Vision-Language Navigation for Indoor Construction Site Inspection using Quadruped Robots

Qihua Chen[1], Xianfei Yin[1,*], Zeying Gong[2], Zihua Zhu[1]

[1]Department of Architecture and Civil Engineering, City University of Hong Kong, Hong Kong SAR, China;
[2]Artificial Intelligence Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China; *Corresponding Author
qi-hua.chen@my.cityu.edu.hk, xianfyin@cityu.edu.hk, zgong313@connect.hkust-gz.edu.cn, zihuazhu2-c@my.cityu.edu.hk

Abstract –

Automated inspection of indoor construction sites is pivotal for enhancing project quality management and operational safety. Mobile inspection robots possess the potential to gradually replace human workers. However, characterized by high dynamics, clutter, and visual monotony, construction environments pose severe navigational challenges. Current SLAM solutions impose heavy computational burdens and face robustness challenges in such environments. Therefore, this study proposes a Vision-Language Navigation framework tailored for indoor construction site inspection. Specifically, we leverage the robust generalization of pre-trained Vision-Language Models (VLMs) to construct a two-stage navigation agent controlling quadruped robots. In the first stage, we utilize building floor plans and scene graphs to plan movement paths within the facility, thereby generating detailed navigation sub-tasks and path descriptions. In the second stage, the navigation agent outputs the subsequent action sequence based on current sub-task information and continuously updated historical navigation memory until the navigation task is completed. The proposed framework was validated within a custom three-story educational building construction scenario simulated in the Habitat3.0 simulator. Experimental results demonstrate that the path generation module achieves high semantic alignment with ground truth. Furthermore, the agent exhibits execution capabilities in challenging scenarios, such as texture-less cross-room traversal and stair climbing. This work provides a new technical pathway for achieving low-cost and flexible intelligent construction inspection, advancing construction automation.

Keywords –

Construction inspection; Vision-language navigation; Vision-language models; Quadruped robots; Building information modeling (BIM)

1 Introduction

The construction industry is widely recognized as one of the most hazardous and labor-intensive sectors[1,2], where safety inspections and quality monitoring necessitate extensive manual supervision [3]. However, traditional manual inspection methods are inefficient, heavily reliant on human labor, and susceptible to errors and fatigue [4,5]. Concurrently, the escalating costs and safety risks associated with the industry's labor shortage have underscored the urgent need for automated inspection solutions. Although previous studies have developed various methods for specific monitoring tasks with notable success, such as personal protective equipment violation detection [2] or wall defect identification[6], the critical aspect of autonomous navigation within the inspection process has often been overlooked. Recently, advancements in quadrupedal robotics have demonstrated significant potential for autonomously executing inspection tasks on construction sites. Mimicking the locomotion of four-legged animals, their versatility and advanced mobility enable effective deployment in complex environments [7].

A critical determinant of these robots' effectiveness is their ability to navigate precisely within complex construction environments. In indoor construction sites where GPS signals are restricted, structural occlusions make acquiring positional information extremely challenging [8]. Previously, Visual SLAM has been widely explored for its efficient indoor positioning performance, primarily involving mapping, localization, path planning, and control: mapping provides the spatial layout, while localization utilizes mapped features to determine the robot's position. However, due to the dynamic evolution of construction structures and layouts, along with temporary interference from moving objects, relying on static features for precise and reliable tracking and mapping proves unreliable. Although LiDAR-based SLAM algorithms have been employed for robot

localization due to higher accuracy [9], they suffer from high hardware costs, a lack of semantic information, operational complexity, and an over-reliance on unique geometric features for place recognition, leading to errors in geometrically similar areas [8,10]. Furthermore, these methods typically require time-consuming pre-scanning and pre-configuration for different environments [3]. Consequently, such rigid localization-based navigation approaches struggle to adapt to dynamic construction settings, necessitating a shift from geometric feature matching to a deeper semantic understanding of unstructured environments.

Vision-Language Navigation (VLN), as a significant domain of Embodied AI, has garnered increasing attention, requiring robots to execute navigation tasks in realistic environments based on natural language instructions [11]. Through immersive environmental interaction, navigation agents transcend traditional learning paradigms, developing profound adaptability and human-like understanding of complex dynamic worlds. Specifically, navigation can be simplified into a discrete direction selection task, typically modeled as a sequential decision-making process where agents act based on visual observations and language instructions. The realization of efficient VLN primarily relies on the cross-modal alignment capabilities of Vision-Language Models (VLMs), endowing them with potential for preliminary spatial scene understanding and logical reasoning; representative models include Qwen3-VL [12] and InternVL3 [13].

