IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Co-NavGPT: Multi-Robot Cooperative Visual Semantic Navigation Using Vision Language Models

Bangguo Yu, Qihao Yuan, Kailai Li, Hamidreza Kasaei, and Ming Cao

***Abstract*—Visual target navigation is a critical capability for autonomous robots operating in unknown environments, particularly in human-robot interaction scenarios. While classical and learning-based methods have shown promise, most existing approaches lack common-sense reasoning and are typically designed for single-robot settings, leading to reduced efficiency and robustness in complex environments. To address these limitations, we introduce Co-NavGPT, a novel framework that integrates a Vision Language Model (VLM) as a global planner to enable common-sense multi-robot visual target navigation. Co-NavGPT aggregates sub-maps from multiple robots with diverse viewpoints into a unified global map, encoding robot states and frontier regions. The VLM uses this information to assign frontiers across the robots, facilitating coordinated and efficient exploration. Experiments on the Habitat-Matterport 3D (HM3D) demonstrate that Co-NavGPT outperforms existing baselines in terms of success rate and navigation efficiency, without requiring task-specific training. Ablation studies further confirm the importance of semantic priors from the VLM. We also validate the framework in real-world scenarios using quadrupedal robots. Supplementary video and code are available at: https://sites.google.com/view/co-navgpt2.**

***Index Terms*—Vision-Based Navigation, Multi-Robot Systems, AI-Enabled Robotics.**

I. Introduction

HUMANS efficiently explore complex environments by leveraging common-sense knowledge of environmental structures and collaborative strategies. Similarly, autonomous robots require reasoning abilities to effectively navigate and explore their environments. For example, consider a person instructing two robots in his house, "Hey, robots, please bring my cell phone to me." In this task, implicitly, the first critical step is for the two robots to locate the specific object, namely the cell phone, collaboratively. Although substantial progress has been made in object-goal navigation and multi-robot exploration, efficiently enabling robots to collaboratively identify and locate target objects from visual inputs remains challenging due to semantic complexity and environmental intricacies. In this paper, we address the task of multi-robot visual target navigation, wherein multiple robots collaboratively explore unknown environments to efficiently locate a specific target object. This capability serves as the enabling functionality for various practical robotic applications.

Recent advances in simulation platforms [1], [2], large-scale 3D scene datasets [2], [3], map-based representations [4], and large foundation models [5], [6] have spurred significant interest in visual target navigation. Existing approaches can be broadly categorized into end-to-end frameworks [7], [8] and modular pipelines [9], [10], which aim to improve navigation performance by incorporating spatial and semantic features through supervised or reinforcement learning. More recently, Vision Language Models (VLMs) have been introduced into navigation pipelines [11], [12], leveraging their strengths in semantic reasoning and object grounding. VLMs provide high-level semantic priors, enabling better understanding of object relationships and scene structure. However, the vast majority of existing methods are designed for single-robot settings. In large or complex environments, single-agent navigation often suffers from poor efficiency due to the need to explore vast unknown areas. Additionally, such systems are vulnerable to failure, as a single incorrect decision or unexpected obstacle can significantly delay or derail the navigation process. These limitations motivate the need for multi-robot collaboration in visual target navigation.

Learning-based approaches typically require large amounts of data and corresponding computation to generalize across diverse environments, a challenge that becomes more profound in multi-robot scenarios [13]. With the emergence of VLMs, new opportunities have arisen for zero-shot and few-shot planning in embodied tasks. VLMs possess inherent world knowledge and have shown promise as high-level planners

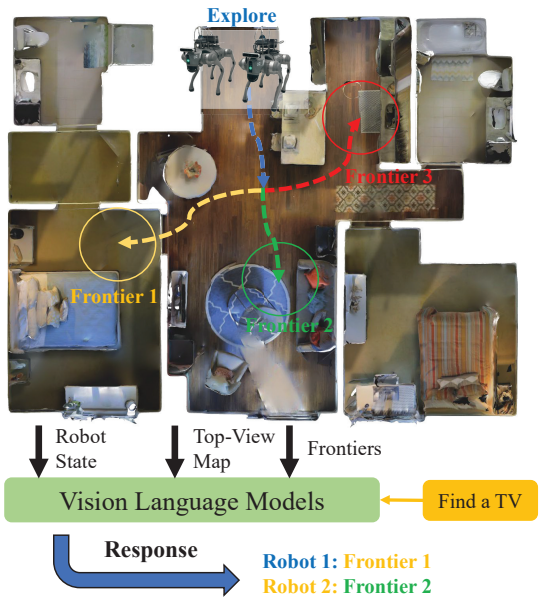


Fig. 1: Two robots visual target navigation example. When multiple unexplored frontiers are detected, the vision language model assigns the frontier for each robot based on the current observation and the target object.

All authors are with the Faculty of Science and Engineering, University of Groningen, the Netherlands. {b.yu, qihao.yuan, kailai.li, hamidreza.kasaei, m.cao}@rug.nl

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

in instruction-following tasks [14], as well as in long-horizon, multi-agent coordination [15]–[18]. These capabilities make VLMs an ideal candidate for enabling common-sense reasoning in various multi-robot tasks.

This paper addresses the challenge of multi-robot visual semantic navigation, wherein multiple robots collaboratively locate a target object in unknown environments. Specifically, we introduce Co-NavGPT, a novel framework that leverages vision-language models to generate efficient exploration and search policies for multi-robot cooperation. Within this framework, a VLM functions as a global planner, strategically assigning unexplored frontier regions to each robot. An illustrative example of visual target navigation, locating a television, is presented in Fig. 1. After mapping their environment, robots must select their next exploration frontier. Utilizing contextual information from the observed environment map, each robot's state, and the given navigation target, a VLM allocates the most relevant frontiers to individual robots, enhancing collaborative search efficiency. We evaluate Co-NavGPT on the Habitat simulation platform [19], comparing its navigation performance against existing multi-robot methods within extensive photorealistic 3D environments from HM3D [3]. Ablation experiments further validate the effectiveness of integrating VLMs in multi-robot navigation tasks. Additionally, we demonstrate our method's practical applicability using two quadrupedal robots in real-world experiments. In contrast to other multi-robot navigation methods, Co-NavGPT uniquely utilizes VLMs to encode frontier-enhanced environment representations, thereby significantly improving scene understanding and cooperative navigation efficiency. Our results highlight the substantial potential of VLMs for managing complex multi-robot collaborative tasks.

