

SSR-ZSON: Zero-Shot Object Navigation via Spatial-Semantic Relations within a Hierarchical Exploration Framework

Xiangyi Meng, Delun Li, Zihao Mao, Yi Yang and Wenjie Song*

Abstract—Zero-shot object navigation in unknown environments presents significant challenges, mainly due to two key limitations: insufficient semantic guidance leads to inefficient exploration, while limited spatial memory resulting from environmental structure causes entrapment in local regions. To address these issues, we propose SSR-ZSON, a spatial-semantic relative zero-shot object navigation method based on the TARE [1] hierarchical exploration framework, integrating a viewpoint generation strategy balancing spatial coverage and semantic density with an LLM-based global guidance mechanism. The performance improvement of the proposed method is due to two key innovations. First, the viewpoint generation strategy prioritizes areas of high semantic density within traversable sub-regions to maximize spatial coverage and minimize invalid exploration. Second, coupled with an LLM-based global guidance mechanism, it assesses semantic associations to direct navigation toward high-value spaces, preventing local entrapment and ensuring efficient exploration. Deployed on hybrid Habitat-Gazebo simulations and physical platforms, SSR-ZSON achieves real-time operation and superior performance. On Matterport3D [2] and Habitat-Matterport3D [3] datasets, it improves the Success Rate(SR) by 18.5% and 11.2%, and the Success weighted by Path Length(SPL) by 0.181 and 0.140, respectively, over state-of-the-art methods.

I. INTRODUCTION

Robotic object navigation algorithms, as one of the core capabilities in industrial, search-and-rescue, and service scenarios, require robots to efficiently search for targets in unknown environments lacking prior maps. Although object navigation technologies have advanced rapidly, existing research still struggles to balance semantic reasoning and spatial exploration efficiency in open-world settings. Traditional geometry-driven exploration strategies often overlook semantic information, failing to achieve true goal-oriented navigation. Classical object navigation algorithms, meanwhile, rely excessively on environmental semantics while neglecting the profound influence of scene structures on exploration. This leads to insufficient adaptability in generalized scenarios and low task success rate.

Recent research in zero-shot object navigation has integrated LLMs into navigation systems to enhance semantic understanding. These approaches typically involve parsing textual instructions to generate relevant target information or modeling object associations via knowledge graphs to

guide robot navigation. However, these methods face two primary challenges: First, they often neglect the impact of scene structure on exploration. While LLMs offer strong open-vocabulary reasoning capabilities, these are not effectively translated into actionable exploration decisions in unknown spaces, leading to a disconnect between semantic associations and path planning. Second, the heuristic value of semantic distributions in guiding target search remains underutilized. Current strategies primarily rely on object or scene correlations and traditional exploration planning methods, such as boundary or sampling point generation, without adequately integrating semantic insights with exploration strategies, resulting in suboptimal navigation efficiency.

To address these challenges, this paper introduces the SSR-ZSON framework, as shown in fig. 1. First, the natural language processing capabilities of LLMs are leveraged to analyze human instructions and identify the target object to be located. Simultaneously, by constructing a dynamic semantic topology map, LLMs are utilized to evaluate in real-time the association weights between visible objects and the target objects, such as reasoning the “co-occurrence probability of a monitor and a keyboard” or the “likelihood of a bed appearing in a bedroom” and so on. Second, the environment is discretized into uniform grid sub-regions, and a regional priority scoring function is designed. This function calculates the unexplored surface coverage rate and semantic association value of each sub-region to determine an optimal exploration sequence, enabling bidirectional exploration of high-semantic-value and geometrically unknown areas. In addition, multiple viewpoints are evenly distributed within each selected region, and a two-stage viewpoint screening mechanism is implemented. Initially, a candidate viewpoint set is generated based on semantic density and uncovered surface coverage; subsequently, connectable viewpoints are sampled from this set and linked to form an exploration path for the region.

In summary, our key contributions are as follows.

1) **Viewpoint Generation Strategy:** We design a viewpoint generation strategy that optimizes both spatial coverage and semantic density, guiding the robot toward areas of semantic relevance but underexplored to significantly reduce ineffective exploration.

2) **Dynamic Semantic Weighting-Based Region Exploration:** We introduce a hierarchical exploration approach that partitions the environment into several regions, employs LLMs to dynamically assign semantic weights and structural priors, and incorporates region-wise state tracking to enhance long-term exploration memory.

This work was partly supported by Program for National Natural Science Foundation of China (Grant No. 62373052), Beijing Natural Science Foundation (Grant No. 4252051), and in part by the National Key Laboratory of Science and Technology on Space Born Intelligent Information Processing TJ-01-22-09.

