across Diverse Missions

To the best of our knowledge, no standard dataset exists for evaluating multi-step, long-horizon multi-robot mission execution under natural language mission descriptions. Therefore, we curate a comprehensive mission suite comprising three groups (nine missions total), each designed to test the distinct capabilities of our system.

**Mission Dataset:** We organize the missions by the complexity of mission analysis and required code generation:

- **Category A: Basic Missions.** These missions comprise multiple atomic tasks executed in a straightforward order, with no trigger conditions. Mission analysis and code generation are direct: all necessary information can be extracted from the natural language descriptions, enabling the LLM to generate input parameters for behavior library functions without additional reasoning. *Examples: Assigning robots to visit predefined points.* (See Missions #1, #2, and #3 in Fig. 3.)
- **Category B: Basic Missions w/ Advanced Code Generation.** While maintaining a simple overall structure, these

TABLE II  
QUANTITATIVE RESULTS.

Mission Category	W/ Template	Avg. Success Rate	Avg. Tokens ( $10^3$ )		Avg. Cost (\$)	Avg. Tokens / Err ( $10^3$ )
			Input	Output		
Category A	✗	0.87	48.70	12.33	0.20	92.48
	✓	1.00	52.83	14.53	0.22	-
Category B	✗	0.73	40.53	9.97	0.16	37.39
	✓	0.93	35.25	8.54	0.14	128.10
Category C	✗	0.73	44.24	10.06	0.17	37.73
	✓	0.87	58.12	14.86	0.24	111.45
<b>Overall</b>		0.78	44.49	10.79	0.18	55.87
		0.93	48.73	12.64	0.20	119.78

\*Avg. Tokens / Error is based on output tokens.

missions require advanced code generation. Certain behaviors are not directly supported by the behavior libraries, necessitating that the LLM synthesizes additional logic or routines. *Example: Generating a spiral trajectory for robots when only straight-line motion is natively supported.* (See Missions #4, #5, and #6 in Fig. 3.)

- **Category C: Advanced Missions.** These missions demand advanced analysis and complex code generation. They may involve intricate dependencies, trigger conditions, or non-trivial mission finish conditions, such as region coverage requirements, dynamic triggers, or adaptive behaviors. (See Missions #7, #8, and #9 in Fig. 3.)

**Evaluation Metrics:** We quantitatively evaluate missions using the metrics: (1) *Avg. Success Rate*: The proportion of missions completed successfully as intended, without human intervention. (2) *Avg. Token Number and Cost*: The total number of LLM tokens used per mission and related cost. (3) *Avg. Tokens per Error*: The average number of tokens consumed per error given the output tokens of the LLM.

**Results:** We evaluate all missions using GPT-4.1, conducting each mission five times. Detailed results are provided in Tab. II. By standardizing raw mission descriptions from users with an expert-provided template, **LAN2CB** achieves a 100% success rate on Category A (basic missions), a 93% success rate on Category B (basic missions with advanced code generation) and 87% success rate on Category C (advanced missions). In addition, the system demonstrates strong robustness, achieving  $\sim 0.12$  million tokens per error across all missions. These results demonstrate that **LAN2CB** reliably analyzes mission descriptions, including identifying task dependencies, extracting essential information, selecting ready-to-run nodes, and dynamically updating behavior trees. Notably, the high success rates in Category B and Category C highlight its ability to perform advanced code generation tasks. These tasks include synthesizing robot trajectories not natively supported by the behavior library, generating custom formations, creating complex execution conditions such as specifying approach directions or target-following duration, and generating both triggers and mission finish conditions. Overall, our results indicate that **LAN2CB** is both robust and adaptable for diverse multi-robot mission execution scenarios.

### B. Impact of standardized template

To quantify the impact of using a template to standardize the mission descriptions on **LAN2CB**, we ablated the