Our contributions are summarized as follows:

- We propose Co-NavGPT, a novel framework that merges multi-robot observations into a global semantic map. The framework uses vision-language models to guide robots toward efficient collaborative exploration and navigation in unknown environments.
- We design a VLM-based global planner that allocates frontier goals based on spatial context and semantic cues, enabling scalable multi-robot coordination without requiring task-specific training.
- Experiments on HM3D demonstrate that our proposed multi-robot cooperative framework significantly improves visual target navigation performance. Furthermore, we validate Co-NavGPT in real-world scenarios using two quadruped robotic platforms and achieve real-time performance in effective multi-robot navigation.

II. RELATED WORK

A. Visual Semantic Navigation

Visual semantic navigation is a fundamental capability for intelligent agents, inspired by human-like semantic reasoning and object search. Early classical approaches constructed metric or topological maps of the environment and planned paths to target objects accordingly. More recently, end-to-end learning methods have gained traction. [7] introduced a reinforcement learning-based policy that encodes RGB observations and target images into a joint embedding space. Subsequent works have enhanced navigation performance through various technologies [5], [8]. However, monolithic learning policies often suffer from poor sample efficiency and generalization to unseen environments. To address these limitations, [9] proposed a modular framework that separates semantic mapping, global planning, and local control, requiring learning only at the high-level planning stage. This design improves both learning efficiency and transferability. Subsequently, diverse feature representations [20], [21] have been employed to train high-level policies. Zero-shot [4], [12], [22] and few-shot [10], [23] multi-modal frameworks further enhance scene generalization. However, single-robot methods still face limited exploration efficiency and low fault tolerance; a detection error or unforeseen obstacle in the environment can disrupt the entire navigation mission or significantly increase task duration, especially in large or complex environments. In this work, we extend visual semantic navigation to multi-robot cooperative scenarios in unknown spaces, aiming to leverage VLMs for faster and more robust multi-robot target search.

B. Multi-Robot Cooperative Navigation

Many studies have addressed the limitations of single-robot systems by examining multi-robot cooperation across various domains, including active mapping [24], [25], exploration [26], and target search [13], [18]. Classical planning-based methods primarily focus on coordinating multiple robots for goal assignment [27], as exemplified by challenges like the multiple traveling salesman problem. In exploration tasks, both multi-agent reinforcement learning [26] and graph-based methods [24] have been proposed to extend single-agent planners to multi-agent settings. [13] emphasized multi-agent visual semantic navigation, leveraging scene prior knowledge to locate objects within maps and subsequently formulating the navigation policy via reinforcement learning. While both planning-based and learning-based techniques have achieved success in multi-robot tasks, they often require real-world common-sense learning for robot assignment. In contrast, vision-language models encode rich, generalized prior knowledge, making them well-suited for multi-robot navigation. [17] employed large language models (LLMs) as planners to facilitate collaboration among agents and humans in dialogue-driven scenarios. Building on this line of work, [16] enabled the robots to discuss and reason about task strategies using LLMs. Recent work by [28] showed that structured annotations can improve VLMs interaction, motivating the adoption of mark-based visual prompting in both manipulation [29] and navigation [30] tasks. Closely related to the approach presented in this paper is the recent work of [18], which also aims to explore unknown environments with multiple robots. Their method employs multimodal chain-of-thought score collaboration to plan cooperative semantic navigation, achieving strong performance on the HM3D and MP3D datasets. However, its real-world deployment remains challenging due to the high computational cost and low execution frequency that is below 0.5 frames per second (FPS) in simulation [18]. In contrast,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

