HuLE-Nav: Human-Like Exploration for Zero-Shot Object Navigation via Vision-Language Models

Anonymous ACL submission

Abstract

Enabling robots to navigate efficiently in unknown environments is a key challenge in embodied intelligence. Human exploration relies on accumulated knowledge, spatio-temporal memory, and scene semantic understanding. Inspired by these principles, we propose HuLE-Nav, a zero-shot object navigation method with two core components: multi-dimensional semantic value maps that emulate human-like memory retention and active exploration mechanisms that mimic human behavior. Specifically, HuLE-Nav utilizes Vision-Language Models (VLMs) and real-time observations to dynamically capture semantic relationships between objects, scene semantics, and spatio-temporal exploration history. This information is then represented and iteratively updated in the multi-dimensional semantic value maps. Using these maps, HuLE-Nav employs active exploration mechanisms that integrate dynamic exploration, replanning, collision avoidance, and target verification, enabling flexible long-term goal selection and real-time adaptation of navigation strategies. Experimental results on the challenging HM3D and Gibson datasets show that HuLE-Nav outperforms the best existing competitors in terms of both success rate and exploration efficiency.

1 Introduction

Understanding how humans navigate efficiently in unfamiliar environments is crucial for developing robots that can effectively mimic human exploration behavior. Typically, efficient human exploration behavior stems from three fundamental cognitive capabilities: accumulated knowledge, spatio-temporal memory, and scene semantic understanding. Therefore, in the context of Zero-Shot Object Navigation (ZSON) (Yu et al., 2023; Wu et al., 2024; Yokoyama et al., 2024), the primary challenge for agents lies in constructing these cognitive capabilities to locate novel target objects in complex environments both accurately and efficiently.

To address these challenges, many ZSON approaches draw inspiration from the concept of "cognitive maps" (Raychaudhuri and Chang, 2025). These historical memories, which help organize spatial information, are crucial for navigation and decision-making (Tolman, 1948; Trullier and Meyer, 2000). Furthermore, recent advances in pre-trained foundation models (Achiam et al., 2023; Kenton and Toutanova, 2019; Touvron et al., 2023) have significantly enhanced decision-making processes by better leveraging these maps, owing to their extensive commonsense knowledge and advanced reasoning abilities (Zhang et al., 2024b). For example, SemExp (Chaplot et al., 2020) pioneered semantic mapping, integrating navigable areas, obstacles, and object categories. Following this, FBE (Gervet et al., 2023) and L3MVN (Yu et al., 2023) improved exploration via language models, encoding object information around map frontiers. ESC (Zhou et al., 2023) refined target localization with room-type semantics, while PixNav (Cai et al., 2023) enabled navigation through pixel-level goals. VLFM (Yokoyama et al., 2024) uses similarity metrics to guide frontier-based decisions.

However, existing methods have several limitations. First, by converting temporary observations into basic textual representations or directly using object information stored in maps for environmental understanding, these approaches struggle with effective scene semantic understanding. Additionally, they neglect the influence of spatio-temporal memory along the trajectory on path planning, causing the agent to engage in inefficient or redundant movements. Furthermore, these methods fail to fully leverage the accumulated knowledge of foundation models to semantically associate information in the environment. Lastly, while these map-based approaches rely on physical topological information to construct maps, foundation models

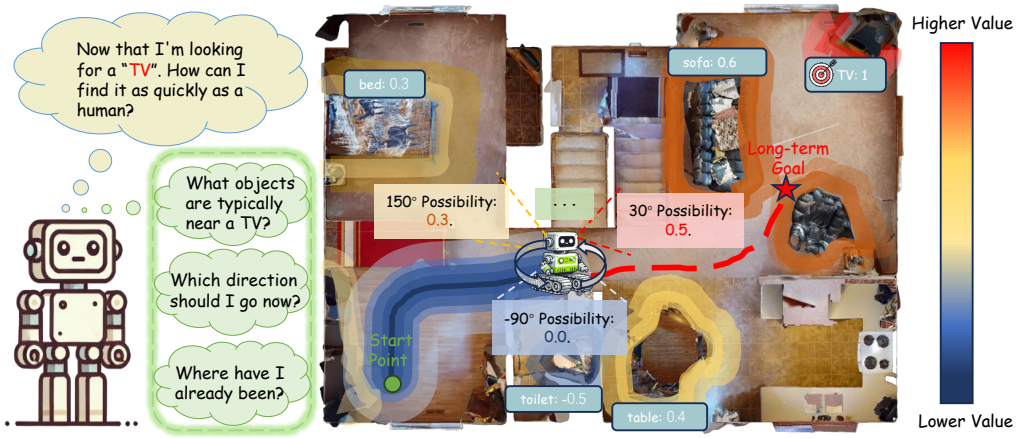

Figure 1: HuLE-Nav records three aspects of the thought process: object-level semantic relevance analysis, scene-level semantic understanding for exploration direction reasoning, and non-redundant path exploration.

are not well-equipped to utilize this map data, hindering the development of a cumulative understanding of the environment. Moreover, these methods lack an active exploration strategy to complement map updates, limiting their ability to efficiently and adaptively gather environmental information. Just imagine humans in a new indoor environment, where they might consider questions like: Which prominent objects are likely to be near the target? Which direction should I take to approach the target now? Where have I already been? Additionally, when humans encounter new environmental information, they may pause to observe and adjust their strategy accordingly.

To address these aspects, we propose Human-Like Exploration for Navigation (HuLE-Nav). We extend the concept of "cognitive maps" within the context of ZSON to not only represent spatial information but also store perception, representation, and action processes (Epstein et al., 2017). Additionally, we strike a balance between navigation and observation, as emphasized in (Kessler et al., 2024). Specifically, HuLE-Nav leverages the accumulated knowledge of foundation models to semantically associate and match object information. We also use their comprehensive semantic understanding of the current scene to determine the optimal direction of movement. To further enhance this, we improve spatio-temporal memory by considering the impact of the agent's trajectory on decision-making. These thought processes are represented as value representations, which are then stored and iteratively updated in real-time, multi-dimensional semantic value maps (see Fig. 1). In parallel, HuLE-Nav incorporates active exploration mechanisms that mimic human behavior—such as dynamic exploration and replanning, collision avoidance, and target validation—ensuring real-time adaptation and accurate exploration. We conduct extensive experiments on the HM3D (Ramakrishnan et al., 2021), Gibson (Xia et al., 2018) datasets on the Habitat platform (Savva et al., 2019). HuLE-Nav achieves state-of-the-art zero-shot performance on all benchmarks, with SPL improving over the previous SOTA by 9.2% on the HM3D dataset and 3.2% on the Gibson dataset. Our contributions can be summarized as follows:

- We store and iteratively update the three-dimensional thought processes in semantic value maps to guide future planning.
- We use active exploration mechanisms to efficiently gather scene information, enhancing exploration accuracy and adaptability.
- We develop a comprehensive map-based navigation method to achieve state-of-the-art performance on the Habitat platform and plan to open-source it.

2 Related Works

Object Navigation. Object Navigation requires an agent, given a textual description of a target object (e.g., "bed"), to locate an instance of that object in a previously unseen environment. Approaches to Object Navigation can be broadly categorized into map-less and map-based methods.Map-less methods use reinforcement learning or imitation learning to directly map visual observations to actions, inferring subsequent actions through the implicit encoding of visual inputs (Chang et al., 2020; Deitke et al., 2022; Maksymets et al., 2021; Ye