However, directly transferring existing VLM technologies to construction indoor navigation tasks faces severe challenges. First, the limitation of planning horizons is prominent: existing path planning is mostly based on local observations restricted to the agent's current field of view, lacking memory and reasoning capabilities for global topological structures, which leads to failure in long-horizon tasks due to the absence of global planning [11,14]. Simultaneously, construction sites present severe issues of "semantic sparsity": environments consist mainly of concrete walls, scaffolding, and unfinished structures with monotonous textures and a lack of distinctive semantic landmarks, making visual features insufficient to support effective self-localization. Furthermore, existing research typically assumes quadruped robots operate on flat indoor terrain [7], lacking exploration of vertical stair scenarios[15]; for instance, Hu and Gan [16] and Sun et al. [11] only considered single-floor navigation, resulting in a current lack of complete construction scenario simulation data for validation.

To apply VLN to the navigation of insufficiently explored indoor construction scenarios, the following critical questions must be addressed: (1) How can existing construction resource data be utilized to plan long-distance navigation paths? (2) How can a vision-language navigation agent tailored specifically for construction scenarios be designed by leveraging the visual understanding capabilities of VLMs? (3) How can a construction scenario navigation simulation environment be constructed for validation purposes?

To address these research gaps, this study proposes a vision-language navigation framework tailored for indoor construction site inspection using quadruped robots. Specifically, this method utilizes scene floor plans derived from Building Information Modeling (BIM) and architectural scene graphs, combined with path planning algorithms to achieve efficient long-distance navigation. Furthermore, we design a VLM-based navigation agent and incorporate a visual-text memory module to overcome memory limitations in long-term sequential navigation. For validation purposes, a multi-story construction scenario is constructed within the Habitat [17] simulator. This study not only expands the potential application of vision-language navigation in unstructured construction environments but also provides a scalable technical pathway for the deployment of quadruped robots in automated construction inspection. By integrating BIM with Embodied AI, this approach is poised to significantly enhance autonomous perception and decision-making capabilities on construction sites, providing a new paradigm for addressing the industry's long-standing safety, efficiency, and labor shortage problems.

2 Methodology

Figure 1 illustrates the proposed vision-language navigation framework tailored for quadruped robots in construction scenarios, which comprises three core modules: (a) navigation data preparation: This module utilizes BIM as the sole source of environmental prior data. Through spatial/component encoding and plan exportation, it generates 2D floor plans containing multi-story information. A scene graph is constructed to abstract complex architectural spaces into a topological network composed of nodes (rooms/areas) and edges (connectivity), thereby providing structured data support for subsequent global planning. (b) Navigation path information generation: Based on the scene graph, this module performs coarse-grained path planning to generate a series of candidate paths. Subsequently, it leverages the reasoning capabilities of VLM to select the optimal path and transform it into a list of detailed spatial descriptions and step-by-step sub-tasks. (c) Vision-language navigation execution: In this stage, the VLM-driven agent receives the generated navigation sub-tasks, scene-specific navigation instructions, and navigation textual memory. During operation, the agent continuously processes visual information captured by

the quadruped robot alongside textual instructions to perform cross-modal reasoning. It then outputs specific action commands to drive the robot's movement within the environment, establishing a closed-loop control system of perception-decision-action. Each component is detailed in the following sections.

