

RATE-Nav: Region-Aware Termination Enhancement for Zero-shot Object Navigation with Vision-Language Models

Junjie Li^{1,2}, Nan Zhang², Xiaoyang Qu², Kai Lu¹, Guokuan Li^{1,†},
Jiguang Wan¹ and Jianzong Wang^{2,†}

¹Huazhong University of Science and Technology, Wuhan, China,

²Ping An Technology (Shenzhen) Co., Ltd., Shenzhen, China

Correspondence: liguokuan@hust.edu.cn and jzwang@188.com

Abstract

Object Navigation (ObjectNav) is a fundamental task in embodied artificial intelligence. Although significant progress has been made in semantic map construction and target direction prediction in current research, redundant exploration and exploration failures remain inevitable. A critical but underexplored direction is the timely termination of exploration to overcome these challenges. We observe a diminishing marginal effect between exploration steps and exploration rates and analyze the cost-benefit relationship of exploration. Inspired by this, we propose RATE-Nav, a Region-Aware Termination-Enhanced method. It includes a geometric predictive region segmentation algorithm and region-Based exploration estimation algorithm for exploration rate calculation. By leveraging the visual question answering capabilities of visual language models (VLMs) and exploration rates enables efficient termination. RATE-Nav achieves a success rate of 67.8% and an SPL of 31.3% on the HM3D dataset. And on the more challenging MP3D dataset, RATE-Nav shows approximately 10% improvement over previous zero-shot methods.

1 Introduction

Object navigation (Dang et al., 2023a; Campari et al., 2020), as one of the core capabilities of embodied agents, aims to enable agents to autonomously locate and navigate to specified target objects in unknown environments (Batra et al., 2020; Sun et al., 2024). This capability is crucial for practical applications such as service robots, for instance, in performing fetch tasks in home environments. However, in real-world applications, robots often operate in unknown and dynamically changing environments, and search targets are not limited to predefined categories, posing significant

[†]Jianzong Wang and Guokuan Li are joint corresponding authors.

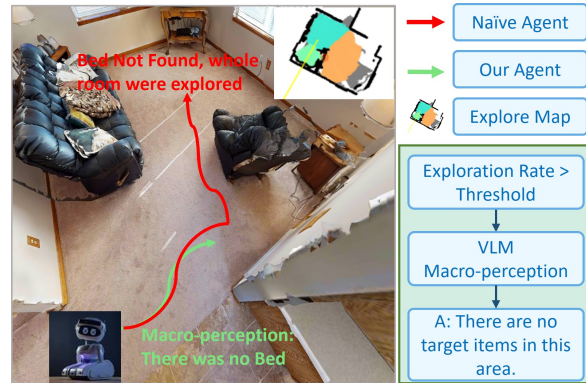


Figure 1: The main difference between our method and existing approaches is the exploration strategy. Naive agent fully explores the current region before moving to the next. In contrast, our agent uses the common-sense reasoning of VLMs to decide whether to continue exploring the current region after achieving a certain exploration level.

challenges to the zero-shot generalization ability of navigation systems (Majumdar et al., 2022).

With the rapid development of Large Language Models (LLMs) (Achiam et al., 2023) and Vision-Language Models (VLMs) (Radford et al., 2021; Li et al., 2023; Liu et al.) with powerful reasoning capabilities in the field of artificial intelligence, researchers (Zhou et al., 2023; Shah et al., 2023) have turned to employing these models for intermediate goal planning: first, by modeling environmental observations with VLMs to construct semantically rich scene representations, and then utilizing the commonsense reasoning abilities embedded in LLMs to predict the possible locations of target objects. Methods (Zhou et al., 2023; Yin et al., 2024) based on LLMs and VLMs have demonstrated notable advantages navigation strategies, thereby enhancing the performance of navigation systems to some extent.

Existing research relies on comprehensive exploration, but there remains a research gap in the

design of exploration termination strategies. As shown in **Figure 1**, traditional navigation strategies struggle to accurately assess the current exploration state when searching specific areas, typically requiring a complete search of the current area before moving to a new one. From a marginal utility perspective, this strategy is inefficient: in the early stages of information gathering, the investment yields significant returns, but as exploration deepens, the marginal value of new information gained from each operation gradually diminishes. This prompts us to consider whether the search in the current area could be terminated earlier. To validate this hypothesis, we conducted hundreds of navigation experiments on the HM3D dataset, thoroughly analyzing the relationship between exploration cost and benefit. Detailed analysis can be found in **Chapter 3.1**.