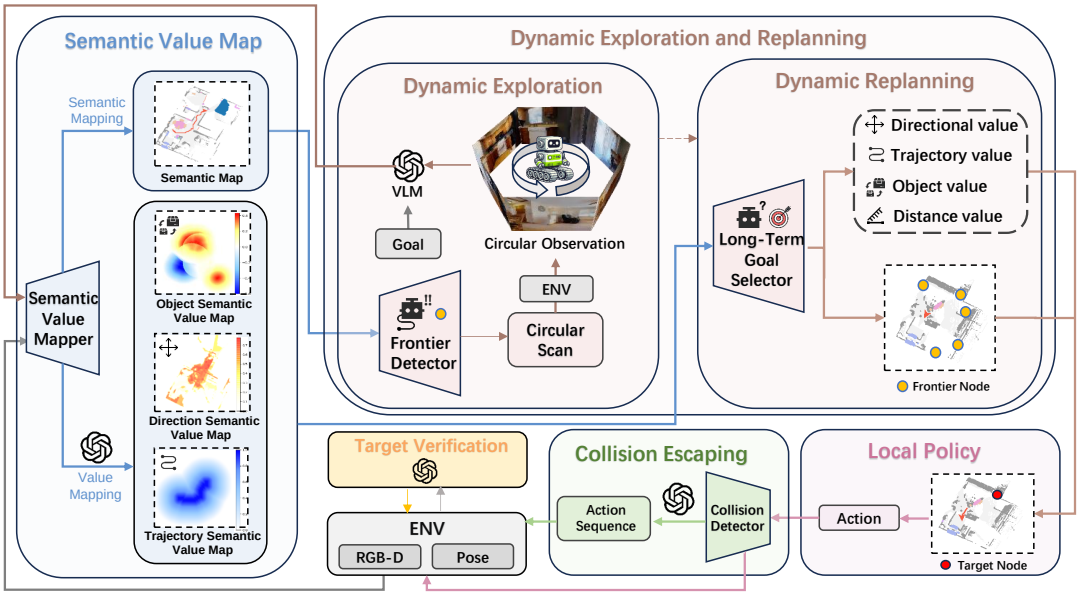

Figure 2: HuLE-Nav consists of two main components: Multi-dimensional Semantic Value Maps for Human-like Exploration Memory and Active Exploration Mechanisms for Human-like Exploration Behaviors.

et al., 2021; Mousavian et al., 2019; Yang et al., 2018). In contrast, map-based methods store historical observations as top-down maps during navigation, enabling the structured storage of topological and semantic information to guide and optimize decision-making (Chaplot et al., 2020; Zhang et al., 2021; Luo et al., 2022; Ramakrishnan et al., 2022; Zhang et al., 2023). However, these task-specific training methods often incur high training costs and are limited to a fixed set of object categories, which restricts their applicability. Moreover, being primarily trained in simulated environments, these methods struggle to generalize effectively to diverse real-world scenarios.

Zero-Shot Object Navigation. Zero-Shot Object Navigation (ZSON) focuses on leveraging prior knowledge to guide an agent's exploration, addressing the challenges of traditional task-specific training-based methods. In recent years, ZSON has made significant progress. Map-less, end-to-end ZSON methods (Majumdar et al., 2022; Zhao et al., 2023; Gadre et al., 2023; Chen et al., 2023a; Gadre et al., 2022) have demonstrated the potential of generalization. However, the implicit scene encoding in these methods often overlooks important contextual details. These methods also require extensive training, suffer from inefficiencies like redundant movements, and lack interpretability in their decision-making processes. In contrast, map-based methods utilize detailed maps to store historical environmental information, which is then used to guide waypoint selection (Kuang et al., 2024; Chen et al., 2023b). Recent advancements in this area have incorporated pre-trained foundation models to enable agents to make frontier-based waypoint decisions based on environmental semantics (Chang et al., 2023; Yu et al., 2023; Shah et al., 2023). Additionally, several methods have improved navigation efficiency and adaptability through the integration of path-planning algorithms (Wu et al., 2024), complementary mapping techniques (Long et al., 2024; Yokoyama et al., 2024), and auxiliary tools (Zhang et al., 2024a). However, these methods have inherent limitations due to their dependence on converting dense visual information into textual representations and discrete landmarks, which leads to sparse and incomplete environmental observations for waypoint selection. Moreover, their decision-making processes are often restricted to fixed intervals or predefined locations, which can result in missed critical context, reduced adaptability, and lower efficiency. We aim to improve our method by maximizing the recording of environmental semantic information and incorporating flexible exploration strategies, enabling better performance in these aspects.

3 Human-Like Exploration for ZSON

In this section, we first provide an overview of HuLE-Nav in Sec .3.1. We then describe the construction of multi-dimensional semantic value maps in Sec .3.2. Finally, we elaborate on the active exploration mechanisms in Sec. 3.3.

3.1 Overview of HuLE-Nav

HuLE-Nav consists of two main components. The first is Human-like Exploration Memory, which uses Multi-dimensional Semantic Value Maps to record environmental observations and multi-dimensional thought processes. The second is Active Exploration Mechanisms, which mimic human-like exploration behaviors, including Dynamic Exploration and Replanning, Collision Escape, and Target Verification (see Fig .2). These mechanisms work together to maximize the acquisition of environmental information, ensuring both efficient and accurate exploration.

The agent takes its position and observations as input and outputs actions. Initially, the agent performs a full panoramic scan and reasons across multiple aspects to construct the Semantic Value Maps, selecting the highest weighted frontier point as the first long-term goal. The agent then uses Local Policy for low-level actions to reach the target. During navigation, when the Dynamic Exploration and Replanning mechanisms detect new frontier points, the agent iteratively updates the maps, selects new goals, and navigates. If the agent gets stuck during exploration, the VLM-based Collision Escape module is activated. Once the target is detected, secondary validation is performed using the VLM-based Target Validation module until the agent successfully finds the target object.

3.2 Multi-dimensional Semantic Value Maps for Human-like Exploration Memory

Semantic Value Maps Overview. To mimic human-like exploration memory, we construct multi-dimensional semantic value maps which is a $K \times M \times M$ matrix initialized to zero. Among them, $M \times M$ represents the map size, and $K = C + 2 + 3$ denotes the number of channels. Specifically, C corresponds to the total number of semantic categories, and the first two channels represent navigable areas and obstacles, respectively. The remaining three channels are uniquely designed to capture (i) *Object Semantic Value*: the semantic relevance between objects, (ii) *Direction Semantic Value*: scene-level semantic information, and (iii) *Trajectory Semantic Value*: the spatio-temporal history of exploration paths.

This $C + 2$ semantic map is constructed following the approach proposed in (Chaplot et al., 2020). Given RGB-D images and the agent's pose at each time step, 3D point clouds are extracted and projected onto the semantic map through height-based filtering. The semantic map is initialized with the agent positioned at the center and evolves dynamically throughout the episode. Point clouds are generated from visual inputs using a geometric method and mapped onto a top-down 2D representation. This map includes obstacle and explored channels derived from depth images, along with semantic channels obtained via semantic segmentation. The semantic masks are aligned with the point clouds, and each channel is accurately projected onto its corresponding position within the semantic map. Collectively, these $C + 2$ layers form a comprehensive historical record of the environment, encompassing navigable areas, obstacles, and semantic categories.

The frontier map is obtained from the explored map and the obstacle map, following the methodology in (Ramakrishnan et al., 2022). It represents the boundaries of the explored area. The process begins by extracting the explored edge from the explored map, identifying the maximum contours. The obstacle map is then dilated, and the frontier map is computed as the difference between the explored and obstacle maps. To refine the results, connected neighborhood analysis is applied to group scattered frontier cells into coherent chains, ensuring spatial continuity. The centroid of each connected frontier region, p_f, is then evaluated and serves as a candidate for the long-term goal.

In addition to preserving environmental context, the remaining 3 channels encode the agent's reasoning process, reflecting its assessment of the necessity and potential value of exploring specific regions. Each pixel in the exploration area is assigned a semantic importance value, quantifying its likelihood of proximity to the target object. The weighted aggregation of the three aforementioned values is then used to evaluate candidate waypoints, dynamically selecting the pixel with the highest value as the next exploration target.

Object Semantic Value Map. To guide the agent in searching around objects with stronger semantic relevance to the target, we radiate the semantic connections between the target and other detected objects to their surrounding areas on the Object Semantic Value Map. At task initiation, the VLM assigns semantic relevance values $S_{o_i} \in [-1, 1]$ between each object instance o_i and the target object as shown in Table 3, with larger positive values indicating a higher likelihood of co-occurrence be-

tween the two objects. These relevance values are then radiated from the object clusters in the semantic map to the Object Semantic Value Map based on their corresponding positions and semantic relationships. For each frontier point p_f within the frontier set on the map, the final semantic value $S_o(p_f)$ is determined by the object with the highest relevance at that point, as shown in Eq. 1.

$$S_o(p_f) = \max_{o_i \in \mathcal{O}} \left(S_{o_i} \cdot \left(1 - \frac{d_o(p_f, o_i)}{r}\right) \cdot \mathbb{I}(d_o(p_f, o_i) \leq r) \right), \tag{1}$$

where $\mathcal{O}$ represents the set of all objects in the semantic map, $d_o(p_f, o_i)$ denotes the distance from the frontier point p_f to the closest point of object o_i, and $\mathbb{I}(d_o(p_f, o_i) \leq r)$ is an indicator function ensuring that only objects within a specified threshold distance r contribute to the relevance score.