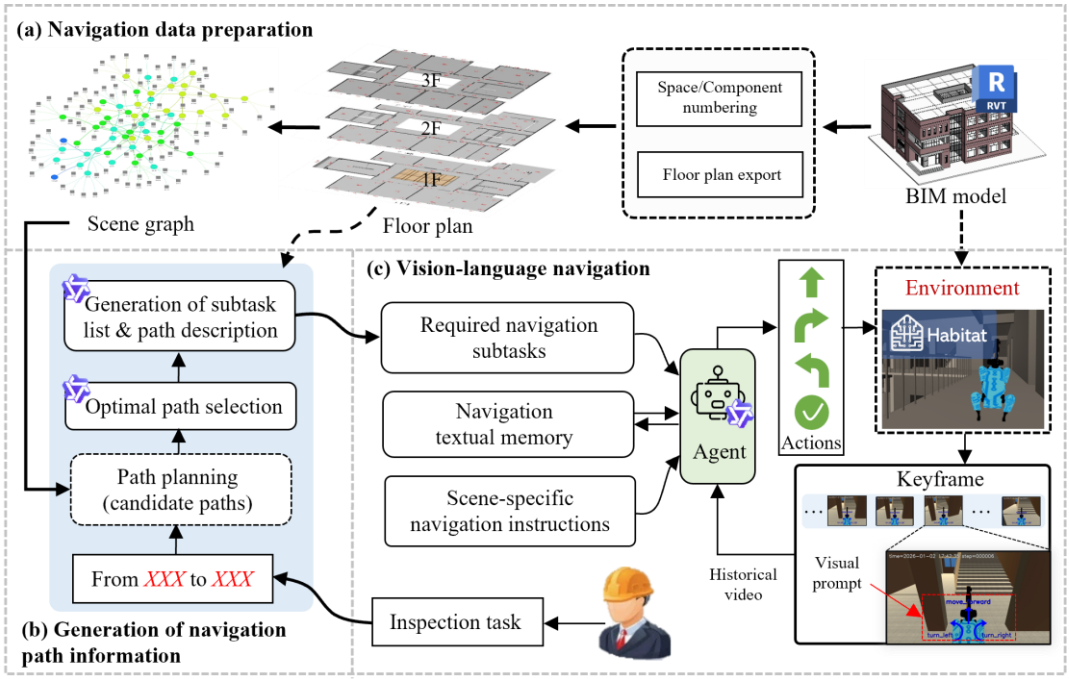


Figure 1. The framework of the proposed method

2.1 Navigation Data Preparation

BIM has emerged as a pivotal resource within the Architecture, Engineering, and Construction (AEC) industry, encapsulating comprehensive architectural information [18]. In this study, we leverage BIM as the primary source of prior environmental knowledge, extracting critical geometric and semantic features to construct structured spatial data required for navigation. Specifically, the data processing workflow comprises two distinct steps: floor plan construction and scene graph construction. For the building floor plan, 2D maps containing information on walls, doors, windows, and obstacles are exported by parsing BIM files, followed by a unified standardized coding of spatial regions and components within the building. We model the architectural environment as a collection of distinct semantic categories. Let $L = \{S, H, R, D, W, C\}$ denote the set of spatial element categories, where the symbols correspond to Staircase, Hallway, Room, Doorway, Window, and Column, respectively. For any arbitrary instance e within the environment, an encoding format of Type-Floor-ID is adopted. Here, the leading digit of the numeric suffix indicates the floor attribute. For instance, H201, R230, and S206 all pertain to the second-floor space; H102 represents a hallway instance on the first floor, while D1 denotes a connectivity interface on the first floor.

Regarding scene graph construction, a heterogeneous scene graph is established based on manually extracted spatial adjacency and inclusion relationships. This process transforms complex BIM models into lightweight, semantically rich navigation maps, providing a precise spatial data foundation for subsequent global path planning and vision-language instruction generation. As shown in Fig. 2, the nodes within the graph structure are categorized into two types of key entities: "Spatial Nodes," representing the centers of navigable areas such as rooms (R), hallways (H), and staircases (S); and "Component Nodes," which serve as critical visual landmarks to aid localization, corresponding to architectural elements like columns (C) and windows (W). Correspondingly, the edges encode two distinct spatial topological relationships: the "Accessibility Edge," which connects two spatial nodes, represents physical connectivity facilitated by doorways (D) and is utilized to support path searching; whereas the "Inclusion Edge," connecting a spatial node to a component node, indicates that the component resides within a specific space. This structured design not only preserves the global topological features required for navigation but also augments the robot's semantic understanding capabilities of the environmental context by associating specific architectural components.