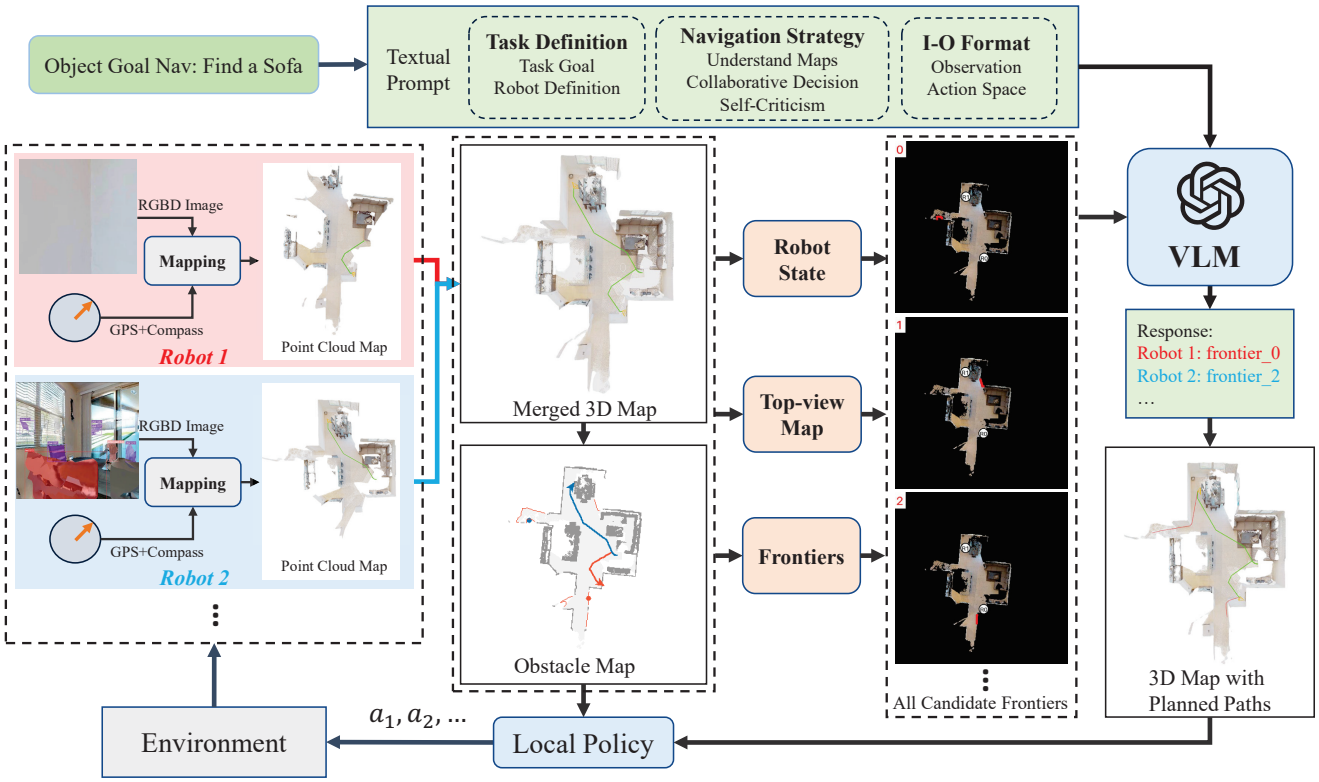


Fig. 2: System architecture of the proposed multi-robot navigation framework. Each robot processes RGB-D observations to generate a local point cloud map, which is then merged into a global 3D map. The merged map, robot states, and candidate frontiers are encoded into a prompt and passed to a vision-language model, which acts as a global planner to assign frontier goals to each robot. A local policy then computes paths to the assigned frontiers based on the obstacle map, enabling coordinated exploration and target search.

our framework achieves an average of 5 FPS onboard mobile robots in the real world while maintaining competitive navigation performance. Building on these insights, our framework integrates a VLM as a global planner using frontier-based visual prompting to enable real-time navigation strategies. This enables contextual reasoning and robot-to-frontier assignment, resulting in more efficient exploration and search.

III. The Proposed Method

A. Task Definition

In the multi-robot visual target navigation task, all robots cooperate to locate an object of a specified category within unknown scenes. The set of categories is described by $C = \{c_1, \ldots, c_m\}$ and the set of scenes can be represented by $S = \{s_1, \ldots, s_k\}$. For each episode, n robots $R = \{r^1, \ldots, r^n\}$ are initialized at the same position p_i within scene s_i, but with different initial orientations. All robots are assigned the same target category c_i. Thus, each episode is defined by $T_i = \{N, s_i, c_i, p_i\}$. At each time step t, each robot r^i receives an observation o_t^i from its own perspective and performs an action a_t^i simultaneously. The observation o^i contains RGB-D images I, the location and orientation p of the robot, and the object category c. The action space $a \in \mathcal{A}$ consists of six discrete actions: `move_forward`, `turn_left`, `turn_right`, `look_up`, `look_down`, and `stop`. Executing `move_forward` advances the robot by 25 cm, while the rotation actions change the robot's orientation by 30° in the corresponding direction. The `stop` action is triggered when the robot is close to the target object. An episode is considered successful if the robot issues the `stop` action while within 0.1 m of the target. Each robot is allowed a maximum of 500 time steps per episode.

B. Overview

As illustrated in Fig. 2, our framework leverages VLMs for goal selection in a multi-robot navigation setting. Each robot collects RGB-D observations to generate a local point cloud map. These individual maps are then merged into a global 3D representation based on the robots' positions. The merged map, along with each robot's state, a top-view map, and the set of detected frontiers, is encoded into a structured prompt and provided to the VLM. The VLM functions as a global planner, assigning distinct frontier goals to each robot based on contextual understanding. Given these long-horizon assignments, a local policy plans collision-free paths over the obstacle map, enabling each robot to explore and search for the target object efficiently.

C. Map Representation

1) 3D Point Cloud Map: For each robot r^i, given a sequence of RGB-D images $\mathcal{I}^i = \{I_1^i, \ldots, I_t^i\}$ and correspond-

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

ing poses $\mathcal{P}^i = \{p_1^i, \ldots, p_t^i\}$, a 3D point cloud map $\mathcal{M}^i$ is constructed incrementally. At each frame I_t^i, an open-vocabulary object detector $Det(\cdot)$ is first applied to identify candidate bounding boxes. A class-agnostic segmentation model $Seg(\cdot)$ then extracts segmentation masks for the detected objects. Depth information from each masked region, along with unsegmented areas, is projected into the global 3D frame using $F_{proj}(\cdot)$ based on the current pose p_t^i. The local map $\mathcal{M}^i$ is updated with each incoming observation. To construct the global point cloud map $\mathcal{M}$, all local maps $\mathcal{M}^i$ are merged into a common global coordinate frame according to each robot's estimated pose:

$$\mathcal{M} = \sum_{i=1}^{n} \sum_{\tau=1}^{t} F_{proj}\left(Seg(Det(I_\tau^i)), p_\tau^i\right). \tag{1}$$

Each semantic mask is accordingly projected onto the global semantic point cloud map. To further refine the map, we apply DBSCAN clustering [31] to filter out noisy points.

