Enhancing Multi-Robot Semantic Navigation Through Multimodal Chain-of-Thought Score Collaboration

Zhixuan Shen, Haonan Luo*, Kexun Chen, Fengmao Lv, Tianrui Li

School of Computing and Artificial Intelligence, Southwest Jiaotong University, China
{shenzx29, chenkexun}@my.swjtu.edu.cn, {lhn, trli}@swjtu.edu.cn, fengmaolv@126.com

Abstract

Understanding how humans cooperatively utilize semantic knowledge to explore unfamiliar environments and decide on navigation directions is critical for house service multi-robot systems. Previous methods primarily focused on single-robot centralized planning strategies, which severely limited exploration efficiency. Recent research has considered decentralized planning strategies for multiple robots, assigning separate planning models to each robot, but these approaches often overlook communication costs. In this work, we propose Multimodal Chain-of-Thought Co-Navigation (MCoCoNav), a modular approach that utilizes multimodal Chain-of-Thought to plan collaborative semantic navigation for multiple robots. MCoCoNav combines visual perception with Vision Language Models (VLMs) to evaluate exploration value through probabilistic scoring, thus reducing time costs and achieving stable outputs. Additionally, a global semantic map is used as a communication bridge, minimizing communication overhead while integrating observational results. Guided by scores that reflect exploration trends, robots utilize this map to assess whether to explore new frontier points or revisit history nodes. Experiments on HM3D_v0.2 and MP3D demonstrate the effectiveness of our approach.

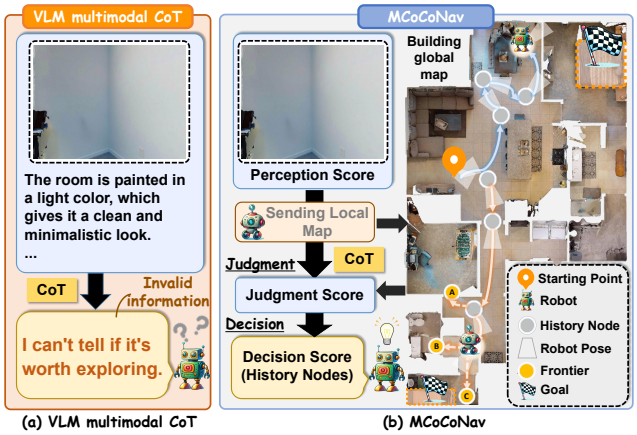

Figure 1: Examples of (a) VLM multimodal CoT reasoning and (b) MCoCoNav cross-image multimodal CoT reasoning. Our MCoCoNav utilizes cross-image multimodal CoT to facilitate robots' simultaneous understanding of scene perspectives and the global semantic map for effective zero-shot multi-robot semantic navigation.

Introduction

The capability to navigate to designated goals is crucial for house service robots, enabling them to effectively find designated objects in unfamiliar indoor environments and complete various subsequent tasks. Consequently, the Object Goal Navigation (ObjectNav) (Du, Yu, and Zheng 2020; Mayo, Hazan, and Tal 2021; Chaplot et al. 2020) task has garnered significant attention. Traditional ObjectNav tasks require robots to be navigated to user-specified categories of objects in an unseen and unmapped environment based on visual observations. Given that the environment is invisible to all robots, they must collaboratively infer potential positions where the goals may appear. This necessitates effective communication and cooperation among multiple robots (e.g., conflict-free communication and global planning after communication), enabling them to make corresponding decisions based on the observed visual cues.

To establish a structured navigation cooperation framework for multiple robots, centralized planning strategies (Zhao et al. 2024; Agashe, Fan, and Wang 2023; Yu, Kasaei, and Cao 2023a; Chen et al. 2023b) map all robots' observations, histories, and other pertinent information to a unified channel, with a single planning model tasked with assigning objectives to each robot group or individual. However, as the complexity of the environment and the number of robots increases, the information processing burden on the planning model significantly escalates. Alternatively, decentralized planning strategies (Chen et al. 2023b; Ying et al. 2024; Liu et al. 2023; Wang, He, and Kantaros 2024) assign each robot an individual "brain" for independent reasoning, allowing the robots to communicate and share information about explored areas similarly to how humans do. These approaches facilitate the connection between new discoveries and previously explored regions, enabling adaptive decisions. By distributing the information generated by multiple robots across various planning models, the burden of decision is alleviated. Nevertheless, the decentralized planning strategy is still limited due to the substantial communication

*Corresponding author

and temporal costs associated.

Utilizing VLMs as tools for multimodal scene understanding and navigation planning provides easily interpretable intermediate representations for robotic systems. Although the narratives generated by VLMs may not be sufficient for navigation, their Chain-of-Thought (CoT) reasoning methods, which simulate human thought processes, can inform or guide the behavior of the underlying navigation stack in multi-robot semantic navigation tasks (e.g., providing intermediate representations for information processing and communication among robots). Consequently, some work (Kuang, Lin, and Jiang 2024; Shah et al. 2023; Ren et al. 2024) has integrated VLM multimodal CoT (Wei et al. 2022; Zhang et al. 2023c; Zheng et al. 2023; Gao et al. 2024) into centralized planning strategies, using the problem decomposition reasoning of multimodal CoT as a heuristic to specify strategies. However, utilizing multimodal CoT for navigation in complex and diverse indoor scenes may be unreliable. As shown in Figure 1 (a), meaningless scene perspectives frequently disrupt the multimodal CoT reasoning process during navigation. In contrast, if both the scene perspectives and the robot's global semantic map in a decentralized planning strategy can be comprehended by multimodal CoT, the reasoning decisions will be more reliable. For instance, the location of the scene perspective on the global semantic map can be used to infer the appropriate answer.