The authors are with the School of Automation, Beijing Institute of Technology, Beijing 100081, China, (Corresponding author: Wenjie Song, email: songwj@bit.edu.cn)

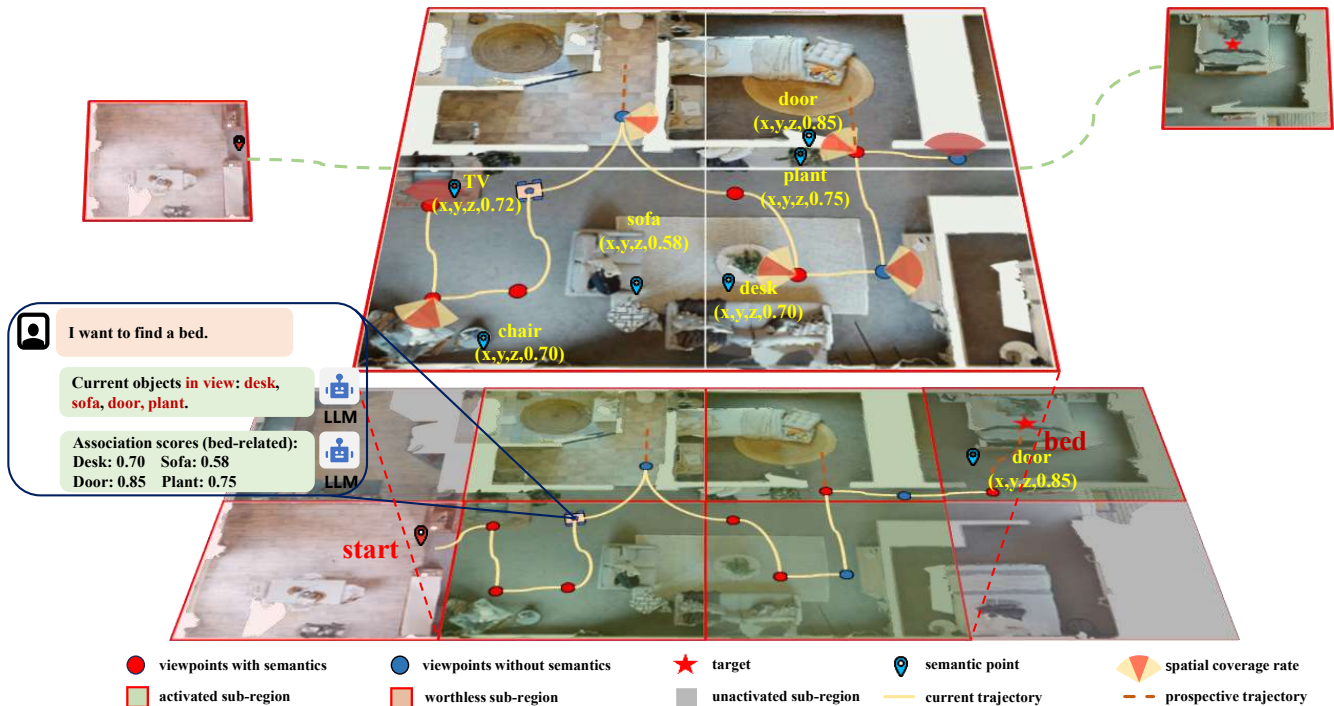


Fig. 1. **Spatial-Semantic Collaborative Exploration Framework.** This framework integrates semantic reasoning based on LLMs with spatial exploration. A dynamic semantic topological map assesses the relevance of objects to the goal in real-time, while the environment is discretized into grid regions. Within each selected region, candidate viewpoints are generated based on local semantic density and coverage. These regions are then prioritized for exploration according to the number of viewpoints and their semantic relevance, with priority given to those exceeding a predefined threshold. The exploration path is formed by connecting traversable candidate viewpoints. This framework drives robotic navigation towards regions with high semantic value and geometrically unknown areas.

3) Viewpoint-Region Joint Object Navigation Framework: We propose a viewpoint-region integrated zero-shot object navigation framework that tightly couples semantic inference with spatial-semantic decision-making, effectively mitigating the spatial-semantic decoupling issue in unknown environments.

II. RELATED WORK

A. Spatial Perception-based Autonomous Exploration

Spatial perception-based autonomous exploration relies on the perception of environmental geometric features. Classical algorithms can be categorized into three types: frontier-based, sampling-based, and hybrid approaches. Frontier-based exploration, first proposed by B. Yamauchi [4], identifies boundaries between known and unknown regions as navigation targets and iteratively expands these boundaries to complete exploration. However, this distance-prioritized strategy often leads to local optima. Subsequent improvements incorporate dynamic replanning and multi-criteria evaluation functions. In 2017, Umari proposed the Rapidly-exploring Random Tree (RRT) method [5] to detect frontiers [6]. When a tree edge straddles both unknown and known regions, the point possessing a path from the unknown region to the known region constitutes the frontier. In complex large-scale scenarios, numerous subsequent approaches have proposed combining frontier-based and sampling-based

strategies by structuring the overall planning process into global and local hierarchies. For instance, to ensure efficient exploration while preventing entrapment in local minima, [7] adopts the Receding Horizon Next-Best-View planning (RH-NBVP) method [8] as the local planner alongside a frontier exploration-based method as the global planner. The TARE method [1] employs viewpoint sampling to jointly optimize path length and unknown space coverage across global and local planning layers. The HPHS method [9] integrates a hybrid frontier sampling approach with hierarchical planning strategies, reducing the complexity of planning while enhancing the efficiency of autonomous exploration. Exploration based on spatial perception is sensitive to geometric structure, but it cannot efficiently navigate for specific target task requirements.