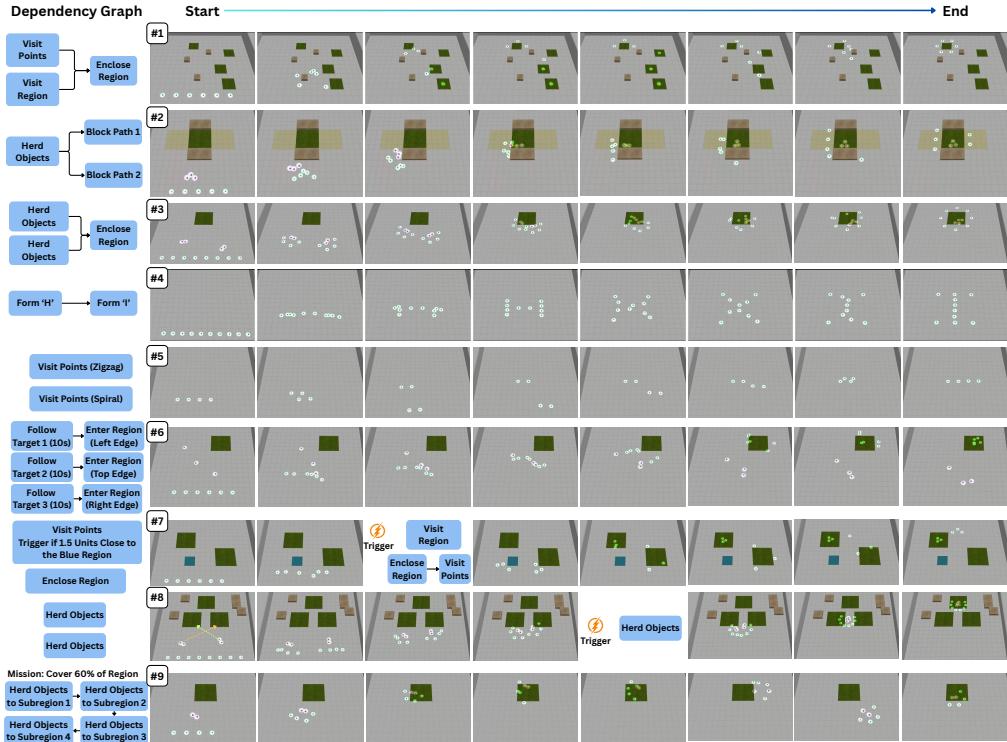


Fig. 3. Quantitative results for nine multi-robot missions. Please check the website for more details.

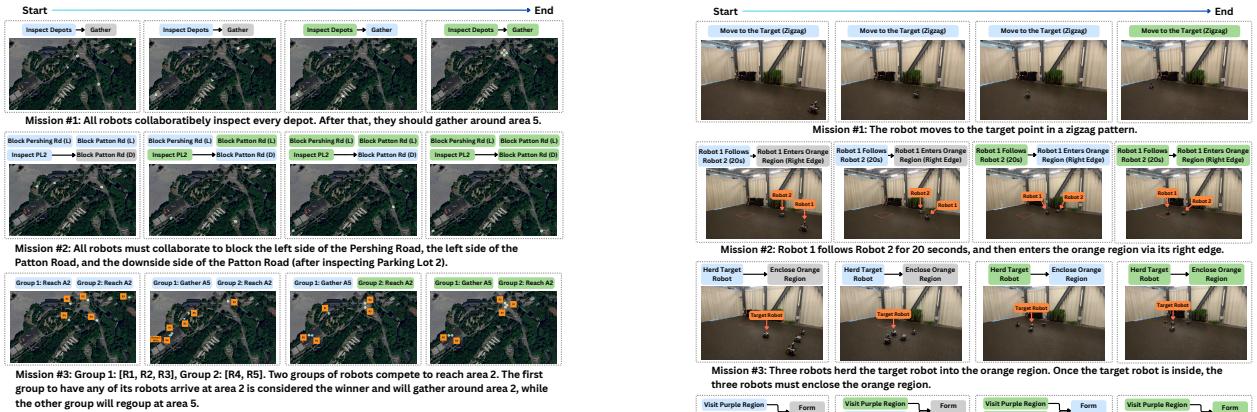


Fig. 4. Realistic-setting experiments.   idle,   running,   finished. template-reformatting stage that aligns user-provided natural language mission descriptions to our predefined mission schema. As reported in Tab. II, this omission results in an average drop of  $\sim 15\%$  in success rates and causes the error rate to become  $2.14\times$  more frequent across all mission categories. These results highlight the critical role of structured prompting in maintaining robust multi-robot performance.

### C. Realistic Simulated Scenarios

We validated our framework in three simulation scenarios designed to approximate real-world conditions (see Fig. 4).

### D. Real World Demonstrations

We validated our framework through four real-world demonstrations<sup>1</sup> across two robot platforms (see Fig. 5).