In order to make high-level information from scene perspectives and global semantic map available for multimodal CoT to understand, we propose Multimodal Chain-of-Thought Co-Navigation (MCoCoNav), a novel framework that utilizes multimodal Chain-of-Thought to develop effective exploration and decision strategies for multi-robot navigation in unfamiliar environments. As shown in Figure 1 (b), given the current scene perspective of a robot, the Perception module of MCoCoNav utilizes multimodal CoT to evaluate its exploration value, predicting the probability of "Yes" as an exploration score. Considering the significant communication costs incurred during navigation planning, MCoCoNav uses the global semantic map as the direct communication bridge between robots, avoiding additional communication cost overhead. All robots jointly maintain a global semantic map that integrates all observations of the unseen environment. Furthermore, to fully leverage the role of history nodes in the global semantic map, for each node being explored by a robot, we examine the VLM's inclination to explore frontier points and design metrics to calculate the horizontal field of view score and history score of the node. The former indicates the robot's tendency to explore the current node, and the latter indicates the exploration possibilities of different history nodes, both of which are synchronized with the update of the global semantic map. Subsequently, each robot selects the frontier point with the highest prediction probability of the Decision module or the history node with the highest history score as its long-term navigation goal on the global semantic map.

Extensive experiments on benchmark navigation metrics and evaluations on HM3D_v0.2 and MP3D demonstrate that MCoCoNav efficiently plans collaborative exploration of multiple robots in unfamiliar indoor environments, outperforming other multi-robot methods based on centralized and decentralized planning strategies. Our contributions can be summarized as follows: (1) We design an innovative planning framework for the multi-robot semantic navigation task that enables local, small-scale VLMs to guide multiple robots through unknown environments for efficient exploration and decision. (2) Our proposed cross-image multimodal CoT facilitates robots' understanding of high-level information from different images and enables low-cost semantic information sharing among robots. (3) Evaluations on HM3D_v0.2 and MP3D demonstrate the superior performance of MCoCoNav, which is fully zero-shot and can be deployed locally at low cost.

Related Work

Zero-Shot Object Goal Navigation Finding specified objects in unfamiliar indoor environments is a widely applied task for embodied intelligent robots. While numerous works have utilized reinforcement learning (Ye et al. 2021; Chang, Gupta, and Gupta 2020), learning from demonstrations (Ramrakhya et al. 2023), or waypoint planners (Chen et al. 2023a; Chaplot et al. 2020; Zhang et al. 2021; Luo et al. 2022; Ramakrishnan et al. 2022; Zhang et al. 2023b) to train robots with semantic navigation capabilities, these task-specific training methods are difficult to generalize to diverse real-world environments. Consequently, many researchers have begun to investigate Zero-Shot Object Navigation (ZSON) techniques. ZSON (Majumdar et al. 2022) and COWs (Gadre et al. 2023) employ CLIP (Radford et al. 2021) features or open-vocabulary object detectors to locate goal objects. Recent studies such as LGX (Dorbala, Mullen Jr, and Manocha 2023), ESC (Zhou et al. 2023), L3mvn (Yu, Kasaei, and Cao 2023b) and Voronav (Wu et al. 2024) involve using Large Language Models (LLMs) for reasoning and decision. However, these works only consider a single robot decision framework and rely on powerful remote large foundation models like GPT-4V (Yang et al. 2023). In contrast, we use only a quantized 9B lightweight VLM to formulate multi-robot collaborative decision strategies.

LLM-based Multi-Robot Systems Compared to more mature approaches such as active mapping (Ye et al. 2022), reinforcement learning planning (Yu et al. 2022), and methods based on prior knowledge (Liu et al. 2022), the application of LLMs in Multi-Robot Systems (MRS) is still in its early stages. Nevertheless, a few contemporary studies have begun to explore the use of generative models in MRS task planning. Some research adopts the centralized planning strategy, where the LLM needs to simultaneously comprehend the observations, histories, and task states of multiple robots and allocate collaborative tasks to each robot. Specifically, Co-NavGPT (Yu, Kasaei, and Cao 2023a) utilizes a single LLM to assign exploration frontiers to each robot. HAS (Zhao et al. 2024) automatically organizes LLM-based robot groups to complete navigation tasks in complex Minecraft environments. Furthermore, other studies employ the decentralized planning strategy, treating each robot as a separate entity that exchanges information through methods similar to human thinking and

communication and makes independent decisions. For instance, CoELA (Zhang et al. 2023a) provides a systematic template for decentralized communication and collaboration, while GOMA (Ying et al. 2024) frames the language interaction between robots and humans as a planning problem.

Multimodal CoT Reasoning Effective prompting techniques are crucial for fully harnessing the capabilities of LLMs and VLMs. Recent advancements (Diao et al. 2023; Ho, Schmid, and Yun 2022; Wang et al. 2022; Zhang et al. 2022) have introduced CoT to further enhance the reasoning abilities of LLMs. Concurrently, researcher (Zhang et al. 2023c; Zheng et al. 2023; Gao et al. 2024)s have focused on maximizing the multimodal reasoning potential of LLMs and VLMs through multimodal CoT. Our work transcends the exploration of simple CoT reasoning strategies, focusing instead on information exchange and collaborative navigation among multiple robots using multimodal CoT.