Direction Semantic Value Map. To break the object semantic information bottleneck, we employ the VLM to extract scene-level semantic cues, integrating comprehensive semantic information to determine the optimal exploration direction. Specifically, we maintain a cumulative record of optimal direction choices on the Direction Semantic Value Map. At task initiation and at each observation point, the agent performs circular scans to capture six equidistant RGB images, $\{I_0, \ldots, I_5\}$, along with corresponding pose information. The VLM evaluates these images for the potential presence S_{d_i} of the target object G (as shown in Eq. 2), and the results are projected onto the corresponding pixels on the map using depth and pose information. When projections from different angles overlap at the same pixel in the direction map, the corresponding values are averaged to update the map as Eq. 3. Additionally, when projections from different time points overlap at the same pixel, the new value is averaged with the old value, ensuring that past observations influence future decisions.

$$\begin{aligned} S_{d_i} &= \text{VLM}(I_i, G), \\ \sum_{i=0}^{5} S_{d_i} &= 1, \quad i = 0, 1, \ldots, 5, \end{aligned} \tag{2}$$

$$S_d(p_f) = \frac{1}{\mathcal{N}_d(p_f)} \sum_{i=0}^{5} \mathbb{I}(p_f \in \text{proj}(I_i)) \cdot S_{d_i}, \tag{3}$$

where $\mathcal{N}_d(p_f)$ is the number of projections contributing to the value at frontier point p_f, $\text{proj}(I_i)$ refers to the projection of image I_i onto the map, determining which pixels in the map correspond to the visual information in the image I_i.

Trajectory Semantic Value Map. To prevent the agent from repeatedly traversing the same paths or getting stuck at the same target point during exploration, we create a Trajectory Semantic Value Map, which assigns lower values S_t around the trajectory T, encouraging the agent to explore new and diverse paths, as shown in Eq. 4 and 5.

$$\begin{aligned} d_t(p_f) &= \min_{t \in T} \|p_f - t\|, \\ \mathcal{N}_t(p_f, r) &= \{t \in T \mid \|p_f - t\| \leq r\} \quad , \end{aligned} \tag{4}$$

$$S_t(p_f) = \left(\frac{d_t(p_f)}{r} - 1\right) \cdot \left(\frac{|\mathcal{N}_t(p_f, r)|}{\lambda + |\mathcal{N}_t(p_f, r)|}\right) \cdot \mathbb{I}(d_t(p_f) \leq r) \quad , \tag{5}$$

where $d_t(p_f)$ is the minimum distance from frontier point p_f to the nearest point on trajectory T, $\mathcal{N}_t(p_f, r)$ is the number of trajectory points within radius r of p_f, λ is a regularization parameter controlling the influence of neighboring points, and $\mathbb{I}(d_t(p_f) \leq r)$ ensures only frontier points within distance r from the trajectory contribute.

3.3 Active Exploration Mechanisms for Human-like Exploration Behaviors

Dynamic Exploration and Replanning. To encourage the agent to actively survey its surroundings and reduce the risk of missing important semantic information, we propose a dynamic exploration and planning mechanism. During navigation, the agent's current location l becomes an observation point p_o if it has a direct line of sight to a frontier point p_f without any intervening obstacles $\mathcal{B}$, as shown in Eq. 6. At p_o, the agent performs a circular scan and updates the semantic value map using the current panoramic observation, along with the newly acquired semantic information. The agent then selects the frontier point p_f with the highest semantic value from the frontier map as its next long-term goal L, as shown in Eq. 7. The semantic value is computed by combining a weighted sum of the three dimensions of the semantic map, as well as the normalized distance between the current position and the candidate target points.

$$p_o = \{l \mid \exists p_f \in P, \text{line}(l, p_f) \cap \mathcal{B} = \emptyset\}, \tag{6}$$

$$L = \arg\max_{p_f \in P} \left(S_d(p_f) + \alpha S_t(p_f) + \beta S_o(p_f) - \gamma d_{\text{norm}}(p_f) \right). \tag{7}$$

Once the long-term goal L is determined, the local policy is employed to navigate the agent toward

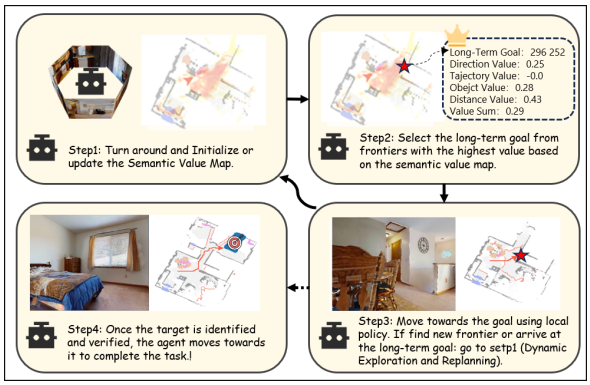

Figure 3: HuLE-Nav method flowchart.

Method	Zero Shot	HM3D		Gibson	
		Success↑	SPL↑	Success↑	SPL↑
Random	✓	0.00	0.00	0.03	0.03
SemExp (Chaplot et al., 2020)	×	37.9	18.8	65.2	33.6
ZSON (Majumdar et al., 2022)	×	25.5	12.6	-	-
Stubborn (Luo et al., 2022)	×	-	-	23.7	9.8
Pixel-Nav (Cai et al., 2023)	×	37.9	20.5	-	-
ESC (Zhou et al., 2023)	✓	39.2	22.3	-	-
COW (Gadre et al., 2023)	✓	32.0	18.1	-	-
FBE (Gervet et al., 2023)	✓	23.7	12.3	41.7	21.4
L3MVN (Yu et al., 2023)	✓	50.4	23.1	76.1	37.7
VoroNav (Wu et al., 2024)	✓	42.0	26.0	-	-
VLFM (Yokoyama et al., 2024)	✓	52.5	30.4	84.0	52.5
SG-Nav (Yin et al., 2024)	✓	54.0	24.9	-	-
HuLE-Nav (Ours)	✓	**55.0** ↑	**33.2**↑	**85.2**↑	**54.2**↑

Table 1: Comparison with SOTA methods on HM3D and Gibson. The best performance for each metric is highlighted in **bold**, and ↑or ↓indicates whether HuLE-Nav outperforms or underperforms compared to other SOTA methods.

the target point. We use the Fast Marching Method (FMM)(Sethian, 1996) for point-to-point navigation, with the agent's current position as the starting point and the long-term goal as the endpoint. At each timestep, FMM computes the optimal path and selects the appropriate action from the action space based on the agent's position. This process ensures real-time, efficient path planning and navigation. By utilizing FMM, we eliminate the need for end-to-end methods like reinforcement learning, allowing our Zero-Shot ObjectNav method to operate without the need for training.

Collision Escape Strategy. Inappropriate selection of long-term goals or poor path planning can cause the agent to become stuck in a situation where it is unable to update its goal or gets trapped in a corner. To mitigate this issue, we propose a VLM-based escape strategy. Specifically, if the long-term goal is not updated, or if the robot's position remains unchanged for a predetermined number of steps, our algorithm triggers the escape strategy. The Vision-Language Model (VLM) uses the current observation I_0 to generate an action plan consisting of 10 actions, $(A_1, A_2, \ldots, A_{10})$, as shown in Eq. 8, which are executed sequentially.

$$(A_1, A_2, \ldots, A_{10}) = \text{VLM}(I_0). \tag{8}$$

Target Verification Strategy. Accurate object recognition is a crucial element for the success of object navigation. To this end, we integrate a Vision-Language Model (VLM) to assist the external object detection module in performing secondary validation of the identified target object. When the agent detects a target object through the detection module, the VLM is then called to verify whether the target object is present in the current observation, considering the full scene. If the VLM confirms the presence of the target object, it is projected onto the semantic map in the target object dimension, serving as the final destination for the local policy navigation. Otherwise, the agent continues its exploration to find the correct target.

4 Experiments

In the section, we focus on demonstrating the superiority of HuLE-Nav compared with counterparts on HM3D dataset (Ramakrishnan et al., 2021) and Gibson (Xia et al., 2018) dataset.

4.1 Experimental Setups

Datasets. The Gibson dataset (Xia et al., 2018) comprises 25 training scenes and 5 validation scenes. The HM3D dataset (Ramakrishnan et al., 2021) includes 75 training scenes and 20 validation scenes. Following the conventional evaluation setup for zero-shot object navigation tasks(Yokoyama et al., 2024), we utilize the validation split of Gibson (1,000 episodes across 5 scenes) and HM3D (2,000 episodes across 20 scenes), respectively.