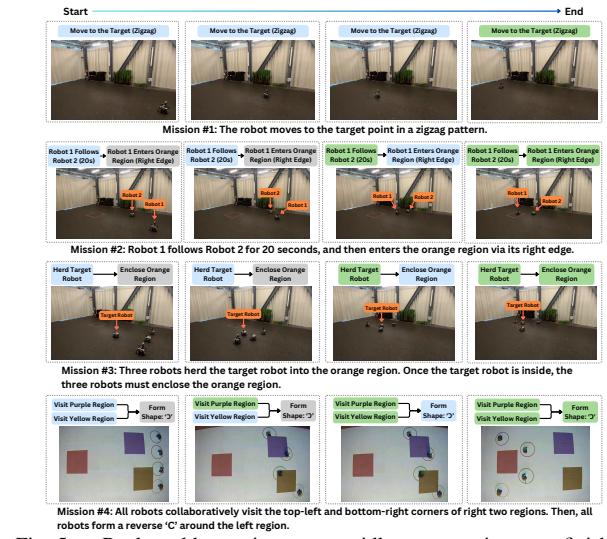


Fig. 5. Real-world experiments.   idle,   running,   finished.

## VI. CONCLUSIONS

We presented an LLM-based framework, Language to Collective Behavior (**LAN2CB**), a pipeline that converts natural language into executable Python code for robot teams. Our experiments suggest that **LAN2CB** enables effective multi-robot coordination and remains flexible and extensible, even for multi-step and long-horizon missions. Additionally, we designed a dataset of natural language mission descriptions for multi-robot coordination. For future work, we aim to improve mission analysis through retrieval-augmented generation (RAG) [50], enabling LLMs to leverage past mission data more effectively. Furthermore, we plan to expand the meta-behavior library to support a broader and more diverse range of capabilities.

<sup>1</sup>We thank Jiazen Liu for assistance with the physical deployment.