Method

Problem Formulation

Multi-Robot Semantic Navigation In the Multi-Robot Semantic Navigation task, the set of scenes can be denoted as $S = \{s_1, ..., s_k\}$, and the set of categories as $C = \{c_1, ..., c_m\}$. In each episode, n robots $R = \{r_1, ..., r_n\}$ are randomly placed in an unfamiliar and invisible scene s_i, with the objective of locating an object of category c_i. Each robot is allowed a maximum of 500 time steps per episode. At each time step t, robot r_i obtains a current observation O, which comprises an RGB-D image, the robot's location, and pose. The action space A consists of six discrete actions: `move_forward`, `turn_left`, `turn_right`, `look_up`, `look_down`, and `stop`. When executing the `move_forward` action, the robot moves $25cm$ ahead, while for `turn_left`, `turn_right`, `look_up`, or `look_down`, the robot rotates 30° in the corresponding direction. The robot must select and execute an action from the action space A. The stop action is triggered when one of the robots r_i approaches the goal object. An episode is considered successful if the distance between robot r_i and the goal is less than 0.1m and robot r_i executes the stop action. In this task robots are not allowed to use purposefully trained models and panoramic 360° FoV sensors.

VLM Predictions VLMs that have undergone large-scale pretraining provide the necessary reasoning capabilities to address the multi-robot semantic navigation task. Given an RGB image I and text T, we query the VLM to predict the probability of the next token y. Specifically, we denote the VLM's prediction as $\hat{f}_y(I, T) \in [0, 1]^{|Y|}$, which represents the softmax scores over a selection set Y.

Overview

Our MCoCoNav methodology is illustrated in Figure 2. The core of the MCoCoNav is the MCoCoNav Planner, which consists of three main components. We begin by employing the Perception module to analyze whether each robot's current perspective warrants exploration, assigning the exploration score. Concurrently, the Judgment module evaluates whether the robot should explore predicted frontier points or return to history nodes for re-exploration based on the global semantic map of all robots, history nodes, predicted frontier points and trajectory information, assigning the judgment score. Subsequently, we combine the exploration and judgment scores, which are weighted to produce the horizontal field of view score and history score for the current history node. Suppose the horizontal field of view score exceeds or is equal to 0.5. In that case, the Decision module integrates information from the Perception and Judgment module to select an appropriate frontier point as the long-term navigation goal for the next time step. If the horizontal field of view score is below 0.5, the robot's long-term navigation goal for the next time step is set to the history node with the highest history score. With long-term navigation goals established for each robot, we logically analyze the navigational feasibility of these goals and employ the local policy to guide each robot in exploration and goal acquisition.

MCoCoNav Planner

The MCoCoNav Planner is a **Decentralized Planning** system that interacts with a single scene view, a single navigation goal text, a global semantic map, and global navigation history nodes to perform long-term goal prediction for individual robot navigation. In this section, we will introduce the planning process of the MCoCoNav Planner.

Preliminaries For robot r_i at time step t, the obtained observation O includes the scene view I_i^t, scene semantic map S_i^t, scene depth map D_i^t and pose p. We first process the visual information of individual robots using Visual Foundation Models (VFMs) to obtain object detection information $\mathcal{O}_i^t$ and semantic maps $\mathcal{M}^t$ for MCoCoNav Planner's Visual Perception and Global Map Exploration Judgment. We provide details of object detection and semantic mapping in Appendix A.1.

Visual Perception We guide the VLM's prediction through a simple multimodal CoT to obtain the exploration score (ES) for the current perspective. Firstly, we use the Perception Instruction $\mathcal{I}_P$, a query template containing the query *Object List*, *Spatial Relationships* and *Additional Context*, to prompt the Perception VLM to provide a description of the spatial relationships in the scene view I_i^t. Specifically, we query the Perception VLM: *"Are they to your left, right, in front of you, or behind you?"* and *"Are they on the floor, mounted on the wall, or placed on top of another piece of furniture?"*, and *"How close are they to you or to each other?"*. Subsequently, we combine VLM output text $VLM(I_i^t, \mathcal{I}_P)$ with the object detection information $\mathcal{O}_i^t$ in pure natural language to form the multimodal CoT to infer whether the current scene is worth further exploration by using the Exploration Instruction $\mathcal{I}_E$ containing *Target of Navigation*, *Scene Objects* and *Decision Criteria* information. It is noteworthy that we only extract the normalized probability output in the "Yes" direction from this CoT to construct

