Overcoming Visibility and Collision Challenges in Construction Robotics with Bird's-Eye-View and Vision-Language Navigation

Maryam Soleymani [1], Mahdi Bonyani[1], and Chao Wang[2]

***Abstract*— We introduce a vision-language navigation framework for simulated 3D construction sites that addresses challenges posed by cluttered layouts, occlusion, and instruction ambiguity. Unlike traditional panoramic or egocentric approaches, our method leverages Bird's-Eye-View (BEV) representations and a spatially grounded scene graph to model both local geometry and global topological structure. A dual-level decision mechanism integrates fine-grained grid-based reasoning with high-level graph-based inference, conditioned on natural language instructions. Additionally, we incorporate a 3D object detection head to enhance spatial perception through awareness. Experimental results in simulated construction environments show that our model significantly improves navigation success, path efficiency, and object grounding accuracy, highlighting the benefits of BEV reasoning for complex construction site scenarios.**

I. INTRODUCTION

The construction industry is increasingly turning to robotic systems to improve efficiency, safety, and precision in various tasks from monitoring and inspection to material delivery and installation. Effective navigation is fundamental to these applications, as construction sites present unique challenges with their dynamic, complex, and often hazardous environments [1], [2]. Also, High-level task planning for construction robots involves identifying required actions, determining their sequence, and matching these actions with robot skills like navigation, picking, and placing [3]. These robot behaviors must leverage contextual information from both real-time sensor-based perception and embedded knowledge from large language models (LLMs). In addition, navigation skills can be activated by providing destination prompts, which then generate and continuously update paths based on observed obstacles [4].

While aerial robots offer unique advantages for construction site inspection and mapping tasks, ensuring their safe navigation near human workers remains challenging. In addition, the limited field of view of camera systems constrained by payload limitations often results in unreliable perception during collision avoidance in dynamic environments [2]. Bird's-Eye-View (BEV) navigation approaches have emerged as effective solutions to the visibility limitations faced by ground robots navigating complex construction environments. The overhead perspective provides a comprehensive view of the site that enables more effective navigation planning and obstacle avoidance. As construction sites are highly cluttered, ground-based robots often struggle with limited visibility, making it difficult to navigate efficiently [5].

Vision-Language Navigation (VLN) represents a significant advancement in embodied AI research, enabling autonomous agents to navigate through complex environments by following natural language instructions while processing visual observations [6]. In construction sites, where environments are dynamic and filled with obstacles, VLN systems provide robots with the capability to understand and follow verbal commands while adjusting their actions based on environmental feedback [7]. One of the key challenges in implementing VLN in construction environments is collision avoidance. Many existing VLN approaches focus primarily on translating discrete navigation methods to continuous environments without adequately addressing collision problems, which can cause robots to deviate from planned paths or become trapped in obstacle areas [8]. This limitation is particularly problematic in construction sites, where unexpected obstacles frequently appear.

The key contributions of our work include:

- We propose a novel vision-language navigation framework that integrates Bird's-Eye-View (BEV) representations and spatial-temporal scene graphs for robust navigation in simulated 3D construction environments.
- We introduce a dual-level decision mechanism that combines grid-level and graph-level reasoning conditioned on natural language instructions, enabling precise and interpretable action selection.
- We enhance spatial understanding through a 3D object detection module that provides object-level supervision to the BEV encoder, significantly improving navigation in occluded and cluttered scenes.

II. METHODOLOGY

This study proposes a vision-language navigation framework designed for simulated 3D construction sites, where agents must interpret natural language instructions to traverse a dynamically generated, object-rich scene. Our approach consists of three main modules: (1) geometric scene encoding from multi-view images, (2) spatio-linguistic graph construction and updating, and (3) hierarchical decision scoring for navigation.

This material is based upon work supported by the National Science Foundation under Grant No. 2222881.