Metrics. The success rate and efficiency of navigation are important metrics to measure the navigation performance of agents, so we select Success Rate (SR) and Success weighted by Path Length (SPL) for evaluation. Intuitively, higher SR and SPL values indicate that the agents have superior navigation performance. Our contributions can be summarized as follows:

- **Success Rate (SR)** measures the percentage of successful episodes, defined as the agent reaching the target. It is calculated as SR $= \frac{1}{N}\sum_{i=1}^{N} S_i$, where $S_i = 1$ for success and

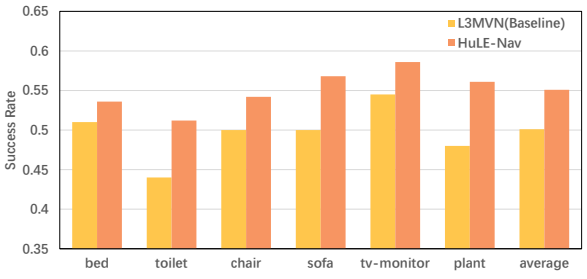

Figure 4: The success rate for each target type in the experiment compared with L3MVN.

$S_i = 0$ for failure, and N is the total number of episodes.

- **Success weighted by Path Length (SPL)** combines success and path efficiency, computed as SPL $= \frac{1}{N}\sum_{i=1}^{N}\frac{S_i \cdot l_i}{\max(p_i, l_i)}$, where S_i indicates success (1) or failure (0), l_i is the shortest path, and p_i is the actual path length. SPL rewards shorter, more efficient paths.

4.2 Experimental Results

To demonstrate the superiority of our method, we compare it with 12 classical algorithms and the experimental results are detailed in Table 1. From Table 1, we can observe that:❶ our method achieves the best performance on both datasets across both metrics; ❷ in terms of SR, HuLE-Nav improves the SR by 1.9% over the current SOTA SG-Nav (Yin et al., 2024) and by 4.8% compared to the second-best VLFM(Yokoyama et al., 2024) on HM3D, and improves the SR by 1.4% over the current SOTA VLFM(Yokoyama et al., 2024) and by 12.0% compared to the second-best L3MVN(Yu et al., 2023) on Gibson; ❸In terms of SPL, HuLE-Nav surpasses all other algorithms by a large margin on both datasets. Its exploration efficiency is both 1.44 times that of the baseline L3MVN(Yu et al., 2023) and improves by 9.2% and 3.2% compared to VLFM(Yokoyama et al., 2024) on the HM3D and Gibson datasets, respectively. The aforementioned results demonstrate the superiority of our method in balancing both success rate and efficiency.

Furthermore, we compare the per-category success rates with the available baseline method L3MVN (Yu et al., 2023) in Figure 4. HuLE-Nav significantly outperforms L3MVN across all target categories, particularly in challenging categories such as toilets and plants, which are difficult to locate or access. This demonstrates that our method can achieve a comprehensive understanding of indoor environments, enabling more effective exploration. In addition, a simple example of the HuLE-Nav navigation process is illustrated in Figure 3 for a more intuitive understanding of our method. More details of the experiment and results can be found in the Appendix.

Experiment Condition	Metrics		Fail Case		
	Success↑	SPL↑	Collision	Exploration	Detection
HuLE-NAV	0.870	0.545	5%	4%	4%
w/o Object Value	0.840 ↓	0.530 ↓	7% ↑	4%	5% ↑
w/o Direction Value	0.760 ↓	0.458 ↓	8% ↑	10% ↑	6% ↑
w/o Trajectory Value	0.740 ↓	0.505 ↓	8% ↑	12% ↑	6% ↑
w/o Distance	0.820 ↓	0.508 ↓	4%	9% ↑	5% ↑
w/o Dynamic Replanning	0.830 ↓	0.508 ↓	4%	9% ↑	4%
w/o Collision Escape	0.810 ↓	0.515 ↓	10% ↑	5% ↑	4%
w/o Target Verification	0.760 ↓	0.503 ↓	2% ↓	8% ↑	14% ↑

Table 2: Ablation Study Results: Evaluating Semantic Value Maps (Upper) and Exploration Mechanisms (Lower). ↑and ↓indicate whether the performance is improved or decreased compared to the original method.

4.3 Ablation & Analysis

After the empirical evaluation of the proposed HuLE-Nav, we are now interested in the following question: Is each component of multi-dimensional semantic value maps and active exploration mechanisms effective? To answer this question, we focus on comparing the approaches that ablate different components on the Gibson validation subset.

Impact of Semantic Value Maps. As shown in Table 2, removing key components from the semantic value maps, such as the Direction Value, Trajectory Value, Object Value, and the distance factor in Eq. 7, results in considerable performance degradation. The absence of the Direction Value impacts the agent's ability to comprehend the overall scene and make informed directional decisions. Most failures arise from the lack of proper direction selection, which leads to exploration issues and a decrease of approximately 10% in both SPL and success rate. Similarly, omitting the Trajectory Value significantly affects the agent's performance, leading to path repetition and, consequently, task failure after hitting the step limit. This causes a 12% decrease in success rate and a 4% reduction in SPL. Furthermore, poor goal point selection due to missing direction and trajectory information can result in unavoidable collisions. Additionally, the Object Value enables the agent to consider object-level relationships in its decision-making, leading to more reliable judgments. Finally, the inclusion of the distance factor contributes to improved ex-

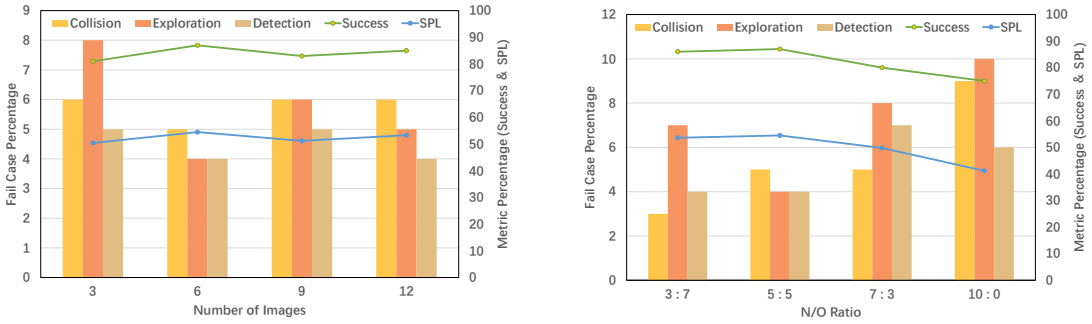

Figure 5: (Left) Ablation study on the number of panoramic images captured during the agent's full rotation. (Right) Ablation study on the ratio of new to old value updates projected onto the same pixel of the Direction Value Map at different timestamps.

ploration performance by allowing the agent to account for the time cost of reaching the target.

Impact of Exploration Mechanisms. The exploration mechanisms—Dynamic Exploration and Replanning, Collision Escape, and Target Verification—are essential for ensuring system resilience and efficiency, as shown in Table 2. Since Dynamic Exploration and Replanning are coupled with the semantic value maps, we adapt the dynamic planning process to a fixed number of steps for replanning, while still allowing the agent to perform dynamic panoramic observations. Even with this adjustment, the results reveal a significant increase in exploration failures, with both success rate and SPL decreasing by approximately 4%. This demonstrates that dynamic exploration and replanning significantly enhance exploration efficiency and navigation performance. Additionally, Collision Escape and Target Verification are designed to address common navigation issues (i.e., collision and detection failures). The results show that these two mechanisms perform well in their respective roles. The absence of the Collision Escape mechanism caused a 10% increase in collision failure rate, highlighting its importance in helping the agent escape from certain predicaments. Moreover, the absence of Target Verification led to a 10% increase in detection error rate. Incorrect object recognition may cause the agent to repeatedly move near the wrong target, resulting in collisions and exploration issues. This underscores the importance of accurate target detection in navigation.

Discussions on Hyperparameters. As shown in Figure 5, we evaluate the impact of the number of panoramic images captured during the agent's full rotation on the direction values (left), as well as the ratio of new to old value updates projected onto the same pixel of the Direction Value Map at different timestamps (right). The results indicate that both the number of captured images and the ratio of new to old memory values significantly affect performance. Capturing too few images may lead to the omission of crucial environmental semantic information, resulting in suboptimal directional decisions. On the other hand, capturing too many images may cause information redundancy, making it difficult for the VLM to process and determine the best course of action. Additionally, during the update process, the values in the direction value map should retain a certain weight for the old values to preserve past decisions and reasoning, providing a reference for future decisions.