Furthermore, existing research primarily relies on exploration maps constructed by visual perception to identify target-free regions. Precision errors and model limitations make it difficult to fully annotate these regions on the map. As shown in **Figure 3**, a situation may occur where a region is largely explored, but repeated boundary settings are triggered due to a small unknown area. This results in redundant exploration, which not only wastes resources but also reduces the adaptability and flexibility of the navigation system. This inspires us to consider whether we can estimate explored region to avoid repeated exploration and thereby improve efficiency.

Based on our analysis, we propose **RATE-Nav** (Region-Aware Termination-Enhanced Navigation), a novel navigation method that enhances exploration efficiency through region-level search and intelligent termination strategies. As one of its key components, we develop a Geometric Predictive Region Segmentation module that assists in partitioning incomplete environmental maps into relatively independent regions. This segmentation module utilizes geometric features to predict unexplored areas, contributing to the transformation from point-by-point search into a region-level search problem. The region-based approach enables holistic evaluation of environmental segments, leading to better estimation of unknown spaces and improved navigation efficiency. For the critical task of region-level evaluation, we leverage VLMs which excel at macroscopic environmental perception. When VLMs determine that a target object is unlikely to exist within a region, the sys-

tem designates it as an "exploration-free zone," immediately terminating current exploration and preventing future redundant searches in that area. This VLM-based termination strategy effectively reduces unnecessary exploration and substantially improves overall efficiency.

Our contributions are summarized as follows:

- The marginal utility between navigation efficiency and information acquisition was investigated, which reveals the information gained per step decreases as exploration progresses, suggesting that comprehensive exploration is not always necessary.
- A novel zero-shot object navigation method called RATE-Nav inspired by marginal utility was proposed, which uses VLMs macro-perception and **Region Aware** to determine exploration **Termination**, thereby achieving effective navigation **Enhancement**.
- Extensive experiments are conducted to demonstrate that RATE-Nav significantly outperforms existing zero-shot object navigation methods.

2 Related Works

2.1 Zero-shot Object Navigation

Research in Object Navigation (Li et al., 2022) has evolved from early end-to-end (Khandelwal et al., 2022) deep learning approaches (based on RL (Dang et al., 2023b) and IL (Ramrakhya et al., 2023)) to modular approaches using dynamic map representations (Chaplot et al., 2020a; Zhang et al., 2023; Ramakrishnan et al., 2022; Chen et al., 2023) to improve computational efficiency and adaptability. With the advancement of large models and increasing demands for generalization in object navigation, researchers have proposed various unsupervised (Majumdar et al., 2022) and zero-shot (Gadre et al., 2023) methods. For instance, zero-shot approaches (Gadre et al., 2023; Zhang et al., 2024) enable agents to navigate to unseen target categories by leveraging general knowledge from training, while ESC (Zhou et al., 2023) and OpenFMNav (Kuang et al., 2024) emphasize the importance of commonsense reasoning and VLM perception. VorNav (Wu et al., 2024) demonstrated the potential of novel map representations.

2.2 Large Pre-trained Models for Zero-shot Object Navigation

Large pre-trained models (Achiam et al., 2023; Bai et al., 2023; Naveed et al., 2023) have revolutionized object navigation by enabling zero-shot decision-making in unknown environments, with works leveraging VLMs for scene understanding (Zhou et al., 2023; Kuang et al., 2024), LLMs for object-room relationship analysis (Yu et al., 2023; Cai et al., 2024), and advanced prompting techniques for navigation planning (Long et al., 2024; Yin et al., 2024). Building upon these advances, we propose to address an even more challenging scenario of zero-shot object navigation without any navigation training data.

3 Motivation

3.1 Marginal Utility in Object Navigation

The concept of marginal utility in economics is used to explain the diminishing returns as consumption increases. We observe a similar relationship between exploration steps and exploration rate in Object Navigation.

To investigate the planning allocation problem in navigation, we conducted a systematic analysis of navigation efficiency and information acquisition using the HM3D dataset. Our study examined the relationship between exploration steps and exploration rate, as illustrated in **Figure 2**. The analysis comprises two key components: a) evaluating the correlation between exploration cost and coverage rate during area exploration, and b) analyzing the distribution of exploration coverage rates in regions where robots successfully locate targets. As shown in **Figure 2a**, our findings reveal a significant diminishing return in new environmental information gained per step as exploration continues. Notably, this exhaustive exploration approach is not always necessary. The results from **Figure 2b** indicate that the detection of targets occurs primarily within two distinct phases: the Efficient Acquisition Phase and the Stable Exploration Phase.