B. Semantic Information-driven Object Navigation

Object navigation typically utilizes object detection for semantic annotation and trains navigation policies through Reinforcement Learning (RL). Zhu et al. first conceived the concept of a target-driven model [10], which serves as the predecessor to object navigation. This approach was developed to tackle challenges concerning goal and scene generalizability while enabling straightforward implementation of end-to-end navigation. Santhosh Kumar Ramakrishnan [11] et al. proposed a novel network architecture

that predicts dual complementary potential functions from semantic maps, leveraging supervised learning to train the network for target-driven navigation. [12] employs a Goal-guided Transformer-based deep reinforcement learning strategy, significantly enhancing data efficiency during learning. Knowledge graph technology structures commonsense knowledge (e.g., object functionalities) to support task-level reasoning. The HOGN-TVGN [13] embeds prior knowledge into reinforcement learning policies. With the continuous progress of large model technology, the field of object navigation has gradually introduced the reasoning ability of LLMs or MLLMs for cost evaluation. At the same time, with the powerful natural language processing ability of large models, human-robot interaction can also be realized. For instance, [14], [15] build a visual-language navigation model based on LLMs, enabling information exchange with open-world environments. Additionally, the maturity of large-scale model technology has also driven the development of various zero-shot object navigation algorithms. The ZSON method [16] pioneers zero-shot object navigation, requiring no labeled data or demonstrations, and completely eliminates dependency on closed vocabulary. [17]–[23] achieve zero-shot object navigation without requiring task-specific training, prebuilt maps, or environmental priors. The above methods are highly dependent on semantic information, but lack complex processing of spatial information, and are easy to fall into local difficulties.

As evidenced by substantial prior research, numerous semantic-based object navigation methods focus their core objectives at the visual-semantic level, while opting for simplified frontier-based exploration or basic path planning algorithms at the spatial-structural level to guide robots. For instance, [24], [25] employ a simplified frontier-based exploration approach integrated with semantic information, while maintaining a 2D environmental map. For each frontier region, they leverage an LLM to evaluate and select the optimal frontier, significantly enhancing the efficiency of object navigation. Specifically, for each candidate frontier region, an LLM-generated relevance score is assigned based on semantic associations (e.g., object-context co-occurrence likelihood). The frontier with the highest composite score is selected as the next exploration target, thereby enhancing target-oriented exploration efficiency. In contrast, our method integrates the coverage of spatial surfaces with the distribution and relevance of open-world semantic information. It operates respectively at both the viewpoint-level and region-level hierarchies, thoroughly integrating spatial and semantic information. These components interlock closely, jointly guiding the robot to complete object navigation tasks under zero-shot conditions.

III. METHOD

Our method proposes a collaborative navigation framework integrating spatial cognition and semantic reasoning. The system leverages a point cloud map constructed and maintained in real-time by laser-based SLAM as its foundation. Through uniform partitioning of the environment,

a spatial grid map is established. Simultaneously, LLMs-driven semantic priority evaluation model (Sec. 3.1) generates relevance scores for target objects, forming a dynamically updated semantic associative topological graph. The core design centers on the hierarchical viewpoint-region collaborative evaluation model (Sec. 3.2), which screens high-value observation poses using dual viewpoint-level metrics: geometric coverage and semantic density, and activates sub-region exploration sequences based on region-level semantic association strength and spatial distribution. Further, we introduce the coverage-Aware sub-regions memory model (Sec. 3.3) that optimizes sub-region traversal logic via a state-switching mechanism based on the camera’s viewing coverage to mitigate redundant exploration. The overall framework establishes a closed-loop “Perception-Reasoning-Decision-Memory” process. The spatial grid map facilitates efficient boundary detection and region segmentation, while the semantic topological map provides global and local information gain. Final viewpoint generation, region activation, and dynamic memory attenuation are coordinated to prioritize exploration of high-semantic value regions, optimizing object navigation efficiency.

A. Semantic Prioritisation Evaluation Based on Large Language Models

This module proposes a target object relevance calculation method that integrates spatiotemporal feature analysis with semantic reasoning, aiming to enhance target navigation efficiency in complex environments. The system employs a three-layer processing architecture: first, fusing multi-frame perceptual data through a spatiotemporal buffering mechanism; then utilizing the DeepSeek-V3 [26] LLM for semantic relevance reasoning; and finally establishing an “object-relevance” memory through dynamic cache optimization to improve real-time performance.