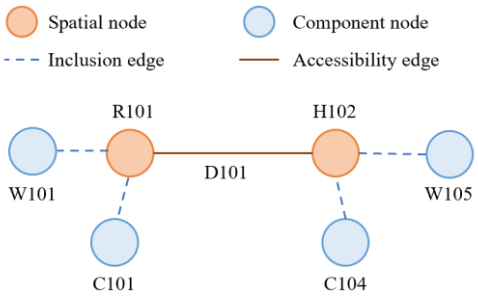


Figure 2. Scene graph example

2.2 Generation of Navigation Path Information

Path planning is a pivotal task and has been extensively explored in numerous studies. It aims to determine a sequence of actions that enables a robot to navigate from a starting position to a target location [19]. This study proposes a hierarchical path generation mechanism that integrates traditional algorithms with the reasoning capabilities of VLMs. First, the system parses the inspection task instruction given in natural language, mapping it to a start node n_{start} and a target node n_{target} within the scene graph. Subsequently, the A* algorithm [20] is employed to perform a global path search on the constructed topological network. To ensure the robustness of path planning and provide diverse options, this stage is not limited to identifying a single geometrically shortest path; instead, it generates a candidate path set $P_{cand} = \{p_1, p_2, ..., p_N\}$ containing N potential routes. Each candidate path p_i consists of a sequence of continuous spatial nodes (e.g., rooms, hallways) and their connecting edges, providing the necessary search space for subsequent optimization based on semantic understanding.

The powerful multi-modal reasoning capabilities of VLMs are utilized to select the optimal path from the candidate set P_{cand}. The 2D floor plan derived from BIM is input into the VLM as a visual prompt, while the node sequences of candidate paths are converted into textual descriptions. The VLM evaluates paths not only based on geometric distance but also by comprehensively reasoning about environmental semantics, including navigation complexity, the saliency of visual landmarks, and spatial accessibility. Through this process, the model can identify and discard paths that, while theoretically having the fewest steps between rooms, may traverse complex obstacle zones or lack distinct visual references. Ultimately, it determines an optimal path p_{opt} that satisfies geometric constraints while possessing high interpretability.

Based on the selected optimal path p_{opt}, the VLM further generates detailed navigation path descriptions and a structured list of sub-tasks. At this stage, the model is configured to generate content from the first-person perspective of the quadruped robot, ensuring that instructions align with the robot's actual perceptual viewpoint. The output comprises two parts: a global path description, which summarizes the overall route; and a navigation sub-task sequence $S = \{s_1, s_2, ..., s_M\}$ arranged in execution order. Each sub-task s_j explicitly specifies the current local navigation goal and specific action instructions. This semantically rich textual information is directly transmitted to the downstream vision-language navigation execution module, serving as high-level guidance for the robot to establish environmental cognition and execute actions.

2.3 Vision-Language Navigation

The navigation execution process constitutes the core component of this framework (as shown in Figure 1(c)), aiming to construct a Vision-Language Model (VLM)-based closed-loop control agent that transforms the abstract sub-task sequences generated in the previous stage into concrete action commands for the quadruped robot. The input and output architecture of the VLM is illustrated in Fig. 3. For each discrete time step t, the agent receives a multi-modal input $I_t = \{V_t, M_{text}, T_j\}$. Here, V_t represents the RGB temporal video stream captured by the front-facing camera, utilized to perceive dynamic environmental structures and visual changes; it consists of m video frames, where each frame contains visual directional prompts (corresponding to each action) to assist in orientation judgment. M_{text} comprises the historical navigation memory, with a maximum record capacity of ω. $T_j \subset T$ denotes the textual description of the specific sub-task currently being executed; based on this task description, the agent selects the required instructions from a Scene-specific Navigation Instruction Set (e.g., if the task involves stair climbing, relevant instruction descriptions for stairs are extracted).

All the variable inputs are integrated into a carefully designed prompt. The system employs a structured Prompt Engineering strategy to dynamically inject critical contextual information, comprising five distinct components. First, the "Task Overview" module includes the "Role & Objective," "Scene Description," "Camera Background," and "Sub-task Description," thereby guiding the model to perform cross-modal reasoning by combining visual perception with textual instructions. Second, to ensure the safety and reliability of the quadruped robot's navigation within complex architectural environments, the "Action Space Definition" strictly constrains operations to a discrete set $A = \{move\_forward, turn\_left, turn\_right, stop\}$. Specifically, *move_forward* denotes advancing forward by a distance l; *turn_left* and *turn_right* represent in-place rotation by an angle θ to the left or right, respectively; and *stop* indicates the termination of the task. Third, the "History Memory" module incorporates memory

descriptions to facilitate accurate utilization, alongside Video Memory (V_t) and Textual Memory (M_{text}). Fourth, "Navigation Strategies" detail the operational precautions during the navigation process, encompassing "Scene-specific Navigation Settings" and "Navigation Constraints." For instance, the Scene-specific Navigation Settings provides critical navigation policies, such as the requirement to prioritize generating turn actions to align the center of the target geometric structure directly to the front before executing critical actions like traversing doorways, entering hallways, or ascending/descending stairs. Fifth, the "Output Format" module defines the specific output structure to guarantee that the results are parsable.