2) 2D Exploration Map: To support efficient exploration, we construct a 2D exploration map used for frontier extraction and path planning. The 2D map is initialized with zeros at the beginning of each episode, with the robot positioned at the center. All 3D point cloud data $\mathcal{M}$ is projected onto a top-down 2D grid map, which consists of two channels: an obstacle map and an explored map. Points above the floor are selected and projected onto the obstacle map, while all 3D points contribute to the explored map. Frontiers are extracted from the boundaries of these two channels. First, the edge of the explored area is identified by detecting the largest contours in the explored map. Then, the frontier map is computed by dilating the obstacle edge and subtracting it from the explored area. Connected neighborhood analysis is applied to group frontier cells into clusters, and small clusters are discarded as noise. The resulting frontiers provide spatial candidates for the VLM-based planner described in the following section.

D. VLM-based Multi-Robot Exploration

After constructing the 3D point cloud and 2D exploration maps, the VLM is used to enable efficient multi-robot exploration by assigning frontiers to each robot. This is achieved via a specifically designed prompting scheme that encodes spatial and semantic context.

1) Multi-Modal Prompting: To handle the complexity of environments, we design both textual and visual prompts to enhance the VLM's understanding of the explored context.

Visual Prompt: The visual prompt is derived from the global 3D point cloud, rendered into a colored top-view map that shares the same coordinate system and resolution as the 2D exploration map. Each robot's position is annotated using a circle and labeled with a unique ID. Since multiple frontiers may exist at each step, we generate one visual map per candidate, each masked with a frontier and labeled with its ID in the top-left corner to assist the VLM in identification. These maps provide both semantic and geometric information, enabling the VLM to reason over spatial configurations. Moreover, the top-view map provides a unified representation that simultaneously captures all robots and candidate frontiers. This global perspective enables the VLM to coordinate decisions across agents for collaborative exploration and target search—capabilities that are difficult to achieve when relying solely on individual camera views. An example of the visual prompt is shown in Fig. 2.

Textual Prompt: The textual prompt encodes the task objective, environment assumptions, and expected output format. It is designed to clarify the task structure and facilitate effective frontier assignment. The prompt includes: (i) a task description, (ii) contextual background, (iii) reasoning requirements, and (iv) a standardized input-output format. An example is shown below:

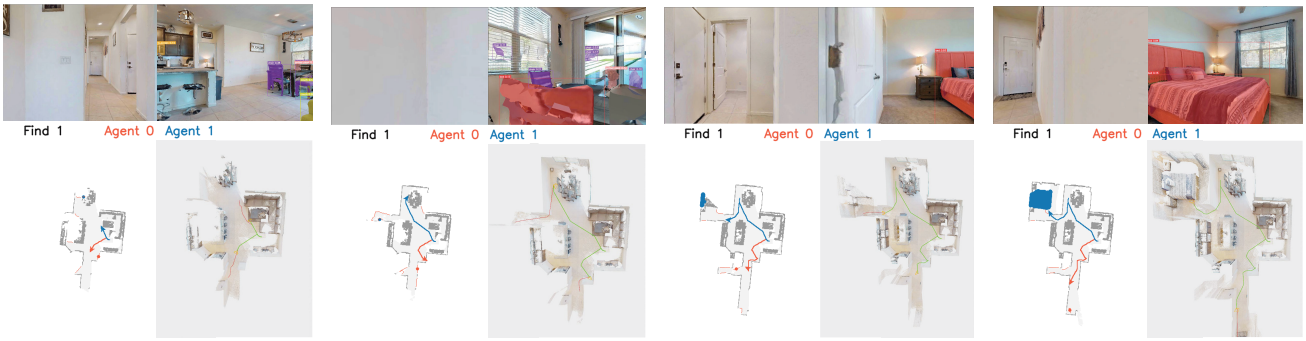
> **Task**: Locate the given target
>
> **Context**:
> We have multiple robots. Each robot perceives the environment and can navigate to explore unknown areas.
> The global top-view map shows the positions of each robot, candidate frontiers, and their IDs.
>
> **Requirements**:
> Understand: the scene layout, the robots' state, and the frontiers.
> Analyze: collaborative exploration and efficient target searching.
> Decide: a frontier assignment policy such that each robot moves to an optimal frontier.
> Justify: Reconsider the decision with a concise explanation.
>
> **Input**: Multiple top-view maps, each containing one candidate frontier. A target object category is provided.
> **Output**: A JSON object indicating the frontier IDs assigned to each robot and the reason.

Given the above, the textual and visual prompts inject structured commonsense into the VLM, improving its ability to generate accurate and explainable frontier assignments. Once the prompt is complete, it is passed to the VLM to obtain the final assignment for each robot.

2) Decision Making: As noted in prior work [14], [15], while vision-language models are effective at generating high-level plans, they are less reliable for fine-grained low-level control. To address this, we use the VLM solely for global goal assignment.

First, at each global step, an updated point cloud map is used to construct a visual prompt that captures both environmental and robot-specific information. The VLM then assigns a frontier waypoint to each robot, as illustrated in Fig. 2. The frontier assignment is guided by two criteria: (i) collaborative exploration of unknown regions, and (ii) semantic relevance to the target object. This encourages the selection of frontiers that not only expand map coverage but are also more likely to be near the target based on the top-view semantic layout. Second, to improve collaborative efficiency and interpretability, the VLM is also required to provide reasoning for its assignment based on the observations and criteria, thereby leveraging its inherent reasoning capabilities. The output is a JSON object specifying the frontier IDs assigned to each robot together with the corresponding rationale. The VLM may hallucinate or violate the required format, such as producing incorrect agent IDs, invalid frontier IDs, or non-JSON outputs. In such cases, the system prompts the VLM to regenerate its decision until a correctly formatted and valid output is obtained. When the number of available frontiers is smaller than the number

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) step 25 (b) step 50 (c) step 75 (d) step 85

Fig. 3: Visualization of the visual target navigation process in the Habitat simulator using two robots to locate a bed. The top row shows first-person RGB observations from both robots (Agent 0 in red, Agent 1 in blue) at different time steps. The bottom row presents the corresponding 2D exploration map and 3D point cloud map. The red and blue lines represent the respective paths of Agent 0 and Agent 1. Red and blue dots denote frontier goals assigned by the vision-language model.

of robots, the VLM is allowed to assign the same frontier to multiple robots, so that robots may temporarily share frontiers in such cases. If no viable frontiers are detected from the exploration map, random points within the explored space are selected as fallback goals. The global planner updates these long-term goals every 25 local steps.