Based on the statistical results, we divide the area exploration process into three characteristic phases. First is the Efficient Acquisition Phase, where robots can quickly gather environmental information with minimal movement due to their field of view advantage. The second is the stable exploration phase, during which robots continuously move to steadily acquire new environmental information. Finally, there is the Edge Comple-

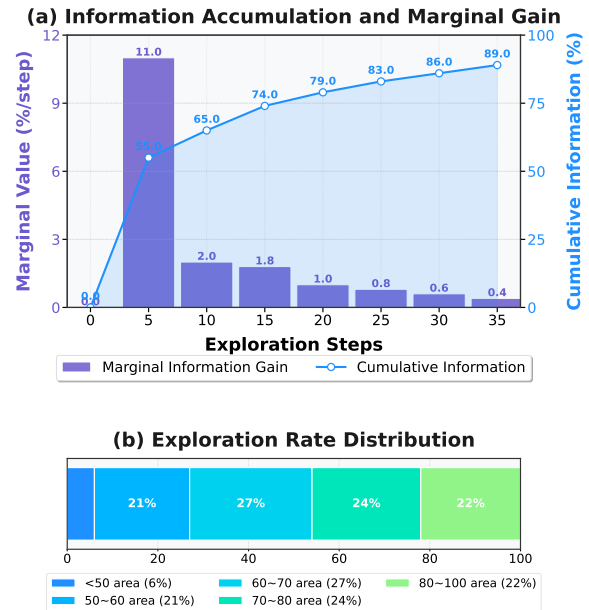


Figure 2: Systematic analysis of navigation efficiency and information acquisition. In (a), the x-axis is exploration steps, and the line graph shows cumulative information. By step 5, exploration reaches 55% of the region. The bar chart displays marginal values (e.g., $55\% \div 5 = 11\%/step$ for steps 0-5). In (b), 6% of samples find the target after exploring 50% of the region, increasing to 78% at 80% exploration.

tion Phase, where robots need to employ complex strategies such as detouring to gather remaining information due to the presence of obstacles.

Through the above research, we introduce the concept of marginal cognition, which identifies two key issues in exploration: 1) Core information collected in the early stages typically holds the highest value, while subsequent marginal information yields diminishing returns; 2) Initial information gathering efforts produce significant gains, but the knowledge acquired per unit time (or cost) tends to decrease as exploration deepens.

3.2 Challenges and Inspiration

While recent studies have attempted to enhance target selection in foundation models through richer environmental information integration, redundant exploration and exploration failures remain inevitable. These challenges underscore the importance of making rapid decisions based on partial exploration data, avoiding the collection of lower-value information at higher costs, and efficiently backtracking from unproductive paths to improve navigation efficiency.

The marginal utility in object navigation inspires

us to explore whether marginal effects can help an agent dynamically adjust its behavior based on changes in the environment. Specifically, in object navigation, the agent could perform region-aware termination based on marginal effects. That is, when the marginal effect of exploring a particular region becomes sufficiently low, the agent may consider halting further exploration in that region and redirecting its efforts to more promising areas.

4 Methodology

4.1 Method Overview

To address the challenges mentioned in **Chapter 3.2**, we propose a macro-perception strategy that leverages VLMs and region exploration rate to evaluate the potential of continuous exploration, enabling efficient exploration termination. The RATE-Nav workflow is illustrated in **Figure 3** and **Algorithm 1**. Our Geometric Predictive Region Segmentation algorithm utilizes geometric features to predict and segment unexplored areas within the map, providing a robust solution to the challenge of ambiguous region references in VLM outputs. This geometric feature-based approach enables more accurate and meaningful region partitioning. Following region segmentation, we introduce a method for estimating explored area that considers both the robot’s visible areas and traversable spaces. When the marginal utility of exploring a specific region drops below a threshold (indicated by the region exploration rate), we filter key frame inputs from RGB observations and feed them into the VLM for macro-environmental perception. The VLM’s logical reasoning capabilities then determine whether to terminate further exploration of that region, preventing redundant exploration in low-value areas.

4.2 Region Semantic Map

The Semantic Map serves as a crucial information tool for evaluating candidate points, playing a dual role in this research: Firstly, it utilizes its rich language information to predict potential locations of target objects based on common sense reasoning. Besides, it builds customized prompts based on identified object features in each Region, enhancing VLMs accuracy in assessing area candidacy.

To construct a high-quality Semantic Map, we employ ConceptGraphs (Gu et al., 2024) for environment modeling. The specific process includes extracting semantic features from RGB-D sequences, accurately projecting them onto 3D

point clouds, and optimizing through multi-view fusion, ultimately generating a complete dataset containing 3D object information and their visual and language descriptors. By combining this information with spatial position data, we can obtain comprehensive semantic information about objects contained in each area, providing reliable prior knowledge support for subsequent target object localization.