The spatiotemporal buffer module uses a sliding window algorithm to store the detection results, and the hash key value $k = (\lfloor \frac{x}{0.1} \rfloor, \lfloor \frac{y}{0.1} \rfloor)$ to spatially discretize the semantic point data, this mechanism effectively solves the problem of dense item group detection interference, such as detecting multiple adjacent chairs in an office scenario, and avoids redundant reasoning about duplicate items by retaining the detection result with the highest confidence within each grid cell.

As shown in fig. 2, the semantic reasoning process integrates two key modules: semantic parsing and relevance evaluation, each handled by specialized models to enhance modularity and inference speed. Semantic parsing is crucial for human-robot interaction. Structured few-shot exemplars are designed to concretize natural human utterances into target object identifiers e_{target} (e.g., parsing “Help me find the fire_extinguisher” to “fire_extinguisher”). For relevance evaluation, a three-level chain-of-thinking evaluation model is constructed: firstly, scene localization is performed to establish a strong mapping relationship between the target items and a typical scene, then the scene topology analysis is performed to determine the topological connectivity

strengths of all the detected items within the scene, and finally, the degree of functional coupling between items is verified, and based on the first two constraints, further scoring is performed according to the degree of functional further scoring based on the degree of association based on the first two constraints. The system adopts a dynamic scoring constraint mechanism to ensure that the scoring results conform to the laws of the physical world through a rigid hierarchical division. The above model is designed to guide the large model in output of standardized scores by designing structured cue word templates, and finally the temperature coefficient of the model is reduced to ensure the stability of the inference.

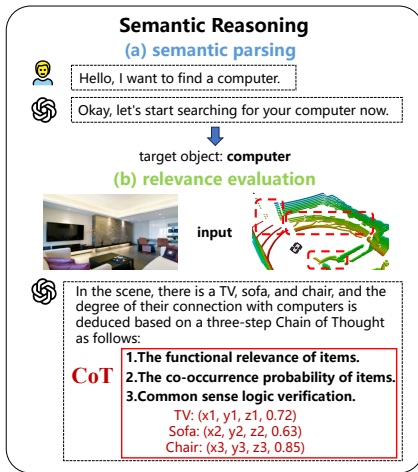


Fig. 2. **semantic reasoning level.** (a) Semantic parsing section: The LLM processes natural language to derive the target item. (b) Relevance evaluation section: Taking the current frame’s visual field image and coordinates as input, this module evaluates the relevance of all detectable items within the view to the target item, along with their global coordinates.

The system adopts a two-layer cache optimization strategy: in order to reduce the reasoning time brought about by the interaction step, the LRU cache is used. In order to reduce the reasoning time that may be brought about by the interaction step, a certain number of recent semantic parsing results are maintained through the LRU caching mechanism, and the timeliness is constrained to be 24 hours, which ensures that the parsing of the system on the language can be run for a long time. Dynamic deletion of records. The real-time processing module fuses the spatiotemporal buffer data with the semantic scores to generate point cloud data containing 3D coordinates and confidence:

$$P = \{(x_i, y_i, z_i, \max(\alpha S_s, I_{\text{target}}))\} \quad (1)$$

where S_s denotes the semantic relevance, α denotes the dynamic relevance parameter, and I_{target} is the minimum semantic relevance, which is guaranteed not to be zero.

B. Hierarchical Viewpoint-Region Collaborative Evaluation Model

This method constructs a two-tier assessment system at the viewpoint level and the regional level, forming a collabora-

tive decision-making assessment model from the viewpoint level to the regional level.

As shown in fig. 3: At the viewpoint screening level, a

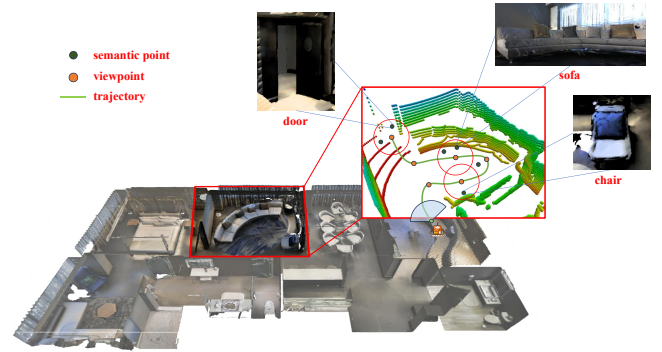


Fig. 3. **viewpoint screening level.** Viewpoints (orange points) are randomly generated within the current accessible area. Each viewpoint undergoes dual screening based on geometric coverage and semantic density. Viewpoints encompassing higher concentrations of semantic points (dark green points) and newly captured uncovered surfaces (voxel point clouds scanned for the first time) within the red circles correlate with a higher probability of selection.