## REFERENCES

- [1] W. Burgard, M. Moors, D. Fox, R. Simmons, and S. Thrun, "Collaborative multi-robot exploration," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No.00CH37065)*, vol. 1, 2000, pp. 476–481 vol.1.
- [2] M. Turpin, N. Michael, and V. Kumar, "Capt: Concurrent assignment and planning of trajectories for multiple robots," *International Journal of Robotics Research*, vol. 33, no. 1, pp. 98 – 112, January 2014.
- [3] L. E. Parker, D. Rus, and G. S. Sukhatme, *Multiple Mobile Robot Systems*. Cham: Springer International Publishing, 2016, pp. 1335–1384.
- [4] H. Garcia de Marina, M. Cao, and B. Jayawardhana, "Controlling rigid formations of mobile agents under inconsistent measurements," *IEEE Transactions on Robotics*, vol. 31, no. 1, pp. 31–39, 2015.
- [5] J. Alonso-Mora, S. Baker, and D. Rus, "Multi-robot navigation in formation via sequential convex programming," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 4634–4641.
- [6] F. Amigoni, J. Banfi, and N. Basilico, "Multirobot exploration of communication-restricted environments: A survey," *IEEE Intelligent Systems*, vol. 32, no. 6, pp. 48–57, 2017.
- [7] M. Santos, Y. Diaz-Mercado, and M. Egerstedt, "Coverage control for multirobot teams with heterogeneous sensing capabilities," *IEEE Robotics and Automation Letters*, vol. 3, no. 2, pp. 919–925, 2018.
- [8] Z. Li, A. Mao, D. Stephens, P. Goel, E. Walpole, A. Dima, J. Fung, and J. Boyd-Graber, "Improving the tenor of labeling: Re-evaluating topic models for content analysis," 2024. [Online]. Available: <https://arxiv.org/abs/2401.16348>
- [9] Z. Li, L. Calvo-Bartolomé, A. Hoyle, P. Xu, A. Dima, J. F. Fung, and J. Boyd-Graber, "Large language models struggle to describe the haystack without human help: Human-in-the-loop evaluation of topic models," 2025. [Online]. Available: <https://arxiv.org/abs/2502.14748>
- [10] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.
- [11] Y. Wu, Y. Tao, P. Li, G. Shi, G. S. Sukhatmem, V. Kumar, and L. Zhou, "Hierarchical llms in-the-loop optimization for real-time multi-robot target tracking under unknown hazards," *arXiv preprint arXiv:2409.12274*, 2024.
- [12] J. Jiang, F. Wang, J. Shen, S. Kim, and S. Kim, "A survey on large language models for code generation," *arXiv preprint arXiv:2406.00515*, 2024.
- [13] F. Zeng, W. Gan, Y. Wang, N. Liu, and P. S. Yu, "Large language models for robotics: A survey," *arXiv preprint arXiv:2311.07226*, 2023.
- [14] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [15] J. Gao, B. Sarkar, F. Xia, T. Xiao, J. Wu, B. Ichter, A. Majumdar, and D. Sadigh, "Physically grounded vision-language models for robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 12 462–12 469.
- [16] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, "A survey of state of the art large vision language models: Benchmark evaluations and challenges," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, June 2025, pp. 1587–1606.
- [17] Z. Li, X. Wu, G. Shi, Y. Qin, H. Du, T. Zhou, D. Manocha, and J. L. Boyd-Graber, "Videohallu: Evaluating and mitigating multi-modal hallucinations on synthetic video understanding," 2025. [Online]. Available: <https://arxiv.org/abs/2505.01481>
- [18] Z. Li, Y. Chang, Y. Zhou, X. Wu, Z. Liang, Y. Y. Sung, and J. L. Boyd-Graber, "Semantically-aware rewards for open-ended rl training in free-form generation," 2025. [Online]. Available: <https://arxiv.org/abs/2506.15068>
- [19] Z. Li, I. Mondal, Y. Liang, H. Nghiem, and J. L. Boyd-Graber, "Pedants: Cheap but effective and interpretable answer equivalence," 2024. [Online]. Available: <https://arxiv.org/abs/2402.11161>
- [20] A. Tagliabue, K. Kondo, T. Zhao, M. Peterson, C. T. Tewari, and J. P. How, "Real: Resilience and adaptation using large language models on autonomous aerial robots," in *2024 IEEE 63rd Conference on Decision and Control (CDC)*, 2024, pp. 1539–1546.
- [21] K. Obata, T. Aoki, T. Horii, T. Taniguchi, and T. Nagai, "Lip-llm: Integrating linear programming and dependency graph with large language models for multi-robot task planning," *IEEE Robotics and Automation Letters*, 2024.
- [22] D. Shah, B. Osinski, B. Ichter, and S. Levine, "LM-nav: Robotic navigation with large pre-trained models of language, vision, and action," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: <https://openreview.net/forum?id=UW5A3SweAH>
- [23] R. Schumann, W. Zhu, W. Feng, T.-J. Fu, S. Riezler, and W. Y. Wang, "Velma: verbalization embodiment of llm agents for vision and language navigation in street view," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. [Online]. Available: <https://doi.org/10.1609/aaai.v38i17.29858>
- [24] Z. Ravichandran, V. Murali, M. Tzes, G. J. Pappas, and V. Kumar, "Spine: Online semantic planning for missions with incomplete natural language specifications in unstructured environments," *International Conference on Robotics and Automation (ICRA)*, 2025. [Online]. Available: <https://arxiv.org/abs/2410.03035>
- [25] R. Sinha, A. Elhafsi, C. Agia, M. Foutter, E. Schmerling, and M. Pavone, "Real-time anomaly detection and reactive planning with large language models," in *Robotics: Science and Systems (RSS)*, 2024.
- [26] A. A. Khan, M. Andrev, M. A. Murtaza, S. Aguilera, R. Zhang, J. Ding, S. Hutchinson, and A. Anwar, "Safety aware task planning via large language models in robotics," 2025. [Online]. Available: <https://arxiv.org/abs/2503.15707>
- [27] Z. Mandi, S. Jain, and S. Song, "Roco: Dialectic multi-robot collaboration with large language models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 286–299.
- [28] S. S. Kannan, V. L. Venkatesh, and B.-C. Min, "Smart-llm: Smart multi-agent robot task planning using large language models," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 12 140–12 147.
- [29] A. Prorok, M. Malencia, L. Carbone, G. S. Sukhatme, B. M. Sadler, and V. Kumar, "Beyond robustness: A taxonomy of approaches towards resilient multi-robot systems," *arXiv preprint arXiv:2109.12343*, 2022.
- [30] J. Desai, J. Ostrowski, and V. Kumar, "Controlling formations of multiple mobile robots," in *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No.98CH36146)*, vol. 4, 1998, pp. 2864–2869 vol.4.
- [31] L. Iocchi, D. Nardi, M. Piaggio, and A. Sgorbissa, "Distributed coordination in heterogeneous multi-robot systems," *Autonomous robots*, vol. 15, pp. 155–168, 2003.
- [32] N. Michael, M. M. Zavlanos, V. Kumar, and G. J. Pappas, "Distributed multi-robot task assignment and formation control," in *2008 IEEE International Conference on Robotics and Automation*, 2008, pp. 128–133.
- [33] Z. Chen, J. Alonso-Mora, X. Bai, D. D. Harabor, and P. J. Stuckey, "Integrated task assignment and path planning for capacitated multi-agent pickup and delivery," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5816–5823, 2021.
- [34] K. Leahy, Z. Serlin, C.-I. Vasile, A. Schoer, A. M. Jones, R. Tron, and C. Belta, "Scalable and robust algorithms for task-based coordination from high-level specifications (scratches)," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2516–2535, 2022.
- [35] J. Motes, R. Sandström, H. Lee, S. Thomas, and N. M. Amato, "Multi-robot task and motion planning with subtask dependencies," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3338–3345, 2020.
- [36] A. Messing, G. Neville, S. Chernova, S. Hutchinson, and H. Ravichandar, "Grstaps: Graphically recursive simultaneous task allocation, planning, and scheduling," *The International Journal of Robotics Research*, vol. 41, no. 2, pp. 232–256, 2022. [Online]. Available: <https://doi.org/10.1177/02783649211052066>
- [37] Z. Wang and M. Gombolay, "Learning scheduling policies for multi-robot coordination with graph attention networks," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4509–4516, 2020.
- [38] E. Seraj, R. Paleja, L. Pimentel, K. M. Lee, Z. Wang, D. Martin, M. Sklar, J. Zhang, Z. Kakish, and M. Gombolay, "Heterogeneous policy networks for composite robot team communication and coor-