5 Conclusion

In this work, we propose HuLE-Nav, a novel zero-shot object navigation method inspired by human-like exploration thinking and behaviors. By constructing multi-dimensional semantic value maps and incorporating human-like exploration mechanisms, HuLE-Nav outperforms its competitors in terms of both navigation success rate and efficiency. The results on challenging datasets demonstrate the superiority of our method in balancing navigation success and efficiency.

Limitations

However, HuLE-Nav has some limitations. These primarily stem from its reliance on external object detection modules. Inaccuracies in object recognition not only affect the final results but may also interfere with intermediate steps during exploration. Additionally, the full potential of Vision-Language Models (VLMs) for navigation tasks has not yet been fully realized. Future work will focus on enhancing VLMs and fully leveraging their capabilities in navigation tasks.

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*.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. 2023. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. *arXiv preprint arXiv:2309.10309*.

Matthew Chang, Theophile Gervet, Mukul Khanna, Sriram Yenamandra, Dhruv Shah, So Yeon Min, Kavit Shah, Chris Paxton, Saurabh Gupta, Dhruv Batra, et al. 2023. Goat: Go to any thing. *arXiv preprint arXiv:2311.06430*.

Matthew Chang, Arjun Gupta, and Saurabh Gupta. 2020. Semantic visual navigation by watching youtube videos. *Advances in Neural Information Processing Systems*, 33:4283–4294.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. 2020. Object goal navigation using goal-oriented semantic exploration. *Advances in Neural Information Processing Systems*, 33:4247–4258.

Hongyi Chen, Ruinian Xu, Shuo Cheng, Patricio A Vela, and Danfei Xu. 2023a. Zero-shot object searching using large-scale object relationship prior. *arXiv preprint arXiv:2303.06228*.

Junting Chen, Guohao Li, Suryansh Kumar, Bernard Ghanem, and Fisher Yu. 2023b. How to not train your dragon: Training-free embodied object goal navigation with semantic frontiers. *arXiv preprint arXiv:2305.16925*.

Matt Deitke, Eli VanderBilt, Alvaro Herrasti, Luca Weihs, Kiana Ehsani, Jordi Salvador, Winson Han, Eric Kolve, Aniruddha Kembhavi, and Roozbeh Mottaghi. 2022. procthor: Large-scale embodied ai using procedural generation. *Advances in Neural Information Processing Systems*, 35:5982–5994.

Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. 2017. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513.

Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. 2022. Clip on wheels: Zero-shot object navigation as object localization and exploration. *arXiv preprint arXiv:2203.10421*, 3(4):7.

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.

Theophile Gervet, Soumith Chintala, Dhruv Batra, Jitendra Malik, and Devendra Singh Chaplot. 2023. Navigating to objects in the real world. *Science Robotics*, 8(79):eadf6991.

Abhinav Gupta, Adithyavairavan Murali, Dhiraj Prakashchand Gandhi, and Lerrel Pinto. 2018. Robot learning in homes: Improving generalization and reducing dataset bias. *Advances in neural information processing systems*, 31.

Jindong Jiang, Lunan Zheng, Fei Luo, and Zhijun Zhang. 2018. Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation. *arXiv preprint arXiv:1806.01054*.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Fabian Kessler, Julia Frankenstein, and Constantin A Rothkopf. 2024. Human navigation strategies and their errors result from dynamic interactions of spatial uncertainties. *Nature Communications*, 15(1):5677.

Yuxuan Kuang, Hai Lin, and Meng Jiang. 2024. Openfmnav: Towards open-set zero-shot object navigation via vision-language foundation models. *arXiv preprint arXiv:2402.10670*.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. 2024. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*.

Haokuan Luo, Albert Yue, Zhang-Wei Hong, and Pulkit Agrawal. 2022. Stubborn: A strong baseline for indoor object navigation. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3287–3293. IEEE.

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.

Oleksandr Maksymets, Vincent Cartillier, Aaron Gokaslan, Erik Wijmans, Wojciech Galuba, Stefan Lee, and Dhruv Batra. 2021. Thda: Treasure hunt data augmentation for semantic navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15374–15383.

Arsalan Mousavian, Alexander Toshev, Marek Fišer, Jana Košecká, Ayzaan Wahid, and James Davidson. 2019. Visual representations for semantic target driven navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8846–8852. IEEE.

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. 2021. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*.

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.

Sonia Raychaudhuri and Angel X Chang. 2025. Semantic mapping in indoor embodied ai–a comprehensive survey and future directions. *arXiv preprint arXiv:2501.05750*.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. 2019. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347.

James A Sethian. 1996. A fast marching level set method for monotonically advancing fronts. *proceedings of the National Academy of Sciences*, 93(4):1591–1595.

Dhruv Shah, Michael Robert Equi, Błażej Osiński, 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.

Edward C Tolman. 1948. Cognitive maps in rats and men. *Psychological review*, 55(4):189.

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*.

Olivier Trullier and Jean-Arcady Meyer. 2000. Animat navigation using a cognitive graph. *Biological Cybernetics*, 83(3):271–285.

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. *arXiv preprint arXiv:2401.02695*.

Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079.

Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, Aaron Gokaslan, Oleksandr Maksymets, Rishabh Jain, Ram Ramrakhya, Angel X Chang, Alexander Clegg, Manolis Savva, Eric Undersander, Devendra Singh Chaplot, and Dhruv Batra. 2022. Habitat challenge 2022. https://aihabitat.org/challenge/2022/.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, et al. 2023. Habitat-matterport 3d semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4927–4936.

Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. 2018. Visual semantic navigation using scene priors. *arXiv preprint arXiv:1810.06543*.

Joel Ye, Dhruv Batra, Abhishek Das, and Erik Wijmans. 2021. Auxiliary tasks and exploration enable objectgoal navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16117–16126.

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. *arXiv preprint arXiv:2410.08189*.

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. L3mvn: 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*, pages 6672–6682.

Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. 2024a. Trihelper: Zero-shot object navigation with dynamic assistance. *arXiv preprint arXiv:2403.15223*.

Min Zhang, Jianye Hao, Xian Fu, Peilong Han, Hao Zhang, Lei Shi, Hongyao Tang, and Yan Zheng. 2024b. Mfe-etp: A comprehensive evaluation benchmark for multi-modal foundation models on embodied task planning. *arXiv preprint arXiv:2407.05047*.

Sixian Zhang, Xinhang Song, Yubing Bai, Weijie Li, Yakui Chu, and Shuqiang Jiang. 2021. Hierarchical

object-to-zone graph for object navigation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15130–15140.

Qianfan Zhao, Lu Zhang, Bin He, Hong Qiao, and Zhiyong Liu. 2023. Zero-shot object goal visual navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2025–2031. IEEE.

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.

A More Details of Preliminary

Task Definition. Zero-Shot Object Navigation (ZSON) is a challenging semantic navigation task where an agent, given a textual description of a target object (e.g., "bed"), must locate an instance of that object in a previously unseen environment, without relying on task-specific training or prior knowledge.

At the beginning of an episode, the agent is initialized at a random starting position p_0 within an unknown environment E and is provided with the target object category $G \in N$, where N represents a novel set of unseen object types. The agent receives egocentric observations comprising RGB-D images I_t and its real-time pose p_t at each time step t. The cumulative observation history is denoted as $O_t = \{(p_0, I_0), \ldots, (p_t, I_t)\}$. The agent operates in a discrete action space A consisting of six actions: *MOVE FORWARD* (0.25m), *TURN LEFT* (30°), *TURN RIGHT* (30°), *LOOK UP* (30°), *LOOK DOWN* (30°), and *STOP*. A successful episode is achieved if the agent executes the *STOP* action within 0.1m of the target object in 500 steps or fewer. Conversely, the episode fails if the agent either exceeds the maximum step limit, stops at an incorrect location, or fails to avoid obstacles. This task emphasizes the agent's ability to utilize semantic reasoning, integrating visual and spatial observations to navigate complex indoor environments efficiently, while generalizing to novel object categories and scenarios.

Episodic Semantic Map. Semantic Map construct and update a $(K+2) \times M \times M$ map using RGB-D images and poses, where M denotes the dimensions of the map's width and height, and $K+2$ represents the total number of channels in the map. Specifically, K channels represent the semantic channels of the detected objects, 2 channels correspond to an obstacle map and an explored map. Given RGB-D images and the agent's poses at each time step, we can obtain 3D point clouds. The 3D point clouds are projected onto a top-down 2D map by judging the height, resulting in an obstacle map and an explored map, which represent navigable areas and non-navigable obstacle areas, respectively. Simultaneously, the RGB images are used to predict the category masks and filter out specific object categories. These are aligned with the 3D semantic point clouds and ultimately projected onto the corresponding K semantic channels.