large number of viewpoints are randomly sampled in the accessible area, and the sampled viewpoints are screened by the dual indexes of geometric coverage and semantic density. Geometric coverage counts the amount of point clouds on uncovered surfaces from viewpoints under a certain perspective, and the higher the number, the higher the spatial exploration value, the expression is shown below:

$$S_{\text{cov}}(v_j) = N_{\text{unobserved}}(v_j) \quad (2)$$

Semantic density counts the number of valid semantic points within a certain range for each viewpoint, the higher the density the higher the semantic exploration value, the expression is shown below:

$$S_{\text{sem-density}}(v_j) = \sum_{p_k \in \mathcal{N}_r(v_j)} I(p_k \in P_{\text{semantic}}) \quad (3)$$

where $\mathcal{N}_r(v_j)$ denotes that within the cylindrical neighbourhood (height set to infinity, height not considered) of viewpoint v_j with radius r , P_{semantic} is the set of all item points detected, and the $I(\bullet)$ function is used to count the number of effective number of items, and the meaningless detection words are eliminated by the large language model. From the above, the evaluation function for each viewpoint is as follows:

$$S_{\text{viewpoint}}(v_j) = \lambda_1 \cdot S_{\text{cov}}(v_j) + \lambda_2 \cdot S_{\text{sem-density}}(v_j) \quad (4)$$

where λ_1 and λ_2 are dynamic coefficients, adjusted based on spatial structures and semantic distributions across scenarios, to achieve a user-defined dynamic balance between geometric coverage and semantic density in the evaluation function.

At the region activation layer, the sub-region scores are determined by the number of viewpoints within the sub-region as well as the semantic association scores. The sub-region semantic association score is weighted by scoring

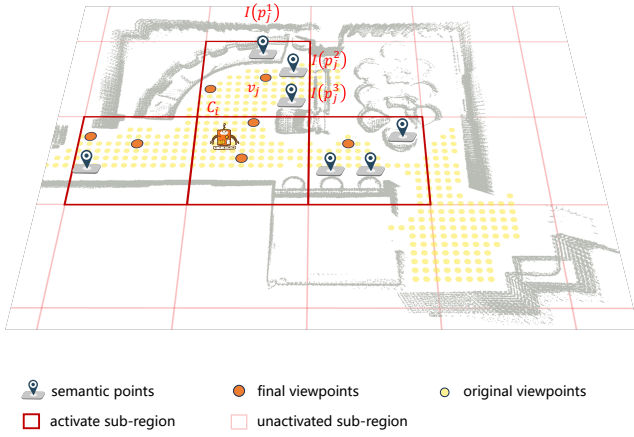


Fig. 4. **region activation layer.** In sub-regions of predetermined size, a higher density of viewpoints correlates with closer proximity of semantic points within the viewpoints' vicinity and elevated relevance scores. This collective metric elevates the sub-region's overall score. Sub-regions persistently exceeding the scoring threshold are designated as activated, with the activated sub-region list undergoing dynamic updates.

each viewpoint within a certain range of neighbourhood, and the total semantic association score is obtained by summing up the scores of all viewpoints in the sub-region, which is divided into two phases; the first phase searches globally with the maximum radius (the connecting circle in the sub-region), and locates the highly-associated targets based on the descending order of the semantic strength, and returns the results directly if there are points exceeding the threshold value; The second stage searches for semantic points within a certain range of neighbourhood of the viewpoint, and calculates the Gaussian decay weight of the viewpoint to get the semantic association score of the viewpoint, and we define ω_j^i as the weight value of the i -th semantic point in the neighbourhood of the viewpoint v_j , and then the formula of the weight function is shown as follows:

$$\omega_j^i = \exp\left(-\frac{\|p_j^i - v_j\|^2}{\sigma^2}\right), \quad \sigma = \frac{\sqrt{2}}{2}r_{xy} \quad (5)$$

where $p_j^i = (p_{jx}^i, p_{jy}^i)$ are the coordinates of the semantic points and $v_j = (v_{jx}, v_{jy})$ are the coordinates of the viewpoints, r_{xy} is the radius of the neighbourhood of the viewpoint.

As a result, the semantic relevance of the i -th semantic point in the neighbourhood of viewpoint v_j is defined as $I(p_j^i)$, and the formula of the semantic relevance evaluation function for each viewpoint is shown below:

$$\bar{S}_{v_j} = \frac{\sum_i \omega_j^i I(p_j^i)}{\sum_i \omega_j^i} \quad (6)$$

As a result, the evaluation function formula for the i th sub-region C_i is shown below:

$$\bar{S}_{C_i} = \left(\sum_{v_j \in C_i} S_{\text{viewpoint}}(v_j) \right) + \bar{S}_{v_j} \quad (7)$$

When \bar{S}_{C_i} exceeds the threshold, the subregion is activated and it is included in the list to be explored. Upon acquiring the list of regions to be explored, in light of the total semantic weighted scores of the sub-regions within the list, a greedy strategy is employed. The traversal sequence of the sub-regions is computed in descending order of scores, thus deriving the regional-level guiding path. The sub-regions in the list of regions to be explored will be updated dynamically in accordance with the scores, constantly steering the robot towards regions featuring denser semantic information, higher semantic relevance weights, and larger unexplored surfaces.