4.3 Geometric Predictive Region Segmentation

To characterize explored areas and define region boundaries, map segmentation is essential. We segment the map based on tall obstacles (primarily walls) that obstruct robot vision into distinct regions. While these regions often correspond to different rooms from an indoor navigation perspective, larger spaces like living rooms may be divided into multiple regions to facilitate description and computation. Given RGB-D images I_0, \dots, I_t and robot poses p_0, \dots, p_t , we use 3D point cloud modeling with height threshold h to identify vision-obstructing obstacles, collectively termed as the wall map W . Based on the wall map, we propose a watershed algorithm-based region segmentation method, as shown in **Algorithm 1 Phase 1**. The method consists of these steps:

- 1) Wall preprocessing: Apply distance transform D_w to the wall map and mark areas within a threshold δ (1.5 units) of walls as wall regions.

- 2) Distance map generation: Perform Euclidean distance transform D_e on the processed binary map to create a distance map.

- 3) Region center detection: Identify potential region centers C , which contains center points from c_1 to c_n , using local maxima detection on the distance map. A point c_i (with coordinates (x, y)) is considered a center if its value $D_e(x, y)$ on the distance map exceeds a threshold τ and is maximal within its neighborhood $N(x, y)$. In other words, $D_e(x, y)$ must be equal to the maximum D_e value among all points (x', y') within its neighborhood $N(x, y)$.

- 4) Region segmentation: Apply watershed algorithm to expand detected region centers, using each center as a seed point s_i and simulating a "flooding" process for automatic segmentation. The region label $R(x, y)$ is assigned as:

$$R(x, y) = \arg \min_i \{P(x, y, s_i)\} \quad (1)$$

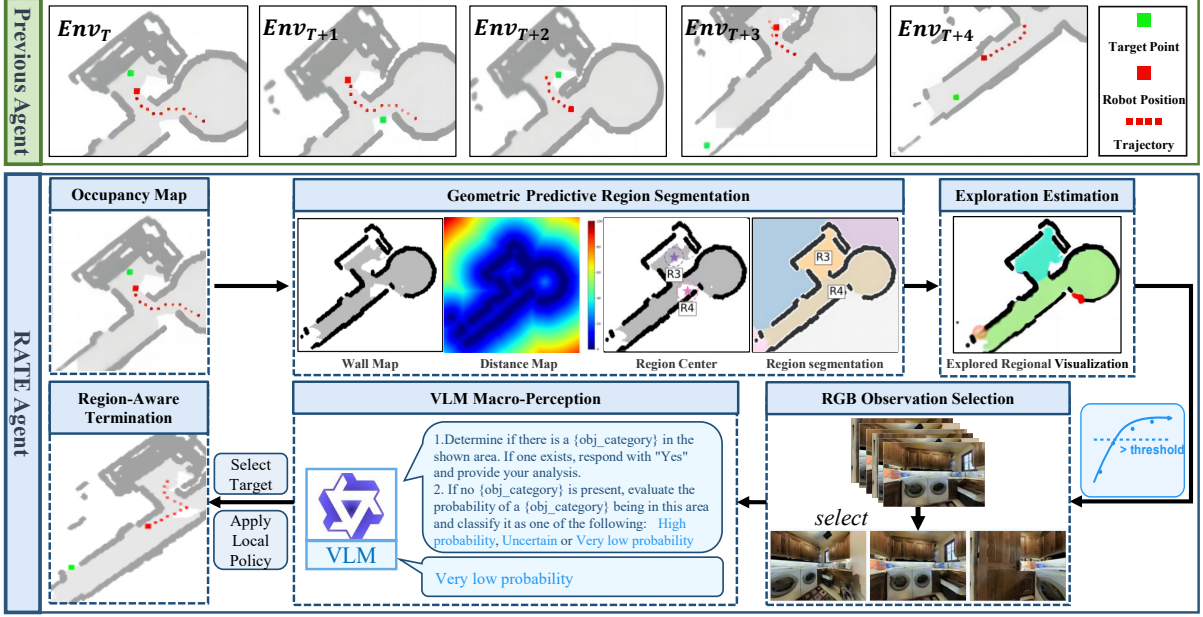


Figure 3: The workflow of RATE-Nav. The previous agent’s requirement for complete spatial exploration often leads to redundant exploration efforts, as evidenced by the repeated exploration of identical regions at Env_T and Env_{T+2} . In contrast, the RATE agent effectively minimizes redundant exploration in low-value areas.

where $P(x, y, s_i)$ represents the topographical distance from point (x, y) to seed point s_i .

5) Post-processing optimization: Refine initial segmentation by merging small regions below a threshold α with adjacent larger regions:

$$R'(x, y) = \begin{cases} \arg \max_{k \in N_R(i)} |R_k| & \text{if } |R_i| < \alpha \\ R(x, y) & \text{otherwise} \end{cases} \quad (2)$$

where $|R_i|$ represents the size of region i and $N_R(i)$ represents the set of regions adjacent to region i .