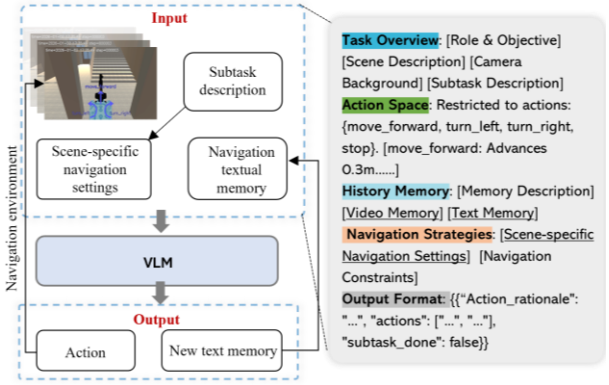


Figure 3. Inputs and outputs of the VLM-based agent during the navigation process

In the output and execution phase, the VLM is required to generate a response in a strict JSON format, encompassing the action sequence $a_{seq} \subset A$, sub-task completion status, and new text memory (Action Rationale). Here, a_{seq} represents the list of impending actions to be executed. To enhance the system's interpretability and debuggability, the model is required to output an "Action Rationale" grounded in strict facts, which is subsequently updated to the textual memory M_{text}. The system then dispatches a_{seq} to drive the robot's movement and proceeds to the next control loop $t+1$. If the completion status (done) is determined to be true, the system updates the current objective and transitions to the subsequent sub-task. This mechanism ensures that the quadruped robot can achieve robust autonomous navigation relying solely on visual perception and semantic understanding, even in scenarios lacking high-precision maps.

3 Case Study

This study selected a typical three-story educational building as the experimental case. Characterized by a complex environment featuring long corridors and stairwells, this building serves as a robust testbed to comprehensively evaluate the navigation capabilities of the proposed method within intricate architectural scenarios. The following sections will detail the Construction of the Navigation Environment, Results and Evaluation, and Practical Implications.

3.1 Construction of the Navigation Environment

3.1.1 Scene data processing

The initialization of the experimental environment was based entirely on the BIM of the educational facility. The original BIM model and its internal perspective views are presented in Fig. 4(a); the model comprehensively includes key architectural structures such as exterior walls, interior walls, beams, columns, and stairs, while retaining only doorways as valid navigable nodes. Vertical connectivity within the building is facilitated by two double-flight staircases that span across all three floors. The first floor covers a gross area of 930 square meters, and its generated 2D floor plan, along with the corresponding spatial encoding, is illustrated in Fig. 4(b). Through the parsing and extraction of BIM data, the experimental scene comprises a total of 60 columns, 39 windows, 29 independent spaces (encompassing corridors, staircases, rooms, and lobbies), and 59 connecting passageways (including channel widths). As shown in the figure, the unique identification numbers for all spaces and components are explicitly labeled in red text on the floor plan, and the ascending/descending positions of the staircases are also marked. Based on the aforementioned planar data and semantic information, the system constructed a topological scene graph for path planning. As demonstrated in the scene graph example in Fig. 4(c), the solid red line visually indicates the optimal path planning result from the starting node H105 to the target node H205.

3.1.2 Simulation Setup and Implementation Details

This study utilizes the Habitat 3.0 simulator as the core simulation platform to construct a high-fidelity three-dimensional architectural navigation environment. Renowned for its efficient physics engine and realistic lighting rendering capabilities, Habitat 3.0 supports high-frame-rate physical interaction simulations, thereby effectively bridging the gap between simulation and the real world. We converted the processed educational building BIM model into a simulator-compatible GLB mesh format via a Revit plugin and imported it into the scene, ensuring that the layout of walls, stairs, and obstacles in the virtual environment maintains high geometric consistency with the real building. Furthermore, the simulation environment enabled

continuous Navigation Mesh computation, precisely defining non-traversable areas and traversable zones, which provides an accurate testing benchmark for validating the robot's obstacle avoidance and path planning capabilities within complex indoor structures.