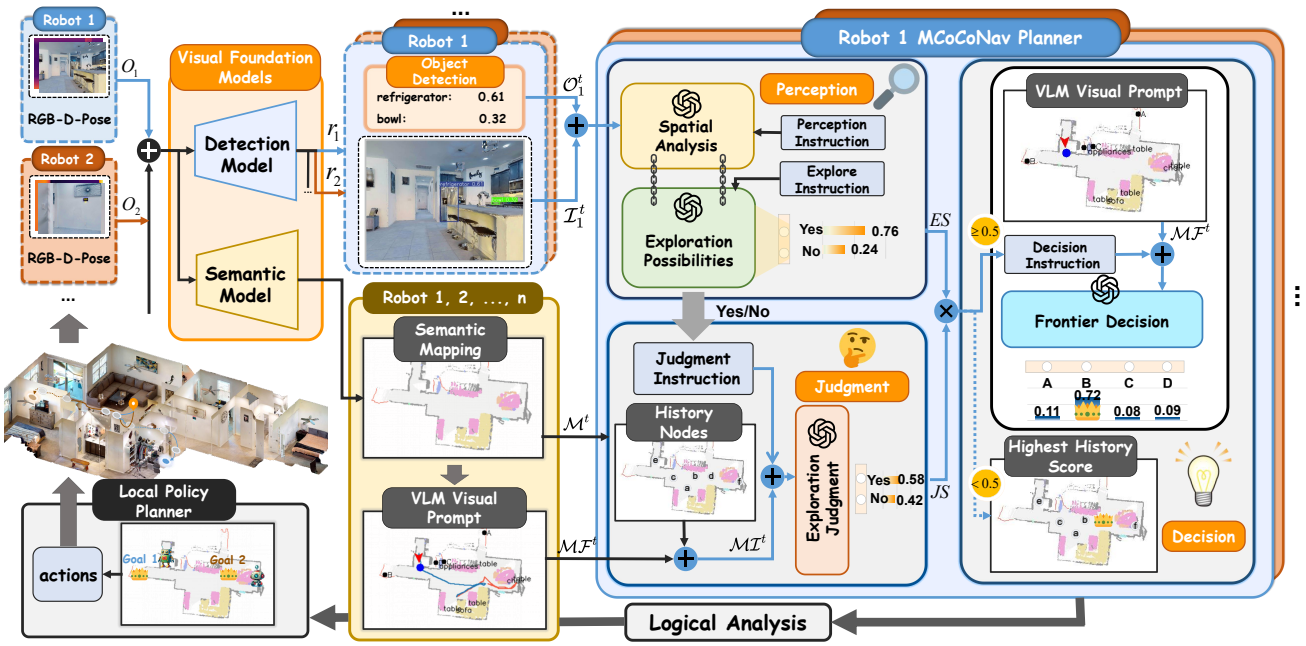

Figure 2: **Components of MCoCoNav.** The MCoCoNav architecture consists of Visual Foundation Models, the MCoCoNav Planner and the Local Policy Planner. At its core is the MCoCoNav Planner, which is composed of three main components: the Perception module, the Judgment module and the Decision module.

the $ES^t_{i,p}$ for the current pose p:

$$ES^t_{i,p} = \hat{f}_{"Yes"}(I^t_i, VLM(I^t_i, \mathcal{I}_P) + \mathcal{O}^t_i + \mathcal{I}_E) \quad (1)$$

Global Map Exploration Judgment Now, we will detail the Judgment module. If the global map $\mathcal{M}^t$ has just been initialized, all robots are set to continue exploring. Otherwise, prompts need to be constructed for Judgment VLM to determine whether it is worthwhile for the robots to continue exploring.

To obtain the visual prompt, first, we annotate the current position coordinates and pose of robot r_i, as well as its last long-term navigation goal, with a red arrow and a blue dot respectively. Simultaneously, we sample the coordinates of all robots' history nodes with lowercase letters in chronological order (in the rare event that the number of history nodes exceeds 26, we use lowercase letters followed by the sequence number minus 26 for those beyond 26). Then, we annotate the predicted frontier points with uppercase letters in the same manner. Finally, we map all annotation information onto the global semantic map $\mathcal{M}^t$, resulting in the visual prompt $\mathcal{MI}^t$ that contains semantic and historical information for all robots.

To obtain the textual prompt, we construct the Judgment Instruction $\mathcal{I}_J$ by incorporating the input-output chain from the Perception module with a query template that includes information on *Frontier Points*, *History Nodes* and *Location and Previous Movement*. Unlike the Perception module, the "yes" output of the VLM corresponds to "explore a frontier point". The judgment score ($JS^t_{i,p}$) of the current pose p can be represented as:

$$JS^t_{i,p} = \hat{f}_{"Yes"}(\mathcal{MI}^t, \mathcal{I}_J) \quad (2)$$

We apply temperature scaling (τ_{ES} and τ_{JS}) for judgment score and exploration score and compute the current horizontal field of view score $HFOVS^t_{i,hfov}$:

$$HFOVS^t_{i,hfov} = exp(\tau_{ES} \cdot JS^t_{i,p} + \tau_{JS} \cdot JS^t_{i,p}) \quad (3)$$

where $hfov$ represents the current horizontal field of view. If $HFOVS^t_{i,hfov}$ is greater than or equal to 0.5, robot r_i continues to select frontier points for exploration. Otherwise, robot r_i returns to history nodes for re-exploration.

Cross-image Multimodal CoT Decision We have been discussing the utilization of VLM to judge exploration possibilities. However, we haven't addressed the most crucial step in completing exploration—determining the next long-term navigation goal, which involves selecting appropriate frontier points or history nodes. VLM's evaluation of suitable long-term navigation goals depends on the current position of robot r_i and the history scores (HS) of all robots' history nodes.

We get the history score HS in the following steps. First, suppose r_i is at a distance greater than or equal to 25 pixels from the nearest history node. In that case, we construct a 360-dimensional zero vector representing the state score of the robot's current position and fill the $HFOVS^t_{i,hfov}$ of robot r_i into this vector corresponding to the robot's angle of horizontal view. We then sum the entire vector and divide by the number of times all robots have explored within a 25-pixel range of the current position, obtaining the HS for