3) Local Policy: Once a long-term goal is assigned, each robot plans a path using the Fast Marching Method (FMM) [32] from its current position to the goal. A short-range local goal is then selected along this path. The final action $a_t \in \mathcal{A}$ is computed to reach this local goal, taking into account obstacles and robot dynamics. At each step, the robot updates its local map and recalculates the local goal based on new sensor observations. This local policy compensates for the limitations of VLMs in fine-grained decision-making, ensuring smooth and efficient navigation.

IV. EXPERIMENTS

In this section, we compare our method with other multi-robot map-based baselines in the simulation to evaluate the performance of our framework. Additionally, we apply our process in two real-world robot platforms to validate its practicality for multi-robot navigational tasks.

A. Simulation Experiment

1) Dataset: We conduct experiments using the HM3D_v0.2 dataset [3], which features high-resolution, photorealistic 3D reconstructions of real-world environments. Following the standard protocol, we use 36 validation scenes comprising 1,000 episodes with semantic object annotations. Although our framework supports open-set detection, we select 6 goal categories with the setting in [9] to enable fair comparison: *chair, sofa, plant, bed, toilet*, and *TV*.

2) Settings: All experiments are conducted using the Habitat simulator [19]. Each robot receives 480×640 RGB-D observations, odometry information, and a goal category encoded as an integer. We use YOLO-World [33] for open-vocabulary detection $Det(\cdot)$ and Mobile-SAM [34] for class-agnostic segmentation $Seg(\cdot)$. The 2D exploration map covers a 24×24 m area with a resolution of 0.05 m. For global planning, we employ GPT-5 [6] via the OpenAI API as the VLM. A visual prompt is generated at each global planning step to represent all robots and the environment state, and the VLM outputs a JSON to assign frontier goals. Our implementation is based on publicly available codes from [35]. Although the framework supports an arbitrary number of robots, for clarity of performance demonstration, we use one, two, and three robots separately for each episode, both initialized at the same location but with different orientations. All the maps and robot states are visualized using Open3D [36].

3) Evaluation Metrics: We follow [9] to evaluate our method using the Success Rate (SR), Success weighted by the Path Length (SPL), and Distance to Goal (DTG) for this task. SR is defined as $\frac{1}{N}\sum_{i=1}^{N} S_i$, and SPL is defined as $\frac{1}{N}\sum_{i=1}^{N} S_i \frac{l_i}{\max(l_i,p_i)}$, where N is the number of episodes, $S_i = 1$ is considered successful if any robot successfully locates the target; otherwise, the episode is deemed a failure. l_i denotes the shortest trajectory length from the start position to one of the success positions, p_i stands for the shortest robot's trajectory length in the current episode i. Lastly, DTG represents the minimal distance between the robots and the target goal when the episode ends.

4) Baselines: To evaluate the navigation performance of our method, we compare it against several baselines. All baselines share the same framework for map construction using object detection; however, they differ in the global policy used for assigning robots to frontiers. After frontier selection, all robots follow the same local planning policy to execute actions.

- **Greedy [37]:** Frontiers are assigned to robots in a greedy manner. Each robot selects the nearest unassigned frontier as its goal.
- **Cost-Utility [38]:** Each frontier cell $f \in F$ is scored using a cost-utility function:

$$S^{CU}(f) = U(f) - \lambda_{CU} C(f)\,, \tag{2}$$

where $U(f)$ denotes the utility based on frontier size, and $C(f)$ represents the distance from the robot to

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I: Results of Comparative Study.

Method	One-Agent			Two-Agent			Three-Agent		
	SR↑	SPL↑	DTG↓	SR↑	SPL↑	DTG↓	SR↑	SPL↑	DTG↓
Greedy [37]	0.581	0.233	2.226	0.611	0.328	2.239	0.633	0.359	1.971
Cost-Utility [38]	0.592	0.228	2.120	0.625	0.323	2.030	0.656	0.355	1.799
Multi-SemExp [9]	0.568	**0.239**	2.440	0.612	0.327	2.234	0.631	0.364	2.080
Co-NavGPT (Ours)	**0.629**	0.236	**2.010**	**0.681**	**0.369**	**1.596**	**0.690**	**0.406**	**1.409**

the frontier. The parameter λ_{CU} balances the trade-off between utility and cost. Each robot selects the frontier with the highest $S^{CU}(f)$ score.

- **Multi-SemExp [9]:** We extend the SemExp framework to a multi-robot setting, in which each robot uses the navigation policy from SemExp to explore the unknown environment and search for the target.

5) Results and Discussion: The quantitative results are summarized in Table I. Among the baselines, Cost-Utility consistently outperforms Greedy, confirming the benefit of jointly considering frontier size and travel cost. The Multi-SemExp baseline also achieves a better SPL across one-agent and three-agent settings, which reflects the effectiveness of its learned global policy. Our method, Co-NavGPT, achieves the best overall performance, with clear improvements in both SR and SPL. In the one-agent case, Co-NavGPT increases SR by 6.3% compared to the strongest baseline (0.629 vs. 0.592) while maintaining competitive SPL. With two agents, the gains become more pronounced: SR increases by 9.0% (0.681 vs. 0.625) and SPL by 12.5% (0.369 vs. 0.328). In the three-agent case, Co-NavGPT achieves an SR improvement of 5.2% (0.690 vs. 0.656) and an SPL improvement of 11.5% (0.406 vs. 0.364). These results demonstrate that VLM-based reasoning enables more effective robot-to-frontier assignment, leading to robustness and scalability in collaborative exploration. An example trajectory of a successful episode in which the target is a bed is illustrated in Fig. 3. The sequence demonstrates how robots collaboratively explore different regions to efficiently locate the target.