- dination,” *IEEE Transactions on Robotics*, vol. 40, pp. 3833–3849, 2024.
- [39] P. Li, Z. An, S. Abrar, and L. Zhou, “Large language models for multi-robot systems: A survey,” *arXiv preprint arXiv:2502.03814*, 2025.
- [40] Y. Chen, J. Arkin, C. Dawson, Y. Zhang, N. Roy, and C. Fan, “Autotamp: Autoregressive task and motion planning with llms as translators and checkers,” in *2024 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2024, pp. 6695–6702.
- [41] Z. Wei, X. Luo, and C. Liu, “Hierarchical temporal logic task and motion planning for multi-robot systems,” *arXiv preprint arXiv:2504.18899*, 2025.
- [42] X. Zhang, H. Qin, F. Wang, Y. Dong, and J. Li, “Lammap: Generalizable multi-agent long-horizon task allocation and planning with lm-driven pdl planner,” 2025. [Online]. Available: <https://arxiv.org/abs/2409.20560>
- [43] J. X. Liu, Z. Yang, I. Idrees, S. Liang, B. Schornstein, S. Tellex, and A. Shah, “Lang2ltl: Translating natural language commands to temporal robot task specification,” in *Conference on Robot Learning (CoRL)*, 2023. [Online]. Available: <https://arxiv.org/abs/2302.11649>
- [44] W. Ji, H. Chen, M. Chen, G. Zhu, L. Xu, R. Groß, R. Zhou, M. Cao, and S. Zhao, “Genswarm: Scalable multi-robot code-policy generation and deployment via language models,” *arXiv preprint arXiv:2503.23875*, 2025.
- [45] I. Singh, V. Blukis, A. Mousavian, A. Goyal, D. Xu, J. Tremblay, D. Fox, J. Thomason, and A. Garg, “Progprompt: Generating situated robot task plans using large language models,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 523–11 530.
- [46] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, “Do as i can, not as i say: Grounding language in robotic affordances,” *arXiv preprint arXiv:2204.01691*, 2022.
- [47] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, “Code as policies: Language model programs for embodied control,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.
- [48] Y. Meng, F. Chen, Y. Chen, and C. Fan, “Audere: Automated strategy decision and realization in robot planning and control via llms,” *arXiv preprint arXiv:2504.03015*, 2025.
- [49] Z. Huang, G. Shi, and G. S. Sukhatme, “Can large language models solve robot routing?” in *International Symposium of Robotics Research*, 2024.
- [50] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.