This method produces semantically meaningful region segmentation with clear boundaries and rich geometric feature descriptions. The approach considers both physical connectivity and practical utility through post-processing, establishing a solid spatial representation foundation for subsequent exploration and navigation tasks.

4.4 Region-Based Exploration Estimation

After obtaining the region segmentation map, we need to estimate the robot’s visible areas to determine the coordinates of explored regions, as presented in **Algorithm 1 Phase 2**. Given the robot’s positions loc_0, \dots, loc_t and orientation information d_0, \dots, d_t , with a maximum visible distance d_{max} defined. Based on the wall map, we calculate the visible area using the following formula:

$$V_t = \{p \mid \|p - loc_t\| \leq d_{max} \& \text{LoS}(loc_t, p) = \text{True}\} \quad (3)$$

Algorithm 1 Region Exploration Algorithm

Input: RGB-D images $\mathcal{I} = \{I_0 \dots I_t\}$

Robot poses $\mathcal{P} = \{p_0, \dots, p_t\}$

Target category c

- 1: **Phase 1: Region Map Construction**
 - 2: $S \leftarrow \text{BuildSemanticMap}(\mathcal{I})$
 - 3: $R \leftarrow \text{SegmentRegions}(S)$
 - 4: **Phase 2: Exploration Estimation**
 - 5: $V \leftarrow \text{CalculateVisibleArea}(\mathcal{P})$
 - 6: $E \leftarrow \text{EstimateRegionRate}(V, R)$
 - 7: **Phase 3: VLM Assessment**
 - 8: $F \leftarrow \text{SelectKeyFrames}(\mathcal{I}, R, E)$
 - 9: $P \leftarrow \text{VLMAssessment}(F, c)$
 - 10: **Phase 4: Decision Making**
 - 11: **if** $\text{IsLowRelevance}(P)$ **then**
 - 12: $R.\text{priority} \leftarrow \text{low}$
 - 13: **else**
 - 14: $\text{ContinueSearch}(R, F)$
 - 15: **end if**
-

where $\text{LoS}(loc_t, p)$ is the line-of-sight function that returns True if there is an unobstructed path between points loc_t and p , implemented using ray tracing Bresenham algorithms.

Additionally, combining with existing mapping methods, we can obtain the traversable area M and obstacle point set O . By taking their intersection, we can get the estimated exploration area point set

E , as shown in the formula:

$$E = \bigcup_{t=0}^T (V_t \cup M_t) \quad (4)$$

This equation calculates the total explored area by taking the union of visible areas and traversable areas at each time step ($(V_t \cup M_t)$), then taking the union of these combined areas over the entire time period T .

After obtaining the exploration area point set, we can calculate the exploration rate r for each region based on the above region segmentation:

$$r = \frac{|E \cap R_i|}{|R_i|} \quad (5)$$

where R_i represents the point set of the i -th region. The exploration rate r indicates the proportion of the region that has been explored, calculated as the ratio of explored points to total points in the region.

4.5 VLMs Macro-Perception of Termination Enhancement

By setting an exploration threshold, when the area’s exploration rate exceeds this threshold, the system triggers VLM for macro-environmental perception. During exploration, the system retains K key frames and records the visible range corresponding to each frame. When initiating VLM environmental macro-perception, the system intelligently filters these K frames based on two core criteria: 1) Field-of-view priority: Select images whose field of view primarily covers the current target area 2) Exploration contribution: Selected images must significantly contribute to the overall exploration rate of the area

The filtered image set is then input to VLM for analysis. Through carefully designed prompts, VLM is guided to make a three-level probability assessment of target object presence in the current area: High probability, Uncertain, and Very low probability. When VLM outputs Very low probability, the system assigns a very low exploration priority score to that area, avoiding redundant exploration in low-value areas and thereby improving overall exploration efficiency, which corresponds to **Phase 3 and 4 in Algorithm 1**.

After obtaining region assessments, we score candidate points by combining the region semantic map with Frontier-based Exploration (FBE), and generate specific actions through local policies such as the Fast Marching Method (FMM)

(Sethian, 1999) once target points are determined. Furthermore, leveraging the powerful capabilities of VLMs, we perform re-perception of discovered targets to enhance detection accuracy.

5 Experiment

5.1 Experimental Setup and Implementation Details

Dataset: We evaluate the effectiveness and navigation efficiency of our proposed method on two widely-used ObjectNav datasets: HM3D (Ramakrishnan et al.) and MP3D (Chang et al., 2017) in the Habitat simulator. The HM3D validation dataset comprises 20 high-fidelity reconstructions of entire buildings and 2K validation episodes for object navigation tasks across six goal object categories. The MP3D validation dataset contains 11 indoor scenes, 21 object goal categories, and 2,195 object-goal navigation episodes.