6) Ablation Study: To evaluate the contribution of individual components in our framework, we conduct ablation studies on the HM3D dataset using the following variants with two robots:

- **VLM Models:** GPT-5 [6] is replaced with the lighter GPT-4o-mini [39] and GPT-4o [40] to assess the impact of VLM scale.
- **DownSam.:** The original resolution of the top-view map is 480×480. The down-sampled versions with 60×60, 120×120, and 240×240 resolution are used in VLM-based global planning to assess the effect of map resolution, while keeping the visual prompting format identical.
- **Obs.:** The top-view semantic map is replaced with an obstacle-only map, while preserving the same visual prompting format.
- **w/o Reason.:** The reasoning component is removed from the VLM output format.

The results in Table II show that our full model achieves the best performance in both success rate and SPL. Substituting GPT-5 with GPT-4o or GPT-4o-mini (**VLM Models**) leads to performance degradation across both metrics, suggesting that the stronger reasoning capability of larger VLMs plays a critical role in task effectiveness. Using either a down-sampled top-view map (**DownSam.**) or an obstacle-only map (**Obs.**) results in moderate drops in success rate and SPL, indicating the importance of semantic context in VLM-based decision making. Notably, the performance reduction with **DownSam.** is relatively minor, implying that the VLM is less sensitive to map resolution, while visual prompts containing frontier and robot positions provide more decisive cues when map details are limited. Finally, removing the explicit rationale-style prompting from the VLM output (**w.o. Reason**) yields the lowest performance, suggesting that requiring the model to produce a step-by-step rationale is beneficial for the VLM-based planner in our setting.

TABLE II: Results of Ablation Study in HM3D.

Ablation	Visual Prompt	VLM	Success↑	SPL↑
VLM Models	$\mathcal{M}_{top}$	GPT-4o-mini	0.641	0.307
	$\mathcal{M}_{top}$	GPT-4o	0.666	0.368
	$\mathcal{M}_{top}$	GPT-5	0.681	0.369
DownSam.	$\mathcal{M}_{top}/2$	GPT-5	0.675	0.360
	$\mathcal{M}_{top}/4$	GPT-5	0.676	0.351
	$\mathcal{M}_{top}/8$	GPT-5	0.675	0.346
w/o Reason.	$\mathcal{M}_{top}$	GPT-5	0.660	0.337
Obs.	$\mathcal{M}_{obs}$	GPT-5	0.679	0.358

7) Failure Cases: Although our framework achieves high success rates across a wide range of scenes, we observe several failure cases that highlight its current limitations and opportunities for future improvement. In the experiments in the HM3D dataset, approximately 14% of failures are attributed to exploration issues, while around 18% stem from object detection errors. Exploration-related failures are primarily caused by situations where robots become trapped in incomplete or poorly reconstructed regions of the scene (e.g., topological holes), or where the target object is either absent, undetectable, or located on a different floor. Multi-floor navigation remains a challenge under the current setup. Detection-related failures mainly result from misclassifications or missed detections, which serve as the primary detection backbone. This issue can be especially pronounced for small-scale objects, which remain difficult to detect reliably with the detector used in this work. In some cases, robots stop prematurely at positions

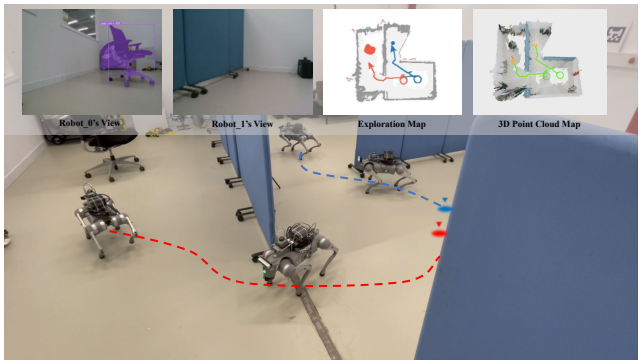
IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

distant from the actual target due to incorrect detections. Addressing these limitations, particularly in robust multi-level planning and reliable open-vocabulary object detection, remains an important direction for future work.

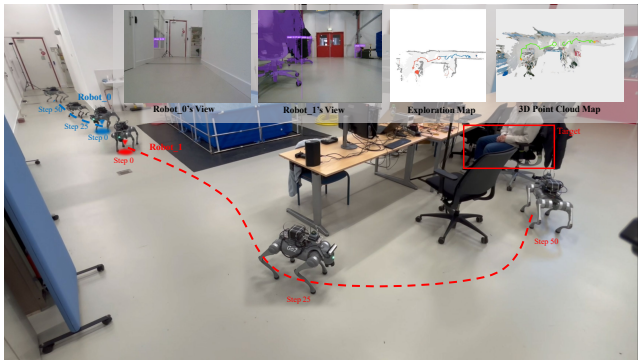
B. Real-World Experiment

We implement our framework for two Unitree Go2 quadruped robots, each equipped with a RealSense D455 RGB-D camera and a Livox Mid-360 LiDAR. The multi-agent system is deployed in a real-world, previously unseen lab environment. The navigation tasks include locating objects, namely, a chair, a sink, and a person. To facilitate the real-world deployment, we configure the sensors and modules according to the settings as Sec. IV-A2 in the simulation. The RGB-D camera is adjusted to match the resolution and depth range as in the simulation. Due to sensor noise caused by lighting and hardware variations, we apply DBSCAN [31] to cluster dense point regions and filter outliers in the point cloud. We exploit the onboard Livox Mid-360 LiDAR with the built-in IMU and employ FAST-LIO2 for localization [41], and conduct camera-to-LiDAR extrinsic calibration [42] to align depth data into the global frame. Before exploration begins, we apply Generalized-ICP [43] to compute initial pose alignment between the two robots, setting Robot 0's initial frame as the global reference. Once the map is constructed from RGB-D observations, frontiers are extracted and assigned by the vision-language model. Based on the map and robot positions, our framework generates waypoints and corresponding discrete navigation actions. Fig. 4 demonstrates three cases, where the proposed Co-NavGPT enabled cooperative navigation of different targets using the two quadruped robots. In each scenario, the VLM successfully assigns complementary frontiers to the robots, enabling cooperative exploration and efficient target search. The entire system runs in real time at approximately 5 FPS, which is sufficient for multi-robot navigation tasks in indoor environments.