B More Details of Method

Semantic Value Map. The Semantic Value Map assigns a value to each pixel in the exploration area, quantifying its semantic importance for locating the target object. This value is a parameterized sum of three dimensions: Direction Semantic Value Map, Trajectory Semantic Value Map, and Object Semantic Value Map. The value map is used to evaluate each frontier, with the highest-valued frontier selected for the next exploration step. The Direction Semantic Value Map is iteratively built using depth and pose information to construct a top-down map, where VLM-provided probabilities are projected onto the corresponding map pixels. When probabilities from different directions are projected onto the same pixel, their average is calculated. The Trajectory Semantic Value Map calculates pixel values based on the agent's trajectory path, while the Object Semantic Value Map computes pixel values based on the most influential value from the object list.

C More Details of Experiment

Experiment Setup. For HM3D(Yadav et al., 2023) and Gibson(Xia et al., 2018) , we use Habitat Simulator(Savva et al., 2019) adhere to the parameters established in the Habitat ObjectNav Challenge (Yadav et al., 2022). For agent, we select a LoCoBot(Gupta et al., 2018) with a base radius of 0.18m, which equipped with an RGB-D camera mounted at a height of 0.88 meters and a pose sensor that provides precise localization. The camera features a 79° Horizontal Field of View (HFoV) and captures frames with dimensions of 480 × 640 pixels. For category prediction across all classes, we employ a finetuned RedNet model(Jiang et al., 2018), following the approach outlined in(Ye et al., 2021). In addition, we employ GPT-4o(Achiam et al., 2023) as the VLM in our method. The parameters for the Trajectory Semantic Value Map and Object Semantic Value Map in Eq. 1, 4, and 5 are set with $r = 30$, and the weights for α, β, and γ in Eq. 7. are set to 0.5, 0.3, and 0.1, respectively. λ is set as 10 in Eq. 5. More details can be found in Sec.4.3

Experiment Baselines. In this work, we compare HuLE-Nav against several baselines:

- **Random Exploration**: A baseline method where the agent selects random target points in

unexplored areas, offering a simple reference for comparison with other methods.

- **FBE (Frontier-Based Exploration)** (Gervet et al., 2023): A traditional exploration approach that builds a map and uses frontier-based exploration to explore unknown environments.

- **SemExp** (Chaplot et al., 2020): A semantic exploration method that integrates reinforcement learning to guide exploration based on semantic maps, relying on pre-trained semantic models for object detection and localization.

- **ZSON** (Majumdar et al., 2022): A zero-shot object navigation method that uses multimodal goal embeddings to navigate to previously unseen objects without specific training for each target category.

- **Stubborn** (Luo et al., 2022): A baseline method that relies on a predefined exploration strategy without active learning, resulting in fixed behavior during the navigation process.

- **Pixel-Nav** (Cai et al., 2023): A foundation model-based approach that selects navigation pixels from panoramic images and trains the agent to navigate towards them using locomotion skills.

- **ESC (Exploration with Soft Commonsense Constraints)** (Zhou et al., 2023): A zero-shot navigation method that integrates object and room type semantics with soft commonsense constraints using large language models to guide the agent's exploration process.

- **COW (Commonsense Object Navigation)** (Gadre et al., 2023): A zero-shot object navigation approach that uses commonsense knowledge to guide exploration, providing improved navigation efficiency without the need for task-specific training.

- **FBE (Frontier-Based Exploration)** (Gervet et al., 2023): As mentioned, FBE employs frontier-based exploration to guide the agent through the environment, aiming to achieve efficient exploration.

- **L3MVN** (Yu et al., 2023): This method uses a large language model (LLM) fine-tuned for frontier-based exploration and reasoning to help the agent make informed decisions in complex environments.

- **VoroNav** (Wu et al., 2024): A zero-shot navigation method that uses Voronoi diagrams combined with reasoning via LLMs for more efficient navigation and path planning.

- **VLFM** (Yokoyama et al., 2024): A novel method that integrates vision-language frontier maps and utilizes LLMs to enhance reasoning about object relationships and navigation goals.

- **SG-Nav** (Yin et al., 2024): A method that combines large language models with 3D scene graph representations to guide zero-shot navigation, leveraging the reasoning abilities of the LLM to enhance decision-making.

Experiment Examples. We analyzed the success rate for each target type in the experiment and compared it with the baseline L3MVN. As shown in Fig. 8, HuLE-Nav achieves a higher exploration success rate than the baseline across all object categories. We present several examples during the experiment. Tab. 3 shows the pairwise relationship degrees between objects provided by GPT-4o, where one row will be used in the object semantic map during the experiment. Fig. 6, 7, 8, illustrate sample prompts given to GPT-4 in our method. Fig. 9, 10, 11, and 12 illustrate some typical examples and processes encountered in various parts of the experiment.

Object	Chair	Sofa	Plant	Bed	Toilet	TV Monitor	Bathtub	Shower	Fireplace	Appliances	Towel	Sink	Chest of Drawers	Table	Stairs
Chair	1	0.75	0.2	0.4	-0.3	0.5	-0.6	-0.5	0.3	0.1	0.1	-0.2	0.5	0.7	0.1
Sofa	0.75	1	0.3	0.5	-0.4	0.6	-0.5	-0.5	0.4	0.2	0.2	-0.2	0.6	0.8	0.2
Plant	0.2	0.3	1	0.1	-0.2	0.2	-0.2	-0.3	0.2	0.1	0.3	0.2	0.2	0.3	0.1
Bed	0.4	0.5	0.1	1	-0.6	0.3	-0.3	-0.4	0.2	0.1	0.1	-0.5	0.6	0.5	0.1
Toilet	-0.3	-0.4	-0.2	-0.6	1	-0.5	0.6	0.7	-0.2	-0.3	0.5	0.6	-0.5	-0.4	0.2
TV Monitor	0.5	0.6	0.2	0.3	-0.5	1	-0.5	-0.4	0.3	0.2	0.1	-0.2	0.5	0.6	0.1
Bathtub	-0.6	-0.5	-0.2	-0.3	0.6	-0.5	1	0.8	-0.2	-0.3	0.4	0.5	-0.5	-0.4	0.1
Shower	-0.5	-0.5	-0.3	-0.4	0.7	-0.4	0.8	1	-0.3	-0.4	0.5	0.6	-0.6	-0.5	0.1
Fireplace	0.3	0.4	0.2	0.2	-0.2	0.3	-0.2	-0.3	1	0.2	0.2	-0.1	0.3	0.4	0.2
Appliances	0.1	0.2	0.1	0.1	-0.3	0.2	-0.3	-0.4	0.2	1	0.2	0.3	0.2	0.2	0.3
Towel	0.1	0.2	0.3	0.1	0.5	0.1	0.4	0.5	0.2	0.2	1	0.5	0.1	0.2	0.1
Sink	-0.2	-0.2	0.2	-0.5	0.6	-0.2	0.5	0.6	-0.1	0.3	0.5	1	-0.4	-0.3	0.2
Chest of Drawers	0.5	0.6	0.2	0.6	-0.5	0.5	-0.5	-0.6	0.3	0.2	0.1	-0.4	1	0.6	0.1
Table	0.7	0.8	0.3	0.5	-0.4	0.6	-0.4	-0.5	0.4	0.2	0.2	-0.3	0.6	1	0.2
Stairs	0.1	0.2	0.1	0.1	0.2	0.1	0.1	0.1	0.2	0.3	0.1	0.2	0.1	0.2	1

Table 3: The Object Correlation Table provided by GPT-4o indicates that higher values represent stronger relationships between objects, meaning they are more likely to appear together, while lower values suggest they are less likely to co-occur.