Evaluation Metrics: We use three metrics to evaluate algorithm performance: 1) Success Rate (SR): percentage of successful episodes; 2) Success weighted by Path Length (SPL): combines success rate and path efficiency; 3) Soft SPL (SSPL): enhanced version of SPL providing finer evaluation by considering final agent-target distance. Higher values indicate better performance for all metrics.

Implementation Details: Experimental setup: max 500 steps per episode, agent camera at 0.88m height with 79° HFOV, discrete actions (0.25m forward step, 30° rotation). Using YOLO-World and GLIP for object detection with 640×640 RGB-D images. Agent maintains 800×800 2D occupancy map (0.05m/cell). Qwen-vl-max (Bai et al., 2023) for complex perception, quantized Llama-Vision (Touvron et al., 2023) 11B for simple reasoning.

5.2 Comparison With Prior Work

Baselines: Based on supervision requirements and zero-shot capabilities, we categorize existing approaches into three groups. For non-zero-shot methods, we selected both supervised approaches (e.g., SemEXP (Chaplot et al., 2020b), PONI (Ramakrishnan et al., 2022)) and unsupervised approaches (e.g., ZSON (Majumdar et al., 2022)) for comparison. In the zero-shot category, where most methods are unsupervised, we included several state-of-the-art approaches for comprehensive comparison, including OpenFMNav (Kuang et al., 2024), ImagineNav (Zhao et al., 2024), and SG-Nav (Yin et al., 2024).

Table 1: Comparison with previous work on MP3D and HM3D.

Method	Unsupervised	Zero-shot	MP3D		HM3D	
			SR \uparrow	SPL \uparrow	SR \uparrow	SPL \uparrow
SemEXP (Chaplot et al., 2020b)	No	No	36.0	14.4	-	-
PONI (Ramakrishnan et al., 2022)	No	No	31.8	12.1	-	-
ZSON (Majumdar et al., 2022)	Yes	No	15.3	4.8	25.5	12.6
CoW (Gadre et al., 2023)	Yes	Yes	7.4	3.7	-	-
TriHelper (Zhang et al., 2024)	Yes	Yes	-	-	56.5	25.3
ImagineNav (Zhao et al., 2024)	Yes	Yes	-	-	53.0	23.8
ESC (Zhou et al., 2023)	Yes	Yes	28.7	14.2	39.2	22.3
L3MVN (Yu et al., 2023)	Yes	Yes	34.9	14.5	48.7	23.0
VLFM (Yokoyama et al., 2024)	Yes	Yes	36.2	15.9	52.4	30.3
OpenFMNav (Kuang et al., 2024)	Yes	Yes	37.2	15.7	52.5	24.1
ImagineNav-Oracle (Zhao et al., 2024)	Yes	Yes	-	-	<u>62.1</u>	<u>31.1</u>
SG-Nav (Yin et al., 2024)	Yes	Yes	<u>40.2</u>	<u>16.1</u>	54.2	24.1
RATE-Nav	Yes	Yes	50.3	20.6	67.8	31.3

Our experimental results demonstrate that RATE-Nav significantly outperforms existing object navigation methods across supervised, unsupervised, and zero-shot categories, as shown in Table 1. On the HM3D dataset, RATE-Nav achieves a success rate of 67.8% and an SPL of 31.3%, substantially surpassing other methods. On the more challenging MP3D dataset, our approach shows approximately 10% improvement over previous zero-shot methods. Notably, the significant improvement in SPL validates that our strategy of transforming point-to-point navigation into region-to-region navigation effectively enhances the robot’s navigation efficiency.

We also analyze the per-category success rate of different zero-shot methods in **Figure 4**. RATE-Nav demonstrates superior performance across most goal categories, significantly outperforming other baseline methods. This consistent improvement across diverse object categories validates the generality and robustness of our approach in object search tasks.

5.3 Ablation Study

To comprehensively evaluate the effectiveness of our key modules, we conducted systematic ablation experiments in Table 2 on three core components: Geometric Predictive Region Segmentation (GPRS), Region-Based Exploration Estimation (REE), and VLM Perception (VP). Here,

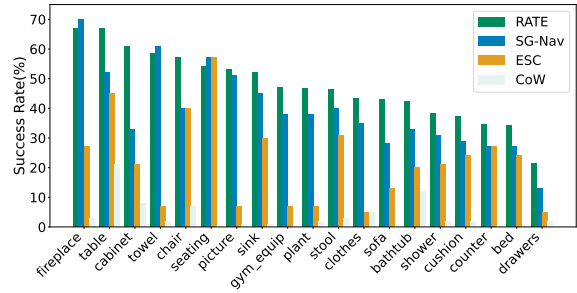


Figure 4: Comparison of our method with previous work across diverse object categories.