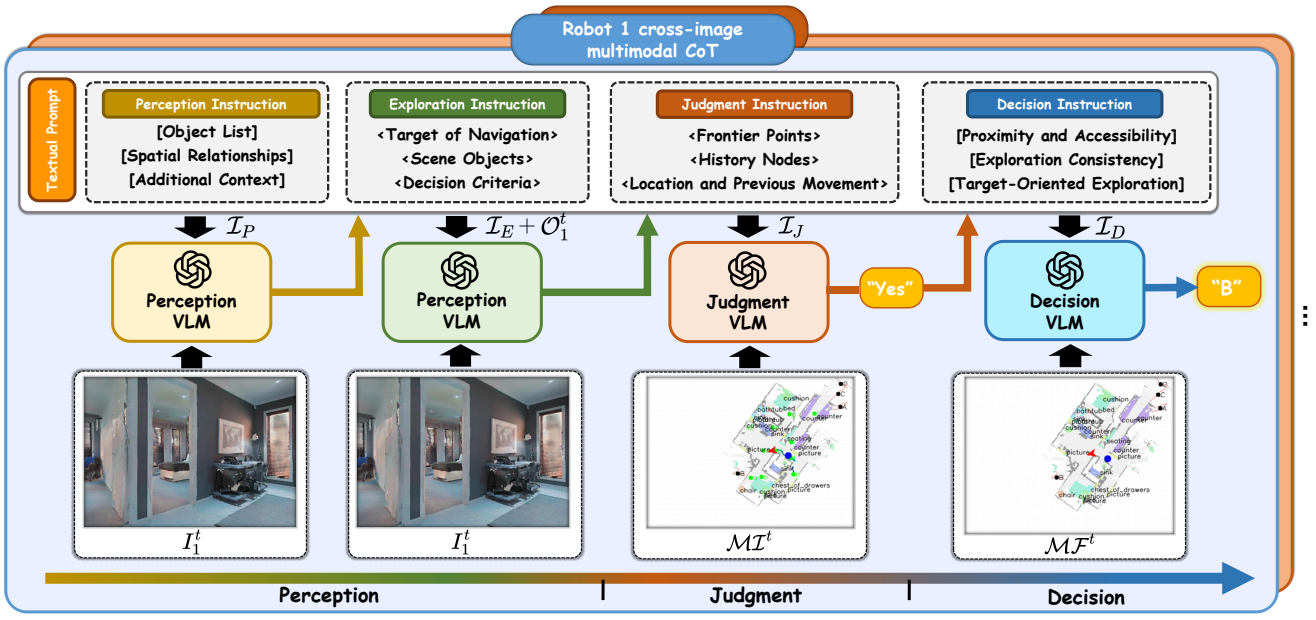

Figure 3: **Workflow of cross-image multimodal CoT.** Information from multiple images is unified into a multimodal CoT through VLM. At each time step, the next action for each robot and the global semantic map are updated based on the observed scene objects. The cross-image multimodal CoT enables semantic label alignment for navigation planning, environmental understanding, and common-sense reasoning.

the current position. If robot r_i is less than 25 pixels from the nearest history node, the $HFOVS_{i,hfov}^t$ is filled into the state score of that history node and the HS is updated. For the readers' convenience, we present the complete history score algorithm in Appendix A.2. After obtaining the HS, if the robot r_i is selected to return to the history nodes, we will choose the history node with the highest HS as the long-term navigation goal. Otherwise, we will construct visual and textual prompts for Decision VLM to select the appropriate frontier.

For the visual prompt of Decision VLM, we remove the history nodes annotations from the visual prompt $\mathcal{MI}^t$ of Judgment VLM, retaining only the annotations for frontier points. For the textual prompt of Decision VLM, we construct the Decision Instruction I_D by integrating the input-output from the Judgment VLM with a query template that includes considerations of *Proximity and Accessibility*, *Exploration Consistency*, and *Target-Oriented Exploration*. As shown in Figure 3, the continuous multimodal CoT ensures that the VLM can fully comprehend information from multiple images, alleviating the potential invalid information that might arise from a single CoT decision. To derive the final decision result, we limit the VLM's output to four or fewer options representing the frontier points "A", "B", "C" and "D", and then obtain the normalized probabilities for these four directions:

$$DS_i^t = \hat{f}_X(\mathcal{MF}^t, \mathcal{I}_D) \tag{4}$$

where $\mathcal{MF}^t$ represents the visual prompt at time step t that includes annotations of frontier points, DS signifies the decision score corresponding to the frontier points, and X is the uppercase letter representing the respective frontier points. We select the frontier point represented by the uppercase letter with the highest decision score as the long-term navigation goal.

Logical Analysis and Local Policy

Even though the long-term navigation goal has been determined, the aforementioned process does not consider two logic issues during navigation: (1) The continuity of exploration towards the long-term navigation goal. (2) Potential collisions.

To address the first issue, each time the MCoCoNav Planner execution ends, we assess whether the robot has reached the previous long-term navigation goal. If the robot is at a distance of 25 pixels or more from the last long-term navigation goal and the $HFOVS_{i,hfov}^t$ is less than 0.5, we consider that the robot has not reached the goal and that the current position is not worth exploring. In this case, the robot needs to continue exploring towards the previous long-term navigation goal. For the second issue, we record the relative distance between the current robot position and the position at the end of the last MCoCoNav Planner execution. If this distance is less than 25 pixels, we consider that the robot is trapped in a location, experiencing collisions with nearby objects or obstacles over several time steps. At this point, the long-term navigation goal will be reset to a randomly sampled point on the global semantic map M^t.

Finally, we utilize a control strategy based on the Fast Marching Method (FMM) (Sethian 1999) to output a low-

Method	Zero-Shot	LLM/VLM	HM3D_v0.2		MP3D	
			SPL↑	SR↑	SPL↑	SR↑
Greedy (Visser et al. 2013)	✓	None	0.328	0.611	0.225	0.406
Cost-Utility (Juliá, Gil, and Reinoso 2012)	✓	None	0.323	0.625	0.212	0.419
Random Sampling	✓	None	0.336	0.636	0.281	0.435
Multi-SemExp (Chaplot et al. 2020)	✗	None	0.327	0.612	-	-
Co-NavGPT (Yu, Kasaei, and Cao 2023a)	✓	GPT-3.5 (Remote)	0.331	0.661	-	-
MCoCoNav (Ours)	✓	GLM-4V (Local)	**0.387**	**0.716**	**0.334**	**0.568**
Co-NavGPT (GT-Seg)	✓	GPT-3.5 (Remote)	0.448	0.757	-	-
MCoCoNav (GT-Seg)	✓	GLM-4V (Local)	**0.464**	**0.872**	**0.410**	**0.655**

Table 1: **Comparison of different baselines for 2-robot on HM3D_v0.2 and MP3D.** MCoCoNav significantly outperforms all baseline methods on all metrics and achieves zero-shot multi-robot semantic navigation for local planning.

level action a_i^t for each robot, which concludes the loop and moves to the next time step $t + 1$.