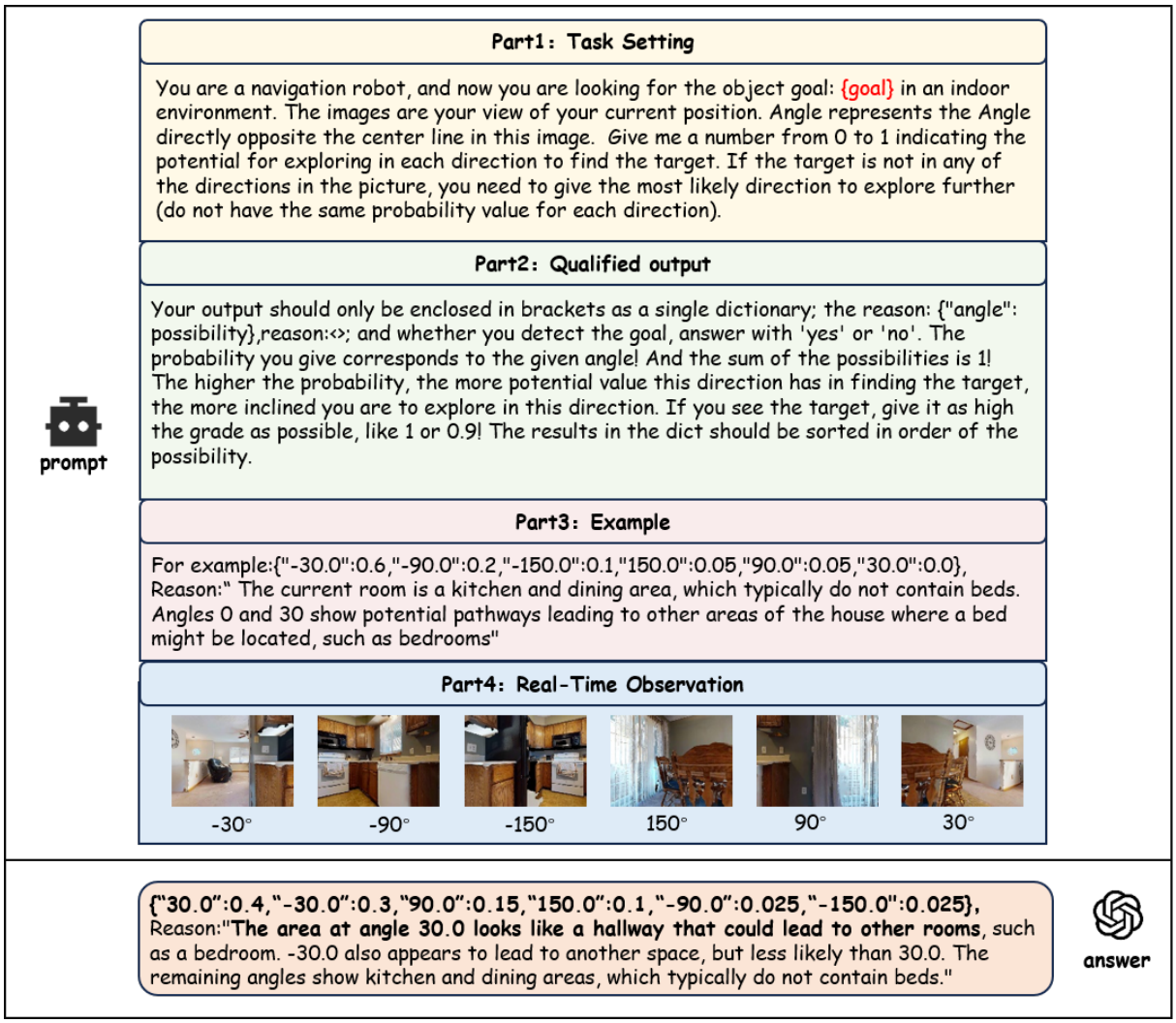

Figure 6: The complete prompt given to GPT-4 for generating the direction value map after obtaining panoramic images in HuLE-Nav, along with a sample response.

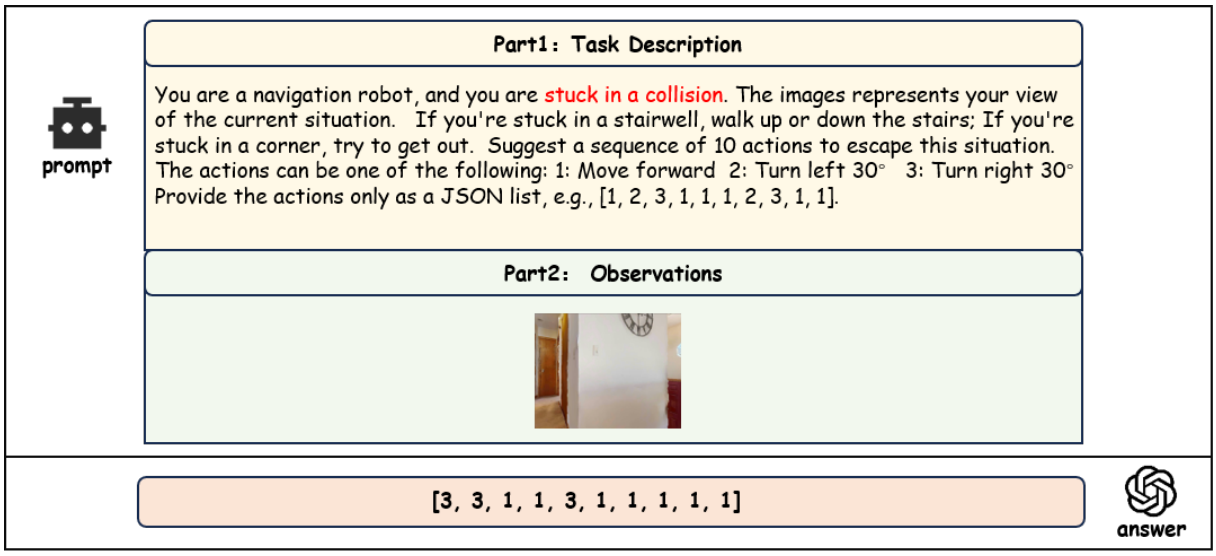

Figure 7: The complete prompt given to GPT-4 for collision escape when HuLE-Nav encounters a predicament, along with a sample response.

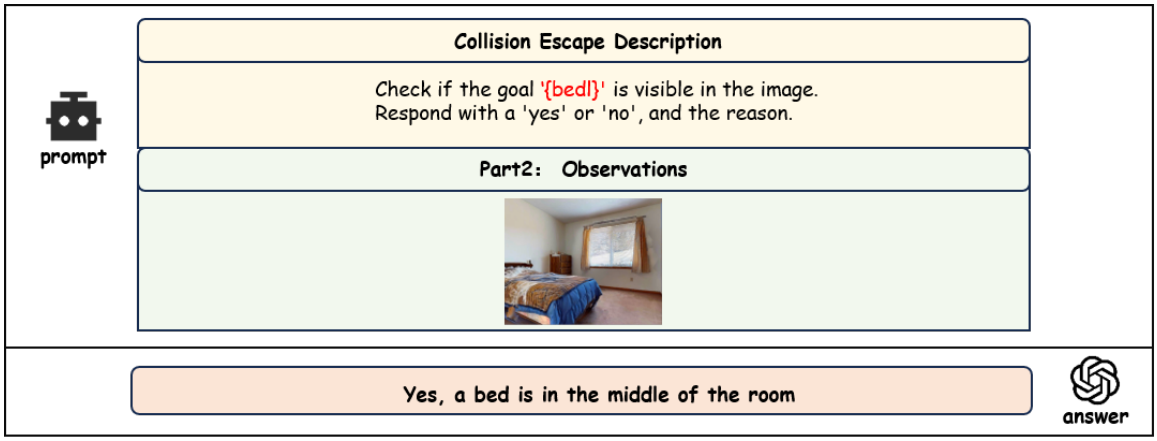

Figure 8: The complete prompt given to GPT-4 for target verification when HuLE-Nav detects a target, along with a sample response.