GPRS is responsible for semantics-based region division, REE handles explored area estimation, and VP provides visual-language model perception capabilities. When all modules are removed, RATE-Nav degrades to a simple random exploration system. With the introduction of the GPRS module, the system can utilize region semantic maps to guide navigation. Without the REE module, the system relies solely on the occupancy map to calculate exploration coverage. Notably, the VP module not only provides conventional visual perception but also includes a crucial Re-perception mechanism - when the system believes it has found the target object, it performs a secondary confirmation through VLM to enhance target identification accuracy.

As shown in the Table 3, we analyzed how the choice of Vision Language Models (VLM) and ex-

Table 2: Ablation study of different component of RATE-Nav on HM3D dataset

GPRS	REE	VP	HM3D		
			SR ↑	SPL ↑	SSPL ↑
×	×	×	45.3	20.2	25.1
✓	×	×	55.2	24.1	32.5
✓	✓	×	57.7	26.7	33.2
✓	×	✓	64.3	25.5	30.8
✓	✓	✓	67.8	31.3	38.6

Table 3: The influence of vlm and exploration rate

VLM	Rate	HM3D	
		SR ↑	SPL ↑
w/o VLM	0.7	35.1	14.7
LLama-vision	0.7	60.1	26.2
Qwen-vl-max	0.5	59.4	26.1
Qwen-vl-max	0.7	67.8	31.3
Qwen-vl-max	0.9	68.1	25.2
LLama w/o re-perception	0.7	54.3	27.5
Qwen w/o re-perception	0.7	60.3	34.2

ploration rates affect navigation performance in the HM3D dataset. Results show that VLMs are crucial for macro-environment perception - Qwen-vl-max with optimal exploration rate significantly outperforms the baseline without VLM. The less capable Llama-vision shows performance degradation when the goal checking mechanism is removed. Exploration rate choice is equally important, as both low (0.5) and high (0.9) rates limit model performance. Additionally, removing the goal verification mechanism leads to significant performance drops in both models, highlighting the importance of goal detection.

Many existing approaches utilize semantic maps to evaluate the probability of candidate points. Our method innovatively introduces a region-based partitioning mechanism, which enables a more natural integration of object semantics into spatial features. Through comparative experiments on HM3D dataset, as shown in Table 4, we found that semantic maps incorporating region information demonstrate significant advantages over traditional methods: they can effectively differentiate between spatially adjacent areas belonging to different rooms, thereby greatly enhancing the system’s understanding of the environment.

Table 4: The effect of region scene map on HM3D

Method	SR ↑	SPL ↑
w/o scene map	62.7	26.3
scene map w/o region info	65.3	30.1
region scene map	67.8	31.3

5.4 VLM Perception Analysis

As shown in Figure 5, through a case study, we analyze the reasoning principles of Visual Language Models (VLM) in environmental perception. The case includes four perception images taken by the robot at different stages. For the target object "bed," due to its large semantic difference from the living room environment, the model can determine its absence using just the first three images. As for "chair," due to its higher probability of appearing in living rooms, the model needs further exploration to make a judgment. In fact, there is a table and chair set behind the viewpoint of the first image, validating the accuracy of VLM’s reasoning. After obtaining more viewpoints and inputs, the model can finally determine that no chair exists in the currently visible area.

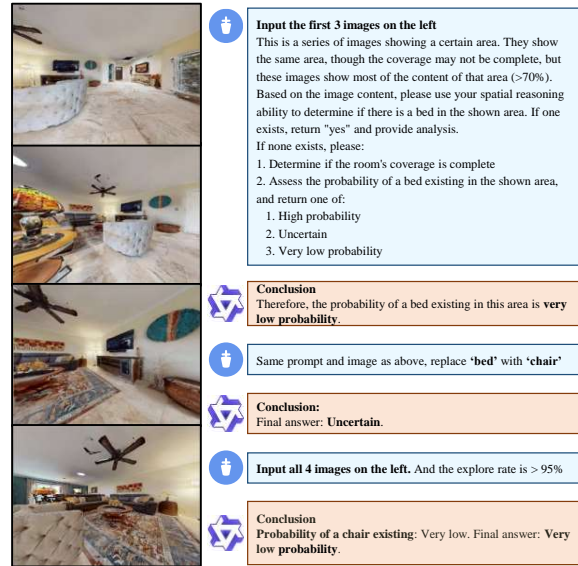


Figure 5: VLM Perception Analysis: Evaluated VLM performance across three input cases to test object recognition and classification capabilities.

6 Conclusion

In this paper, we introduced RATE-Nav, a novel navigation method based on the law of diminishing returns. Our geometric predictive region segmentation method predicts unexplored areas using

geometric features for region partitioning. Converting point-to-point to region-based processing with VLM spatial reasoning improves navigation accuracy and efficiency. While our method effectively handles region-based references, future work should focus on enhancing spatial understanding for terms like "ahead" or "to the right" to enable more natural vision-language navigation.