Experiment

Experimental Setup

Datasets. We evaluate our approach using the Habitat (Savva et al. 2019) simulator and validate it on two different real-world environment 3D scan datasets: HM3D_v0.2 (Ramakrishnan et al. 2021) and MP3D (Chang et al. 2017). The validation split for HM3D_v0.2 includes 1000 episodes, spanning 36 scenes and 6 object categories, while the validation split for MP3D includes 2195 episodes, spanning 11 scenes and 21 object categories.

Metrics. We employ the navigation success rate (SR) and the success rate weighted by navigation path length (SPL) as evaluation metrics. **SR** represents the percentage of successful episodes out of the total number of episodes. **SPL** is calculated as the inverse ratio of the actual path length to the optimal path length weighted by the success rate.

Implementation Details. We employ YOLOV10m(Wang et al. 2024) as the Detection model and RedNet(Jiang et al. 2018) as the Semantic model. All VLMs utilize the locally deployed INT4-quantized GLM-4V-9B (GLM et al. 2024). To learn more about the details of the experiments and the VLM Prompt setup, see Appendix B.

Baselines

We evaluate MCoCoNav by comparing it with several multi-robot Semantic Navigation baselines: Greedy (Visser et al. 2013), Cost-Utility (Juliá, Gil, and Reinoso 2012), Random Samples, Multi-SemExp (Chaplot et al. 2020), and Co-NavGPT (Yu, Kasaei, and Cao 2023a). In the Greedy strategy, each robot selects its nearest designated frontier as its goal location. The Cost-Utility strategy conducts a cost-utility assessment for each frontier cell after obtaining the frontiers, choosing the highest-scoring frontier as the goal location. The Random Samples strategy randomly samples long-term navigation goals on the map. Multi-SemExp is a baseline extended from (Chaplot et al. 2020) for multi-robot settings. Co-NavGPT employs a centralized planning approach, encoding the explored environmental data as prompts for LLM.

Results and Analysis

Comparison with baseline methods The performance of MCoCoNav with 2-robot in comparison to other methods on the HM3D_v0.2 and MP3D datasets is summarized in Table 1. Firstly, compared to Multi-SemExp, which is trained end-to-end on local maps, MCoCoNav increases +6.0% SPL and +10.4% SR on HM3D_v0.2. The results demonstrate that our method constructs a global map shared by the multi-robot system, enriching the semantic information of the environment and thereby achieving better performance. Compared to other zero-shot methods, such as Co-NavGPT, MCoCoNav achieves +5.6% SPL and +5.5% SR on HM3D_v0.2. Meanwhile, +5.3% SPL and +13.3% SR on MP3D compared to Random Sampling. Our MCoCoNav method further improves the navigation process by constructing a comprehensive analysis framework for scene images and global map information, striking a balance between exploration and return, and is better adapted to untrained complex environments and navigation goals.

Furthermore, we evaluate the importance of navigation performance alone without considering the accuracy of semantic segmentation. We replace the semantic segmentation algorithm with ground-truth (GT-Seg) and achieve an increase of +1.6% SPL and +11.5% SR on HM3D_v0.2 compared to Co-NavGPT and +12.9% SPL and +22.0% SR on MP3D compared to Random Sampling. This establishes new state-of-the-art metrics for these datasets. Figure 4 illustrates a successful case of MCoCoNav navigating to the goal "TV" using multimodal CoT, visualizing the observations of four key global decisions along with the synchronously updated details of the VLM Visual Prompt and global semantic map.

Effect of the number of robots and semantic accuracy We investigate the impact of varying robot numbers and semantic accuracy on navigation performance on HM3D_v0.2. As shown in Table 2, significant differences are observed under different semantic segmentation conditions. We attribute the limitations of VLM in addressing the multi-robot semantic navigation task to two primary factors: the accuracy of semantic segmentation and the quality of 3D scanning. From the results, we observed that both SR and SPL increase