C. Coverage-Aware Sub-regions Memory Model

For the activated sub-regions, we set the state switching function, adding the “worthless” state in the inactive and active states, and realize the memory management of the environment sub-regions through real-time perception and coverage evaluation based on Field of View (FOV) during the robot navigation. When the robot updated its pose each time, only the current sub-region was accumulated by discrete sampling according to the plane sector field of view of the camera. When the coverage ratio reaches the threshold when it is detected to leave the sub-region, it will be marked as “worthless” at the moment of leaving, so that it will not be selected in the subsequent global planning.

Each sub-region is uniformly subdivided into $N \times N$ grid cells. In every update of robot pose, we only process the current grid cell: using a planar sector FOV model with radius R and angle α , we mark grid cells that fall within the sector as “seen”. This yields a monotonically increasing set of seen grid cells, used to approximate how much of the grid cell has actually been scanned. The coverage ratio is defined as:

$$\text{cov} = \frac{|\mathcal{S}_{\text{seen}}|}{N^2} \quad (8)$$

where $\mathcal{S}_{\text{seen}}$ is the set of grid cells that have been marked as seen. A grid cell is considered seen by the current FOV if it lies within the planar sector (ignoring height):

$$\text{Visible}(g) = \mathbb{I}[\|\mathbf{v}_g\| \leq R \wedge \angle(\mathbf{h}_t, \mathbf{v}_g) \leq \alpha/2] \quad (9)$$

where \mathbf{v}_g is the 2D vector from the robot to the grid cell center and \mathbf{h}_t is the 2D heading estimated from the last two robot positions (if the displacement is too small, the previous heading is reused). State switching is checked only at the moment the robot leaves a sub-region. If the accumulated coverage ratio of the just-left sub-region reaches the threshold τ , that sub-region is set to “worthless”; otherwise its state remains unchanged: If cov is greater than or equal to τ , we switch the sub-region to the “worthless” state. Upon entering a new sub-region, we activate its state, reset the coverage bitmap to prevent interference between adjacent grids, and promptly perform a FOV accumulation step based on the current pose. The state transition based on view scanning can enhance the robot's memory during the target navigation process, prevent redundant exploration of the same areas, and thereby improve the overall efficiency of object navigation.

IV. EXPERIMENTAL RESULTS

The proposed method was validated in both simulation environments and real-world experiments. Through comparisons with multiple current state-of-the-art methods, the metrics utilized by our approach were superior to those of the current mainstream methods.

A. Simulation Experiment

We carried out numerous tests within the combined simulation environment of Habitat and Gazebo, employing the MP3D [2] dataset and the HM3D [3] dataset. For the simulation experiments, semantic information and a simulated spatial environment are essential. Therefore, it is necessary to convert the semantic datasets into point clouds to acquire solid models. The semantic information is loaded in Habitat, while the spatial information is synchronously retrieved in Gazebo. The outcomes of the simulation experiments are presented in fig. 5.

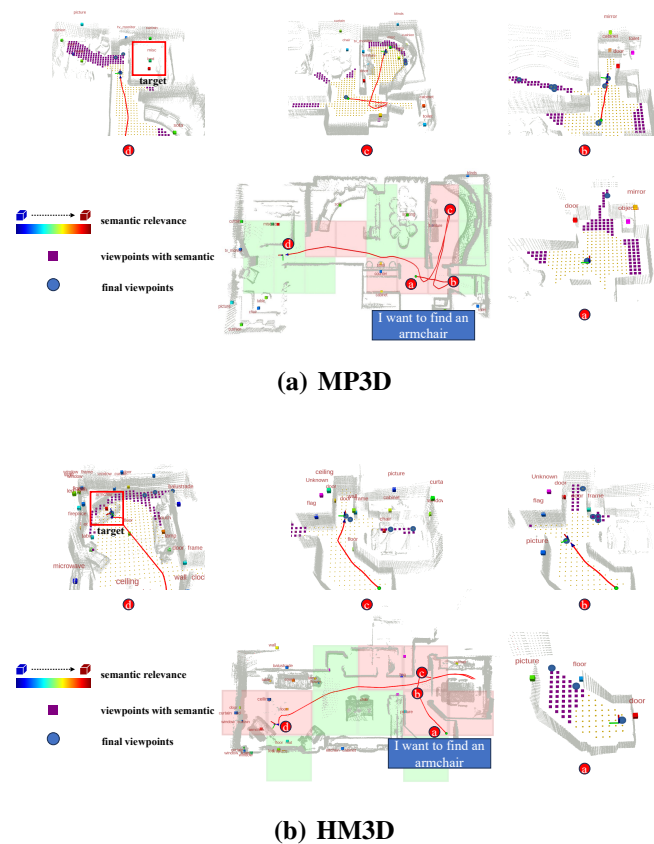


Fig. 5. **the outcomes of the simulation experiments.** In the figure, the yellow point clouds on the ground denote the original sampling viewpoints, the purple point clouds represent the sampling viewpoints under semantic weighting, and the blue spheres indicate the final viewpoints obtained from the current screening process. The path will be connected based on the final viewpoints and updated dynamically.