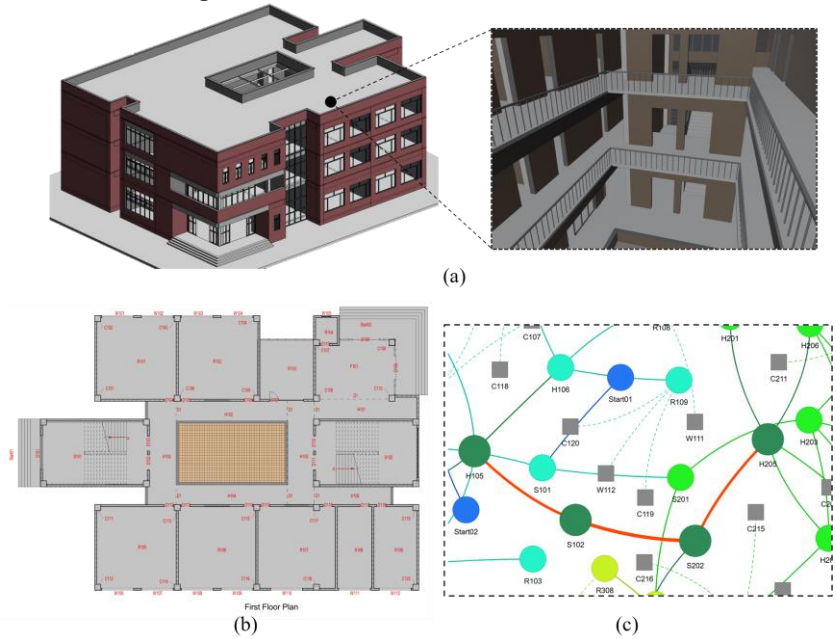


Figure 4. BIM data and scene graph examples of the case building. (a) Original BIM model and its internal view; (b) Ground floor plan with annotations; and (c) Example scene graph with the shortest path indicated by the red line

Regarding agent configuration, the experiment employed a quadruped robot model as the navigation subject, with parameterized settings tailored to the physical characteristics of a real robot. To simulate the realistic perspective of a quadruped robot, we configured an RGB camera above the robot's center, as shown in Table 1. The camera installation height was set to 1.0 meter above ground, with a Field of View of 90 degrees and an image resolution of 1280×720 pixels, as shown in Fig. 5(a). Additionally, a rear-facing camera with a resolution of 640×480 and a downward tilt of 36 degrees was utilized solely for recording the experimental process and was not involved in the navigation control loop. The robot's physical collision radius was set to 0.28 meters, and the height to 0.55 meters; these dimensional constraints compel the robot to perform precise path planning when traversing narrow doorways or maneuvering in proximity to obstacles. The action space parameters were defined as a discrete control mode: the step length l for a single "move_forward" action is 0.3 meters, and the angle θ for a single "turn" action is 15 degrees. This configuration effectively mitigates cumulative errors in log-sequence navigation while maintaining navigation precision.

During the navigation process, the number of video frames m was set to 20, the length of the action sequence a_{seq} was limited to 2, and the parameter ω was set to 10. The VLM utilized the cloud-deployed Qwen3-VL-8B model, with a model temperature set to 0.7. The local Habitat environment was hosted on an Ubuntu 22.04.5 operating system equipped with an NVIDIA RTX 3060.

Table 1. Parameter settings for the quadrupedal robot

Module	Parameter	Value
Front Camera	Resolution	1280×720
	Field of view	90◦
	Pitch angle	0◦
	Mounting height	1.0 m
Back Camera	Resolution	640×480
	Field of view	90◦
	Pitch angle	36◦ (downward)
	Mounting height	1.6 m
Action Space	Forward step Length (l)	0.3 m
	Turn Angle (θ)	15◦

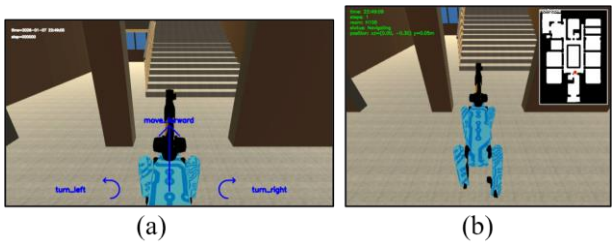


Figure 5. Camera views. (a) Front-facing (quadruped robot's perspective) and (b) Rear-facing (global perspective)