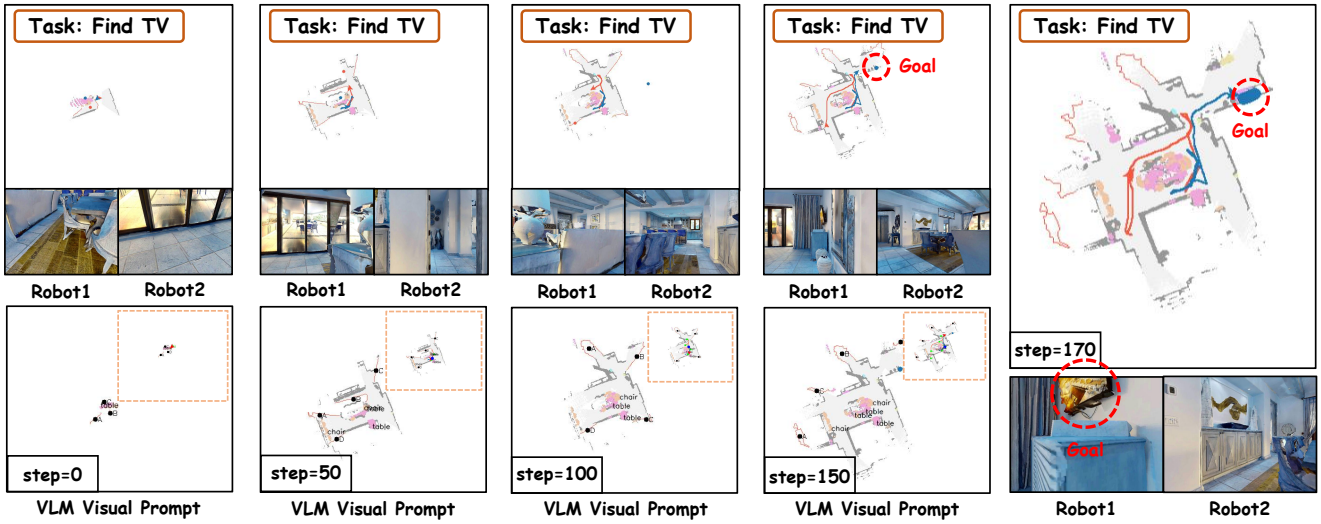

Figure 4: **2-Robot navigation episode with MCoCoNav in HM3D_v0.2.** The top row shows the robots' RGB scene views and the global semantic map. The bottom row shows the VLM Visual Prompt used for Decision (large image) and the VLM Visual Prompt used for Judgment (small image). Best viewed when zoomed in.

Seg.	Robot Num.	T ↓	DT↓	SPL↑	SR↑
Pred-Seg	1	744s	132	0.297	0.634
Pred-Seg	2	883s	156	0.387	0.716
Pred-Seg	3	1110s	192	0.442	0.720
GT-Seg	1	342s	0	0.298	0.736
GT-Seg	2	582s	0	0.464	0.872
GT-Seg	3	961s	0	0.485	0.911

Table 2: **Effect of the number of robots and semantic accuracy on HM3D_v0.2.** Pred-Seg represents the semantic segmentation predicted by the model and DT denotes the number of the Detection Trap. T is the average time spent per episode.

Modules					SPL↑	SR↑
Per	PCoT	Jud	Dec	Log		
			✓		0.332	0.676
✓			✓		0.349	0.677
✓	✓		✓		0.353	0.698
		✓	✓		0.341	0.692
✓	✓	✓	✓		0.381	0.707
				✓	0.286	0.648
✓	✓	✓	✓	✓	**0.387**	**0.716**

Table 3: **Ablation Study.** Per means the Perception module without multimodal CoT and PCoT denotes the Perception module's multimodal CoT. Jug represents the Judgment module; Dec is the Decision module. Log denotes Logic Analysis.

with the number of robots, although the rate of improvement gradually decreases. However, the navigation time cost significantly increases. The frequency of detection traps (DT), where robots incorrectly identify the wrong objects as goals, also rises with an increasing number of robots. Notably, when there are two robots, the improvements in SR and SPL are the greatest, while the increase in detection traps and navigation time is minimal. In addition, since the 3D scan quality of the MP3D is significantly lower than that of the HM3D_v0.2, its performance in Table 1 is lower than that of the HM3D_v0.2.

Ablation Study We present the impact of the three modules and Logical Analysis on MCoCoNav's performance in Table 3. The third row of Table 3 shows that the multimodal CoT of the Perception module is effective, which can alleviate the invalid information that VLM's prediction may cause. Additionally, the Logical Analysis module plays a crucial role in resolving navigation logic problems such as collisions, and when it works together with the other three modules, the performance is significantly improved. The ablation studies demonstrate the effectiveness of each component in our method.

Conclusion

We propose MCoCoNav, a novel framework leveraging multimodal Chain-of-Thought for efficient exploration and decision in multi-robot semantic navigation. To reduce communication overhead, MCoCoNav utilizes a global semantic map for direct communication among robots. Robots maintain the shared map, integrating observations and using node scores to guide exploration decisions. Experimental results demonstrate MCoCoNav's superior performance compared to previous multi-robot methods.

Acknowledgments

The authors express their gratitude to the National Natural Science Foundation of China (62306247), China Postdoctoral Science Foundation (2022M722630), the Natural Science Foundation of Sichuan Province (2024NSFSC1474, 2024ZHCG0166).

References

Agashe, S.; Fan, Y.; and Wang, X. E. 2023. Evaluating multi-agent coordination abilities in large language models. *arXiv preprint arXiv:2310.03903*.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Chang, M.; Gupta, A.; and Gupta, S. 2020. Semantic visual navigation by watching youtube videos. *Advances in Neural Information Processing Systems*, 33: 4283–4294.

Chaplot, D. S.; Gandhi, D. P.; Gupta, A.; and Salakhutdinov, R. R. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33: 4247–4258.

Chen, J.; Li, G.; Kumar, S.; Ghanem, B.; and Yu, F. 2023a. How To Not Train Your Dragon: Training-free Embodied Object Goal Navigation with Semantic Frontiers. In *Robotics: Science and Systems*.

Chen, Y.; Arkin, J.; Zhang, Y.; Roy, N.; and Fan, C. 2023b. Scalable Multi-Robot Collaboration with Large Language Models: Centralized or Decentralized Systems? *CoRR*, abs/2309.15943.

Diao, S.; Wang, P.; Lin, Y.; and Zhang, T. 2023. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*.

Dorbala, V. S.; Mullen Jr, J. F.; and Manocha, D. 2023. Can an embodied agent find your "cat-shaped mug"? llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*.

Du, H.; Yu, X.; and Zheng, L. 2020. Learning object relation graph and tentative policy for visual navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, 19–34. Springer.

Gadre, S. Y.; Wortsman, M.; Ilharco, G.; Schmidt, L.; and Song, S. 2023. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23171–23181.

Gao, T.; Chen, P.; Zhang, M.; Fu, C.; Shen, Y.; Zhang, Y.; Zhang, S.; Zheng, X.; Sun, X.; Cao, L.; et al. 2024. Cantor: Inspiring Multimodal Chain-of-Thought of MLLM. *arXiv preprint arXiv:2404.16033*.