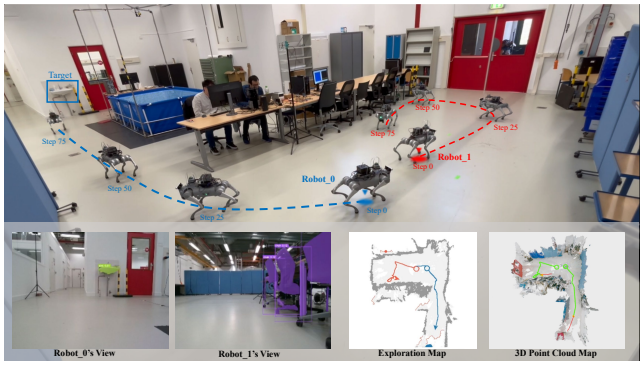
All perception and planning modules, including YOLO-World [33], Mobile-SAM [34], and our VLM-based planner, are deployed on a workstation equipped with an NVIDIA RTX 4090 GPU. Sensor processing, LiDAR-based localization, and low-level control modules run onboard the robots. Our framework is hardware-agnostic and can be flexibly deployed, as it only requires RGB-D observations and pose estimates as input, and outputs discrete navigation actions.



(a) Target: A sink.



(b) Target: A person.



(c) Target: A chair.

Fig. 4: Real-world multi-robot visual target navigation. Each scene shows the first-person RGB images, the exploration map, and the 3D point cloud map. In (a), the robots search for a sink; (b) search for a person; (c) search for a chair. Red and blue lines denote the trajectories of the two robots.

V. Conclusions

We presented Co-NavGPT, a novel framework that leverages vision-language models for multi-robot cooperative visual target navigation. By encoding scene-level information into structured visual prompts, a VLM functions as a global planner, enabling efficient frontier assignment for collaborative exploration and object search. Extensive experiments in simulation demonstrate that our approach significantly outperforms existing multi-robot baselines in terms of success rate and path efficiency, without relying on any task-specific learning. Real-world experiments also validate its practicality for multi-robot navigational tasks. These results highlight the strong potential of VLMs in coordinating complex multi-agent behaviors. Future work includes investigating tighter integration between VLMs and embodied agents in 3D environments, particularly toward interactive decision-making, dynamic replanning, and closed-loop real-time control.

References

[1] A. Szot, A. Clegg, E. Undersander, E. Wijmans, Y. Zhao, J. Turner, N. Maestre, M. Mukadam, D. Chaplot, O. Maksymets, A. Gokaslan, V. Vondrus, S. Dharur, F. Meier, W. Galuba, A. Chang, Z. Kira, V. Koltun, J. Malik, M. Savva, and D. Batra, "Habitat 2.0: Training Home Assistants to Rearrange their Habitat," *Advances in Neural Information Processing Systems*, vol. 1, pp. 251–266, 2021.

[2] F. Xia, A. R. Zamir, Z. He, A. Sax, J. Malik, and S. Savarese, "Gibson Env: Real-World Perception for Embodied Agents," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9068–9079, IEEE, jun 2018.

[3] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-Matterport 3D Dataset (HM3D): 1000 Large-scale 3D Environments for Embodied AI," *ArXiv*, sep 2021.

[4] X. Lei, M. Wang, W. Zhou, and H. Li, "Gaussnav: Gaussian splatting for visual navigation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 5, pp. 4108–4121, 2025.

[5] A. Khandelwal, L. Weihs, R. Mottaghi, and A. Kembhavi, "Simple but Effective: CLIP Embeddings for Embodied AI," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14809–14818, IEEE, jun 2022.

[6] OpenAI, "Gpt-5." https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2025-08-21.

[7] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3357–3364, IEEE, may 2017.

[8] H. Wang, A. H. Tan, and G. Nejat, "Navformer: A transformer architecture for robot target-driven navigation in unknown and dynamic environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 6808–6815, 2024.

[9] D. S. Chaplot, D. Gandhi, A. Gupta, and R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 2020-Decem, pp. 1–12, 2020.

[10] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "PONI: Potential Functions for ObjectGoal Navigation with Interaction-free Learning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 18868–18878, 2022.

[11] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, "Instructnav: Zero-shot system for generic instruction navigation in unexplored environment," in *Proceedings of The 8th Conference on Robot Learning* (P. Agrawal, O. Kroemer, and W. Burgard, eds.), vol. 270 of *Proceedings of Machine Learning Research*, pp. 2049–2060, PMLR, 06–09 Nov 2025.

[12] G. Zhou, Y. Hong, and Q. Wu, "NavGPT: Explicit Reasoning in Vision-and-Language Navigation with Large Language Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, pp. 7641–7649, 2024.

[13] X. Liu, D. Guo, H. Liu, and F. Sun, "Multi-Agent Embodied Visual Semantic Navigation with Scene Prior Knowledge," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3154–3161, 2022.