7 Limitations

Although our method divides the occupancy map into distinct regions, enabling VLM outputs to correspond to specific map regions, this segmentation approach has limitations in incorporating VLM's more general and semantically rich descriptions. In our current work, VLM's spatial descriptions are confined to fixed, specific regions with associated exploration rates. However, in broader contexts, VLM's spatial descriptions tend to be more general, such as "forward" or "turn right." Future work could focus on localizing these linguistically described regions, which would allow us to extend our method to fields like vision-language navigation and image navigation, ultimately working toward a unified navigation framework. Additionally, while our method has only been tested on the Habitat simulator dataset, further exploration is needed to effectively adapt it to real-world applications.

Acknowledgments

This work was sponsored by the Key Research and Development Program of Guangdong Province under grant No. 2021B0101400003, the National Key Research and Development Program of China under grant No.2023YFB4502701.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.
- Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. 2020. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*.
- Wenzhe Cai, Siyuan Huang, Guanran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. 2024. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5228–5234. IEEE.
- Tommaso Campari, Paolo Eccher, Luciano Serafini, and Lamberto Ballan. 2020. Exploiting scene-specific features for object goal navigation. In *European Conference on Computer Vision*, pages 406–421. Springer.
- Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*.
- Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020a. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, volume 33, pages 4247–4258. Curran Associates, Inc.
- Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020b. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258.
- Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. 2023. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. *Proceedings of Robotics: Science and System XIX*, page 075.
- Ronghao Dang, Liuyi Wang, Zongtao He, Shuai Su, Jiagui Tang, Chengju Liu, and Qijun Chen. 2023a. Search for or navigate to? dual adaptive thinking for object navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8250–8259.
- Ronghao Dang, Liuyi Wang, Zongtao He, Shuai Su, Jiagui Tang, Chengju Liu, and Qijun Chen. 2023b. Search for or navigate to? dual adaptive thinking for object navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8250–8259.
- Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 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*, pages 23171–23181.
- Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya

- Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. 2024. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE.
- Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838.
- Yuxuan Kuang, Hai Lin, and Meng Jiang. 2024. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 338–351.
- Baosheng Li, Jishui Han, Yuan Cheng, Chong Tan, Peng Qi, Jianping Zhang, and Xiaolei Li. 2022. Object goal navigation in embodied ai: A survey. In *Proceedings of the 2022 4th International Conference on Video, Signal and Image Processing*, pages 87–92.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge (january 2024). URL <https://llava-vl.github.io/blog/2024-01-30-llava-next>, 2(5):8.
- Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. 2024. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. In *8th Annual Conference on Robot Learning*.
- Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. 2022. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems*, 35:32340–32352.
- Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. 2022. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18890–18900.
- Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.
- Ram Ramrakhya, Dhruv Batra, Erik Wijmans, and Abhishek Das. 2023. Pirlnav: Pretraining with imitation and rl finetuning for objectnav. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17896–17906.
- James A Sethian. 1999. Fast marching methods. *SIAM review*, 41(2):199–235.
- Dhruv Shah, Michael Robert Equi, Błażej Osipiński, Fei Xia, Brian Ichter, and Sergey Levine. 2023. Navigation with large language models: Semantic guesswork as a heuristic for planning. In *Conference on Robot Learning*, pages 2683–2699. PMLR.
- Jingwen Sun, Jing Wu, Ze Ji, and Yu-Kun Lai. 2024. A survey of object goal navigation. *IEEE Transactions on Automation Science and Engineering*.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. 2024. Voronav: voronoi-based zero-shot object navigation with large language model. In *Proceedings of the 41st International Conference on Machine Learning*, pages 53737–53775.
- Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. 2024. SG-nav: Online 3d scene graph prompting for LLM-based zero-shot object navigation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. 2024. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 42–48. IEEE.
- Bangguo Yu, Hamidreza Kasaei, and Ming Cao. 2023. L3mvm: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3554–3560. IEEE.

- Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, and He Wang. 2023. 3d-aware object goal navigation via simultaneous exploration and identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6672–6682.
- Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. 2024. Trihelper: Zero-shot object navigation with dynamic assistance. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2024, Abu Dhabi, United Arab Emirates, October 14-18, 2024*, pages 10035–10042. IEEE.
- Xinxin Zhao, Wenzhe Cai, Likun Tang, and Teng Wang. 2024. Imaginnav: Prompting vision-language models as embodied navigator through scene imagination. *arXiv preprint arXiv:2410.09874*.
- Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. 2023. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *International Conference on Machine Learning*, pages 42829–42842. PMLR.