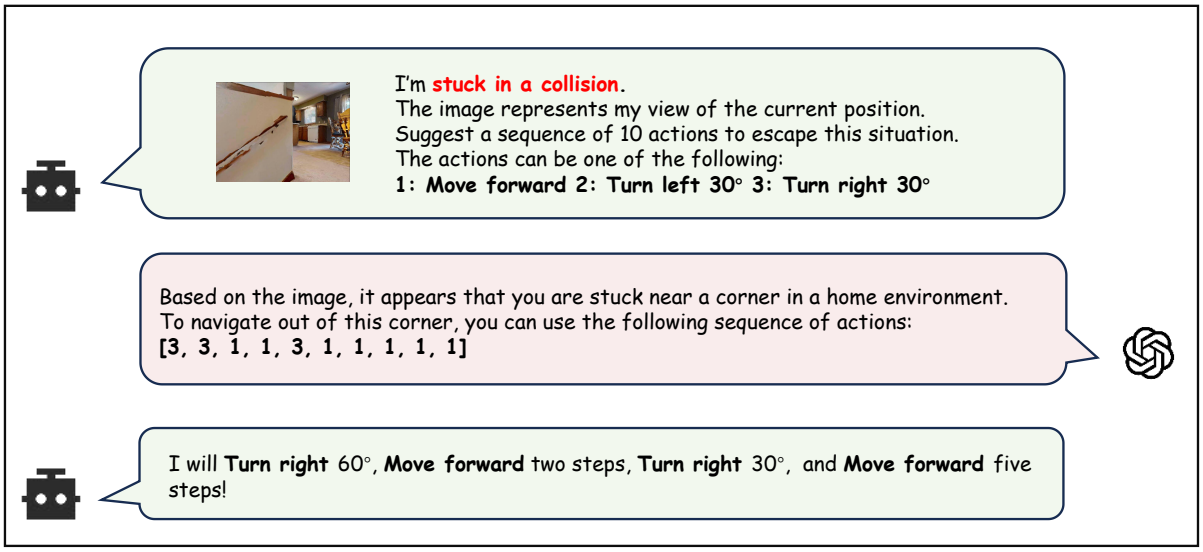

Figure 9: Example of HuLE-Nav collision escape: When the robot encounters a deadlock, GPT-4o generates an action sequence based on the robot's current observations. The robot then executes this sequence to successfully escape the trapped situation.

Figure 10: Example of HuLE-Nav navigation process: The robot's main steps in a task to find a "bed," from initialization to task completion.

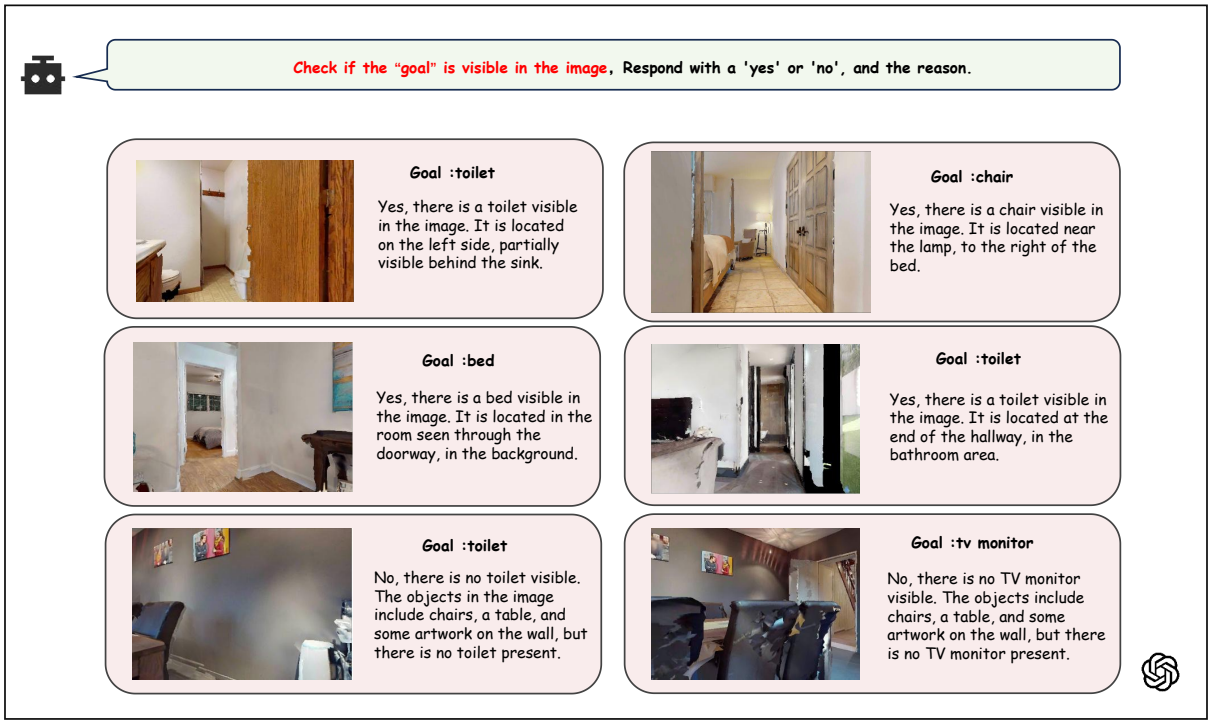

Figure 11: Example of HuLE-Nav target verification: After the target detector identifies the object, GPT-4o is used to verify and confirm the detection.

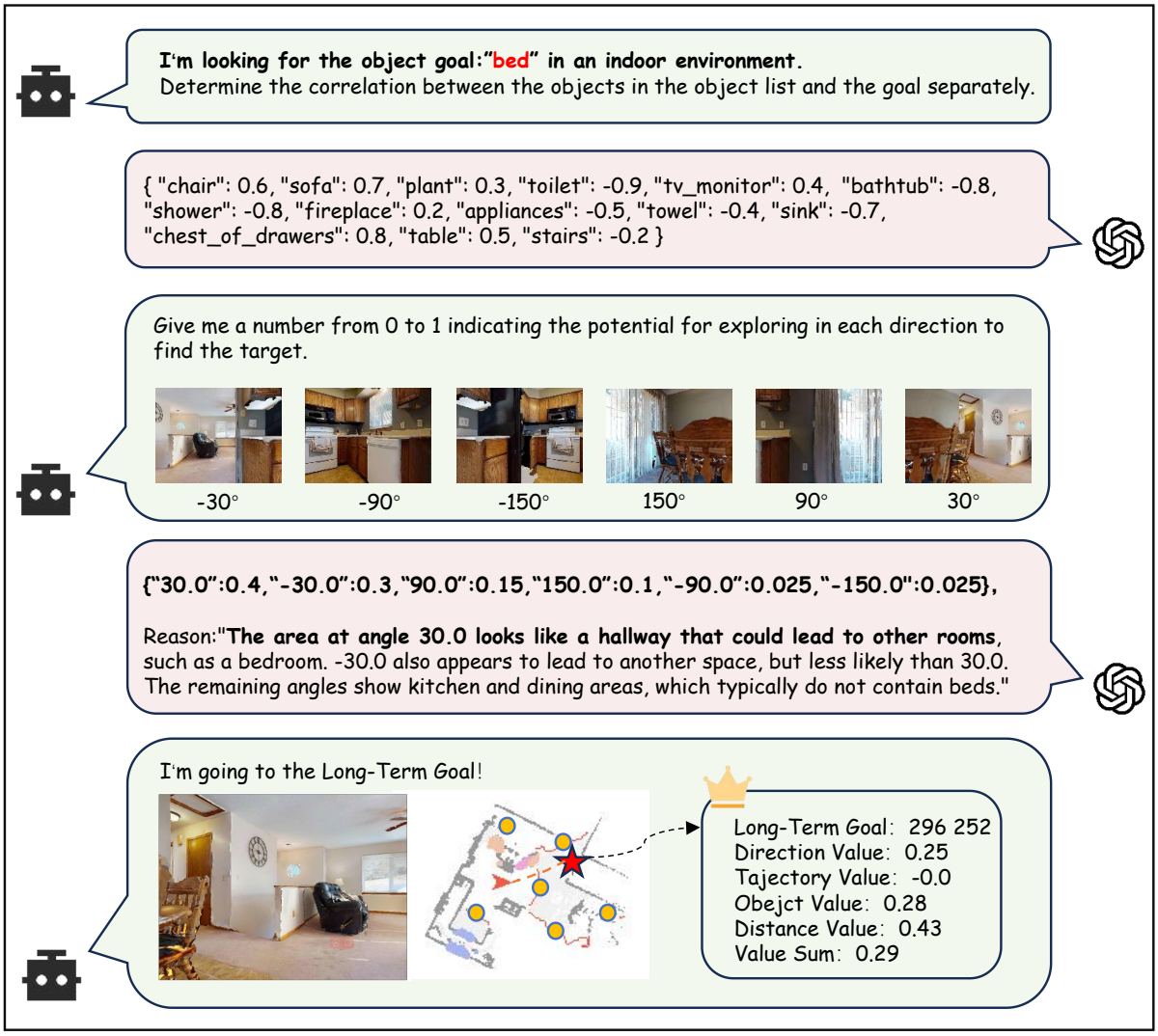

Figure 12: Example of HuLE-Nav circular scan initialization for the semantic value map decision-making based on the updated semantic value map.