[14] Z. Yang, C. Garrett, D. Fox, T. Lozano-Pérez, and L. P. Kaelbling, "Guiding long-horizon task and motion planning with vision language models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 16847–16853, 2025.

[15] T. Yang, P. Feng, Q. Guo, J. Zhang, X. Zhang, J. Ning, X. Wang, and Z. Mao, "Autohma-llm: Efficient task coordination and execution in heterogeneous multi-agent systems using hybrid large language models," *IEEE Transactions on Cognitive Communications and Networking*, vol. 11, no. 2, pp. 987–998, 2025.

[16] Z. Mandi, S. Jain, and S. Song, "RoCo: Dialectic Multi-Robot Collaboration with Large Language Models," in *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 286–299, 2024.

[17] H. Zhang, W. Du, J. Shan, Q. Zhou, Y. Du, J. B. Tenenbaum, T. Shu, and C. Gan, "Building Cooperative Embodied Agents Modularly With Large Language Models," *12th International Conference on Learning Representations, ICLR 2024*, pp. 1–22, 2024.

[18] Z. Shen, H. Luo, K. Chen, F. Lv, and T. Li, "Enhancing Multi-Robot Semantic Navigation Through Multimodal Chain-of-Thought Score Collaboration," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 14, pp. 14664–14672, 2025.

[19] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik, D. Parikh, and D. Batra, "Habitat: A Platform for Embodied AI Research," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, vol. 2019-Octob, pp. 9338–9346, IEEE, oct 2019.

[20] D. S. Chaplot, D. Gandhi, S. Gupta, A. Gupta, and R. Salakhutdinov, "Learning To Explore Using Active Neural Slam," in *8th International Conference on Learning Representations, ICLR 2020*, apr 2020.

[21] S. Zhang, X. Yu, X. Song, X. Wang, and S. Jiang, "Imagine before go: Self-supervised generative map for object goal navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16414–16425, 2024.

[22] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, "Vlfm: Vision-language frontier maps for zero-shot semantic navigation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48, IEEE, 2024.

[23] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging Large Language Models for Visual Target Navigation," *IEEE International Conference on Intelligent Robots and Systems*, pp. 3554–3560, 2023.

[24] K. Ye, S. Dong, Q. Fan, H. Wang, L. Yi, F. Xia, J. Wang, and B. Chen, "Multi-Robot Active Mapping via Neural Bipartite Graph Matching," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2022-June, pp. 14819–14828, 2022.

[25] A. Asgharivaskasi, F. Girke, and N. Atanasov, "Riemannian optimization for active mapping with robot teams," *IEEE Transactions on Robotics*, vol. 41, pp. 1077–1097, 2025.

[26] C. Yu, X. Yang, J. Gao, H. Yang, Y. Wang, and Y. Wu, "Learning efficient multi-agent cooperative visual exploration," in *European Conference on Computer Vision*, pp. 497–515, Springer, 2022.

[27] D. Puig, M. A. Garcia, and L. Wu, "A new global optimization strategy for coordinated multi-robot exploration: Development and comparative evaluation," *Robotics and Autonomous Systems*, vol. 59, no. 9, pp. 635–653, 2011.

[28] A. Shtedritski, C. Rupprecht, and A. Vedaldi, "What does clip know about a red circle? visual prompt engineering for vlms," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11987–11997, 2023.

[29] K. Fang, F. Liu, P. Abbeel, and S. Levine, "MOKA: Open-World Robotic Manipulation through Mark-Based Visual Prompting," in *Robotics: Science and Systems*, 2024.

[30] D. Song, J. Liang, X. Xiao, and D. Manocha, "Vl-tgs: Trajectory generation and selection using vision language models in mapless outdoor environments," *IEEE Robotics and Automation Letters*, vol. 10, no. 6, pp. 5791–5798, 2025.

[31] M. Daszykowski and B. Walczak, "Density-Based Clustering Methods," in *Comprehensive Chemometrics*, vol. 2, pp. 565–580, Elsevier, 2009.

[32] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts.," *Proceedings of the National Academy of Sciences*, vol. 93, pp. 1591–1595, feb 1996.

[33] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911, June 2024.

[34] C. Zhang, D. Han, Y. Qiao, J. U. Kim, S. H. Bae, S. Lee, and C. S. Hong, "Faster segment anything: Towards lightweight sam for mobile applications," *arXiv preprint arXiv:2306.14289*, 2023.

[35] B. Yu, Y. Liu, L. Han, H. Kasaei, T. Li, and M. Cao, "VLN-Game: Vision-Language Equilibrium Search for Zero-Shot Semantic Navigation," pp. 1–15, nov 2024.

[36] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A modern library for 3D data processing," *arXiv:1801.09847*, 2018.

[37] A. Visser and J. D. Hoog, "Discussion of multi-robot exploration in communication-limited environments," in *2013 IEEE International Conference on Robotics and Automation*, pp. 1–5, 2013.

[38] M. Juliá, A. Gil, and O. Reinoso, "A comparison of path planning strategies for autonomous exploration and mapping of unknown environments," *Autonomous Robots*, vol. 33, pp. 427–444, nov 2012.

[39] OpenAI, "GPT-4o mini: advancing cost-efficient intelligence." https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/, July 2024. Accessed: 2025-05-06.

[40] OpenAI, "Hello GPT-4o." https://openai.com/index/hello-gpt-4o/, May 2024. Accessed: 2025-05-06.

[41] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "Fast-lio2: Fast direct lidar-inertial odometry," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2053–2073, 2022.

[42] K. Koide, S. Oishi, M. Yokozuka, and A. Banno, "General, single-shot, target-less, and automatic lidar-camera extrinsic calibration toolbox," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11301–11307, 2023.

[43] A. V. Segal, D. Haehnel, and S. Thrun, "Generalized-ICP," in *Robotics: Science and Systems*, vol. 5, pp. 161–168, 2010.