GLM, T.; Zeng, A.; Xu, B.; Wang, B.; Zhang, C.; Yin, D.; Rojas, D.; Feng, G.; Zhao, H.; Lai, H.; et al. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*.

Ho, N.; Schmid, L.; and Yun, S.-Y. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Jiang, J.; Zheng, L.; Luo, F.; and Zhang, Z. 2018. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*.

Juliá, M.; Gil, A.; and Reinoso, O. 2012. A comparison of path planning strategies for autonomous exploration and mapping of unknown environments. *Autonomous Robots*, 33: 427–444.

Kuang, Y.; Lin, H.; and Jiang, M. 2024. OpenFMNav: Towards Open-Set Zero-Shot Object Navigation via Vision-Language Foundation Models. *arXiv preprint arXiv:2402.10670*.

Liu, J.; Yu, C.; Gao, J.; Xie, Y.; Liao, Q.; Wu, Y.; and Wang, Y. 2023. Llm-powered hierarchical language agent for real-time human-ai coordination. *arXiv preprint arXiv:2312.15224*.

Liu, X.; Guo, D.; Liu, H.; and Sun, F. 2022. Multi-agent embodied visual semantic navigation with scene prior knowledge. *IEEE Robotics and Automation Letters*, 7(2): 3154–3161.

Luo, H.; Yue, A.; Hong, Z.-W.; and Agrawal, P. 2022. Stubborn: A strong baseline for indoor object navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3287–3293. IEEE.

Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; and Batra, D. 2022. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35: 32340–32352.

Mayo, B.; Hazan, T.; and Tal, A. 2021. Visual navigation with spatial attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16898–16907.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramakrishnan, S. K.; Chaplot, D. S.; Al-Halah, Z.; Malik, J.; and Grauman, K. 2022. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18890–18900.

Ramakrishnan, S. K.; Gokaslan, A.; Wijmans, E.; Maksymets, O.; Clegg, A.; Turner, J.; Undersander, E.; Galuba, W.; Westbury, A.; Chang, A. X.; et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

Ramrakhya, R.; Batra, D.; Wijmans, E.; and Das, A. 2023. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17896–17906.

Ren, A. Z.; Clark, J.; Dixit, A.; Itkina, M.; Majumdar, A.; and Sadigh, D. 2024. Explore until Confident: Efficient Exploration for Embodied Question Answering. *arXiv preprint arXiv:2403.15941*.

Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9339–9347.

Sethian, J. A. 1999. Fast marching methods. *SIAM review*, 41(2): 199–235.

Shah, D.; Equi, M. R.; Osiński, B.; Xia, F.; Ichter, B.; and Levine, S. 2023. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, 2683–2699. PMLR.

Visser, A.; De Hoog, J.; Jiménez-González, A.; and de Dios, J.-R. M. 2013. Discussion of multi-robot exploration in communication-limited environments. In *Workshop" Towards Fully Decentralized Multi-Robot Systems: Hardware, Software and Integration" at the ICRA Conference*. Citeseer.

Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; and Ding, G. 2024. Yolov10: Real-time end-to-end object detection. *arXiv preprint arXiv:2405.14458*.

Wang, J.; He, G.; and Kantaros, Y. 2024. Safe Task Planning for Language-Instructed Multi-Robot Systems using Conformal Prediction. *arXiv preprint arXiv:2402.15368*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Wu, P.; Mu, Y.; Wu, B.; Hou, Y.; Ma, J.; Zhang, S.; and Liu, C. 2024. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*.

Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.-C.; Liu, Z.; and Wang, L. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1): 1.

Ye, J.; Batra, D.; Das, A.; and Wijmans, E. 2021. Auxiliary tasks and exploration enable objectnav. *arXiv preprint arXiv:2104.04112*.

Ye, K.; Dong, S.; Fan, Q.; Wang, H.; Yi, L.; Xia, F.; Wang, J.; and Chen, B. 2022. Multi-robot active mapping via neural bipartite graph matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14839–14848.

Ying, L.; Jha, K.; Aarya, S.; Tenenbaum, J. B.; Torralba, A.; and Shu, T. 2024. GOMA: Proactive Embodied Cooperative Communication via Goal-Oriented Mental Alignment. *arXiv preprint arXiv:2403.11075*.

Yu, B.; Kasaei, H.; and Cao, M. 2023a. Co-NavGPT: Multi-robot cooperative visual semantic navigation using large language models. *arXiv preprint arXiv:2310.07937*.

Yu, B.; Kasaei, H.; and Cao, M. 2023b. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3554–3560. IEEE.

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

Zhang, H.; Du, W.; Shan, J.; Zhou, Q.; Du, Y.; Tenenbaum, J. B.; Shu, T.; and Gan, C. 2023a. Building cooperative embodied agents modularly with large language models. *arXiv preprint arXiv:2307.02485*.

Zhang, J.; Dai, L.; Meng, F.; Fan, Q.; Chen, X.; Xu, K.; and Wang, H. 2023b. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6672–6682.

Zhang, S.; Song, X.; Bai, Y.; Li, W.; Chu, Y.; and Jiang, S. 2021. Hierarchical object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 15130–15140.

Zhang, Z.; Zhang, A.; Li, M.; and Smola, A. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023c. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhao, Z.; Chen, K.; Guo, D.; Chai, W.; Ye, T.; Zhang, Y.; and Wang, G. 2024. Hierarchical auto-organizing system for open-ended multi-agent navigation. *arXiv preprint arXiv:2403.08282*.

Zheng, G.; Yang, B.; Tang, J.; Zhou, H.-Y.; and Yang, S. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 5168–5191.

Zhou, K.; Zheng, K.; Pryor, C.; Shen, Y.; Jin, H.; Getoor, L.; and Wang, X. E. 2023. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, 42829–42842. PMLR.