Our method was compared against the following five previous zero-shot goal navigation methods, evaluated using the metrics SR and SPL. The comparison methods included:

ZSON [16]: Proposes an open-world navigation novel paradigm that supports free-form linguistic instructions, en-

abling the accomplishment of target navigation tasks without requiring training.

InstructNav [17]: A general instruction navigation system capable of executing different types of instruction in unknown environments.

VLFM [18]: Proposed a zero-shot method capable of performing goal-driven semantic navigation without task-specific training, pre-built maps, or environmental prior knowledge.

VoroNav [19]: A zero-shot goal navigation framework based on Voronoi diagrams.

UniGoal [20]: A universal zero-shot goal-oriented navigation framework that unifies different goal types (including object categories, instance images, and text descriptions) via a unified graph representation method.

Considering the characteristics of the MP3D [2] and HM3D [3] datasets, more than five trials were carried out per scene, each targeting a different object. That is, over five distinct target items were selected for experimentation per scene. The experimental results are shown in table I.

To test the system’s capability for natural language understanding, several different descriptive formats were selected. For instance, goal instructions such as “Help me find the monitor”, “I want to find a table”, and “Someone might need a chair” were used. This ensured that the system could correctly interpret the natural language commands and identify the target object.

TABLE I
VISION-LANGUAGE NAVIGATION PERFORMANCE (SR/SPL)

Method	SR (%)		SPL	
	MP3D	HM3D	MP3D	HM3D
ZSON	15.3	25.5	0.048	0.126
VoroNav	–	42.0	–	0.260
VLFM	36.4	52.5	0.175	0.304
InstructNav	–	58.0	–	0.209
UniGoal	41.0	54.5	0.164	0.251
SSR-ZSON (Ours)	59.5	65.7	0.345	0.391

B. Real-World Experiment

Our real physical platform is equipped with a 12-core Arm Cortex-A78AE v8.2 64-bit processor, 64GB of RAM, 275TOPS of AI computing power, and Ubuntu 20.04 LTS. An Intel RealSense D435i depth camera and a Livox-Mid360 lidar sensor were installed on the platform.

The process and outcomes of the physical verification conducted on the experimental platform within a long corridor featuring multiple branches and rooms are illustrated in fig. 6. Upon receiving natural language instructions—such as the task of locating a desk—the platform demonstrates a preference for navigating toward areas that contain doors, entrance signs, and office-related objects. During the experiment, regions with closed doors or obstructed entrances are not designated as exploration zones, such as the goal of locating a desk—the platform demonstrates a preference for navigating toward areas that contain doors, entrance signs,

and office-related objects. During the experiment, regions with closed doors or obstructed entrances are not designated as exploration zones, and target search operations are not initiated in these areas. This approach enhances both the computational efficiency of the platform and the overall efficiency of object navigation.

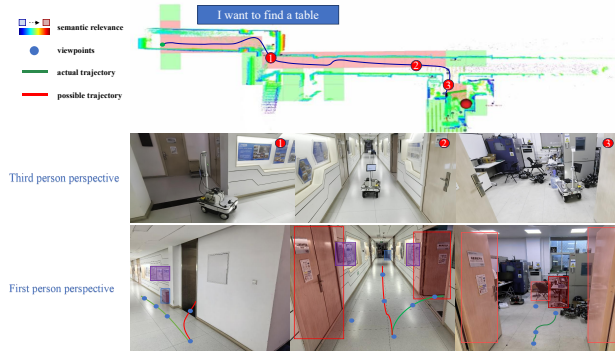


Fig. 6. **the process and outcomes of the physical verification.** The experimental platform is physically validated in a long corridor containing multiple fork roads and rooms.

V. CONCLUSIONS

We presented the SSR-ZSON, a hierarchical framework for zero-shot object navigation that integrates LLMs with spatial-semantic reasoning to address exploration in unknown open-world. The core innovation lies in its viewpoint-region joint assessment mechanism, which combines spatial coverage and semantic density at the viewpoint level, and LLM-derived target relevance weights at the region level, to prioritize exploration toward semantically significant areas. Additionally, a collaborative memory model reduces redundant exploration by tracking spatial coverage integrity and semantic consistency. Experimental results demonstrate the effectiveness of SSR-ZSON. On the MP3D and HM3D datasets, SSR-ZSON achieves success rates of 59.5% and 65.7%, respectively, while the corresponding SPL are 0.345 and 0.391 that outperform existing state-of-the-art methods. These results validate that SSR-ZSON significantly enhances both navigation success and efficiency, particularly in semantically sparse environments. The framework contributes to embodied AI by deeply combining LLM-driven semantic comprehension with robust spatial analysis, enabling real-time operation in combined simulation (Habitat-Gazebo) and physical platforms. However, the current approach relies on predefined semantic associations and static environmental representations, which may limit adaptability in highly dynamic settings. Future work will focus on extending SSR-ZSON's scene reasoning capabilities, incorporating a scene prediction module, and further improving the semantic association model's adaptability to dynamic environmental variations.