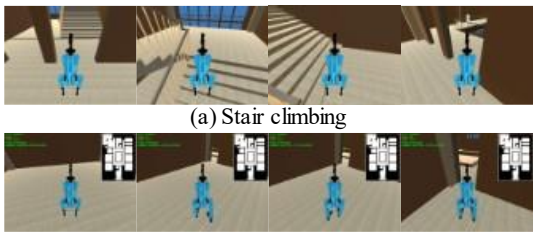
3.2 Results and evaluation

The effectiveness of the path planning and natural language description generation in the first stage was assessed through a quantitative evaluation of 20 generated navigation task descriptions. Model-generated text sequences were compared against human-provided ground truth path descriptions. The evaluation employed three metrics widely established in natural language processing to provide a comprehensive assessment: BLEU, ROUGE-L, and Token F1 score. Specifically, BLEU quantifies the n-gram precision match, assessing the exact overlap of vocabulary between the generated and reference texts. ROUGE-L focuses on the longest common subsequence to reflect structural recall, effectively measuring the sentence-level fluency and the correct sequential ordering of instructions. Furthermore, Token F1 evaluates the coverage accuracy of key tokens, ensuring that critical navigational entities are accurately retained. As presented in Table 2, the method achieved a BLEU score of 64.24, indicating lexical precision in the generated descriptions. Additionally, scores in ROUGE-L F1 (0.795) and Token F1 (0.803) confirm that the instructions are highly aligned with standard answers regarding semantic coherence and key information retention. These results demonstrate that the first-stage module effectively transforms abstract BIM topological paths into semantically rich instructions, establishing a planning foundation for the robot's execution.

Table 1. Quantitative evaluation of navigation path description generation

Metric	Score
BLEU	64.24
ROUGE-L F1	0.795
Token F1	0.803

The execution capability of the navigation agent in complex environments was subsequently verified using two representative scenarios, "Room-to-Room Navigation" and "Stair Climbing," with the navigation processes recorded by the back-facing camera visualized in Fig. 6. In the Room-to-Room task, the agent accurately identified the target doorway within a texture-less room, utilized the VLM's spatial understanding to align its heading with the center, and traversed the doorway to enter the adjacent room. For the more challenging Stair Climbing task, the agent exhibited vertical mobility; notably, at the stair landing, the agent successfully recognized spatial geometric changes and generated correct turning commands, facilitating smooth floor navigation. This performance indicates that the framework possesses superior capabilities when dealing with spatial bottlenecks and structural transitions.



(a) Stair climbing

(b) Room-to-room navigation

Figure 6. Navigation process in representative scenarios

4 Conclusion

This study proposes a novel Vision-Language Navigation framework tailored for indoor construction site inspection. To the best of our knowledge, this work presents one of the first integrations of Vision-Language Navigation with quadruped robots in unstructured indoor construction environments. By integrating the semantic reasoning capabilities of VLMs with topological scene graphs derived from BIM, we established a robust two-stage navigation mechanism. The first stage transforms abstract architectural information into semantically rich, executable natural language sub-tasks, achieving high lexical and structural alignment with ground-truth descriptions. The second stage enables the quadrupedal agent to execute closed-loop control based on visual perception and historical memory. Validation within the high-fidelity Habitat 3.0 simulation environment demonstrated the system's capability to navigate complex, texture-less scenarios without requiring pre-built high-precision maps. The theoretical and practical implications of this work are significant for the automation of the construction industry. By empowering quadruped robots with human-like semantic understanding and logical reasoning, this approach provides a scalable, low-cost technical pathway for replacing manual labor in hazardous inspection tasks, thereby enhancing operational safety and facilitating the digital transformation of construction management.

Despite these promising results, several limitations remain. First, the current validation is confined to the simulation environment; while visually realistic, it may not fully capture the chaotic sensory noise, unpredictable dynamic obstacles (e.g., moving workers), and unexpected layout variations prevalent in real-world construction sites. Consequently, the "Sim-to-Real" transfer gap remains a critical challenge. Second, the system currently relies on cloud-deployed VLMs, where network latency could potentially hinder real-time decision-making during high-speed maneuvers. Third, the inherent issues of hallucination and output instability in pre-trained VLMs have not been fully resolved; future efforts could involve fine-tuning domain-specific large models to mitigate these risks. Finally, evaluation on large-scale datasets is lacking. Future work will focus on

deploying the framework on physical quadruped robots to test its robustness in real-world scenarios, automating the extraction of scene graphs, optimizing the inference efficiency of edge-side models, developing a more flexible and context-aware navigation framework to handle dynamic elements and layout changes, and benchmarking against broader datasets to further enhance building indoor navigation safety and reliability.

Acknowledgements

The authors gratefully acknowledge the financial support provided by the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515010190) and the National Natural Science Foundation of China (Grant No. 72404233).

References

[1] Q. Chen, X. Yin. Tailored vision-language framework for automated hazard identification and report generation in construction sites. *Advanced Engineering Informatics*, 66 103478, 2025.

[2] Q. Chen, D. Long, S. Wang, Q. Chen, B. Yuan. Real-Time Detection of Personal Protective Equipment Violations for Construction Workers Using Semisupervised Learning and Video Clips. *Journal of Construction Engineering and Management*, 151 (3): 04024213, 2025.

[3] X. Zhao, C.C. Cheah. BIM-based indoor mobile robot initialization for construction automation using object detection. *Automation in Construction*, 146 104647, 2023.

[4] A. Yarovoi, Y.K. Cho. Review of simultaneous localization and mapping (SLAM) for construction robotics applications. *Automation in Construction*, 162 105344, 2024.

[5] Q. Chen, X. Yin. Tailored Vision-Language Solutions for Comprehensive Hazard Identification on Construction Sites. In *Proceedings of the 42nd ISARC*, pages 548-555, Montreal, Canada, 2025.

[6] W. Ren, Z. Zhong. Building construction crack detection with BCCD YOLO enhanced feature fusion and attention mechanisms. *Scientific Reports*, 15 (1): 23167, 2025.

[7] Z. Chen, C. Song, B. Wang, X. Tao, X. Zhang, F. Lin, J.C.P. Cheng. Automated reality capture for indoor inspection using BIM and a multi-sensor quadruped robot. *Automation in Construction*, 170 105930, 2025.

[8] L. Yang, H. Cai. Enhanced visual SLAM for construction robots by efficient integration of dynamic object segmentation and scene semantics. *Advanced Engineering Informatics*, 59 102313, 2024.

[9] X. Chen, Y. Yu. An unsupervised low-light image enhancement method for improving V-SLAM localization in uneven low-light construction sites. *Automation in Construction*, 162 105404, 2024.

[10] L. Xu, C. Feng, V.R. Kamat, C.C. Menassa. An Occupancy Grid Mapping enhanced visual SLAM for real-time locating applications in indoor GPS-denied environments. *Automation in Construction*, 104 230-245, 2019.

[11] L. Sun, A. Kanezaki, G. Caron, Y. Yoshiyasu. Enhancing multimodal-input object goal navigation by leveraging large language models for inferring room–object relationship knowledge. *Advanced Engineering Informatics*, 65 103135, 2025.

[12] Q. Team. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*. 2025.

[13] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*. 2025.

[14] D. An, H. Wang, W. Wang, Z. Wang, Y. Huang, K. He, L. Wang. ETPNav: Evolving Topological Planning for Vision-Language Navigation in Continuous Environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47 (7): 5130-5145, 2025.

[15] Z. Gong, R. Li, T. Hu, R. Qiu, L. Kong, L. Zhang, Y. Ding, L. Zhang, J. Liang. Stairway to Success: Zero-Shot Floor-Aware Object-Goal Navigation via LLM-Driven Coarse-to-Fine Exploration. *arXiv preprint arXiv:2505.23019*. 2025.

[16] D. Hu, V.J.L. Gan. Semantic navigation for automated robotic inspection and indoor environment quality monitoring. *Automation in Construction*. 170 105949, 2025.

[17] X. Puig, E. Undersander, A. Szot, M.D. Cote, T.-Y. Yang, R. Partsey, R. Desai, A.W. Clegg, M. Hlavac, S.Y. Min. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*. 2023.

[18] Z. Chen, H. Wang, K. Chen, C. Song, X. Zhang, B. Wang, J.C.P. Cheng. Improved coverage path planning for indoor robots based on BIM and robotic configurations. *Automation in Construction*. 158 105160, 2024.

[19] J. Cai, A. Du, X. Liang, S. Li. Prediction-Based Path Planning for Safe and Efficient Human–Robot Collaboration in Construction via Deep Reinforcement Learning. *Journal of Computing in Civil Engineering*. 37 (1): 04022046, 2023.

[20] F. Duchoň, A. Babinec, M. Kajan, P. Beňo, M. Florek, T. Fico, L. Jurišica. Path planning with modified a star algorithm for a mobile robot. *Procedia Eng.* 96 59-69, 2014.