REFERENCES

[1] C. Cao, H. Zhu, H. Choset, and J. Zhang, "Tare: A hierarchical framework for efficiently exploring complex 3d environments," in *Robotics: Science and Systems*, 2021.

[2] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3D: Learning from RGB-D data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[3] S. K. Ramakrishnan, A. Gokaslan, E. Wijmans, O. Maksymets, A. Clegg, J. M. Turner, E. Undersander, W. Galuba, A. Westbury, A. X. Chang, M. Savva, Y. Zhao, and D. Batra, "Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, vol. 1, 2021.

[4] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. Towards New Computational Principles for Robotics and Automation*, 1997, pp. 146–151.

[5] S. M. Lavalle, "Rapidly-exploring random trees: A new tool for path planning," *The annual research report*, 1998.

[6] H. Umari and S. Mukhopadhyay, "Autonomous robotic exploration based on multiple rapidly-exploring randomized trees," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1396–1402.

[7] M. Selin, M. Tiger, D. Duberg, F. Heintz, and P. Jensfelt, "Efficient autonomous exploration planning of large-scale 3-d environments," *IEEE Robotics and Automation Letters*, vol. 4, pp. 1699–1706, 2019.

[8] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, and R. Siegwart, "Receding horizon "next-best-view" planner for 3d exploration," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 1462–1468.

[9] S. Long, Y. Li, C. Wu, B. Xu, and W. Fan, "Hphs: Hierarchical planning based on hybrid frontier sampling for unknown environments exploration," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 12056–12063.

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

[11] S. K. Ramakrishnan, D. S. Chaplot, Z. Al-Halah, J. Malik, and K. Grauman, "Poni: Potential functions for objectgoal navigation with interaction-free learning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18 868–18 878.

[12] W. Huang, Y. Zhou, X. He, and C. Lv, "Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1832–1845, 2024.

[13] B. Yang, X. Yuan, Z. Ying, J. Zhang, B. Song, Y. Song, F. Zhou, and W. Sheng, "Hogn-tvgn: Human-inspired embodied object goal navigation based on time-varying knowledge graph inference networks for robots," *Advanced Engineering Informatics*, vol. 62, p. 102671, 2024.

[14] B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang, "Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 7, pp. 5945–5957, 2025.

[15] G. Zhou, Y. Hong, and Q. Wu, "Navgpt: Explicit reasoning in vision-and-language navigation with large language models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 7, 2024, pp. 7641–7649.

[16] A. Majumdar, G. Aggarwal, B. Devnani, J. Hoffman, and D. Batra, "Zson: zero-shot object-goal navigation using multimodal goal embeddings," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2022, pp. 38 096–38 108.

[17] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, "Instructnav: Zero-shot system for generic instruction navigation in unexplored environment," in *8th Annual Conference on Robot Learning*, 2024.

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

[19] P. Wu, Y. Mu, B. Wu, Y. Hou, J. Ma, S. Zhang, and C. Liu, "VoroNav: Voronoi-based zero-shot object navigation with large language model," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235, 21–27 Jul 2024, pp. 53 757–53 775.

- [20] H. Yin, X. Xu, L. Zhao, Z. Wang, J. Zhou, and J. Lu, "Unigoal: Towards universal zero-shot goal-oriented navigation," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, June 2025, pp. 19 057–19 066.
- [21] S. Y. Gadre, M. Wortsman, G. Ilharco, L. Schmidt, and S. Song, "Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 23 171–23 181.
- [22] V. S. Dorbala, J. F. Mullen, and D. Manocha, "Can an embodied agent find your "cat-shaped mug"? Llm-based zero-shot object navigation," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4083–4090, 2024.
- [23] W. Cai, S. Huang, G. Cheng, Y. Long, P. Gao, C. Sun, and H. Dong, "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 5228–5234.
- [24] B. Yu, H. Kasaei, and M. Cao, "L3MVN: Leveraging large language models for visual target navigation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 3554–3560.
- [25] D. Shah, M. R. Equi, B. Osiński, F. Xia, B. Ichter, and S. Levine, "Navigation with large language models: Semantic guesswork as a heuristic for planning," in *Proceedings of The 7th Conference on Robot Learning*, 2023, pp. 2683–2699.
- [26] DeepSeek-AI, "DeepSeek-V3 Technical Report," 2024. [Online]. Available: <https://arxiv.org/abs/2412.19437>