[1] are Ph.D. Students, Bert S. Turner Department of Construction Management, Louisiana State University, USA msoley1@lsu.edu, mbonya1@lsu.edu

[2]Associate Professor and Graduate Program Advisor, Bert S. Turner Department of Construction Management, Louisiana State University, USA chaowang@lsu.edu

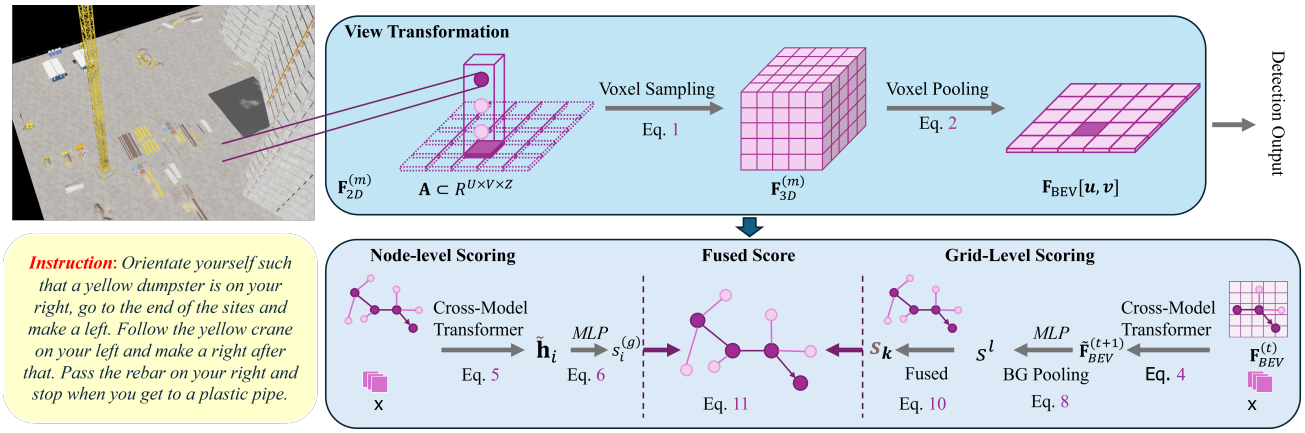

Fig. 1. An overview of the proposed method. The process begins by transforming multi-view images into a BEV plane format. Following this, the BEV features undergo encoding through 3D detection methods. During navigation, the system incorporates BEV representations to generate two decision scores: one at the node-level scoring and another at the grid-level scoring based on BEV data. These complementary scores are then combined to enable more effective decision-making processes.

A. Geometric Representation from Multi-View Inputs

At each simulation step t, the agent captures a set of M RGB images $\{\mathcal{I}_m\}_{m=1}^{M}$ from different viewing angles at its current camera. Each image is passed through a CNN-based encoder ϕ to produce a set of 2D feature maps $\{\mathbf{F}_{2D}^{(m)}\} \subset R^{H\times W\times C}$. To obtain a top-down geometric layout, we define a set of 3D anchors $\mathbf{A} \subset R^{U\times V\times Z}$, distributed uniformly within a local 3D region around the agent.

Each anchor point $\mathbf{a}_{uvw}$ is projected to each image view using known extrinsic and intrinsic camera parameters, and queried via cross-view attention:

$$\mathbf{F}_{3D}[u,v,z] = \sum_{m=1}^{M} \text{Attn}\left(\mathbf{a}_{uvw}, \mathbf{F}_{2D}^{(m)}\right), \tag{1}$$

where Attn denotes a differentiable sampling-and-weighting operator. The 3D tensor $\mathbf{F}_{3D} \in R^{U\times V\times Z\times C}$ is collapsed along the height axis to form a 2D BEV feature map:

$$\mathbf{F}_{BEV}[u,v] = \text{Pool}_z\left(\mathbf{F}_{3D}[u,v,:]\right). \tag{2}$$

B. Constructing the Spatial-Language Scene Graph

To reason over spatial context and instruction grounding, we incrementally build a spatio-linguistic graph $\mathcal{S}_t = (\mathcal{N}_t, \mathcal{E}_t)$ over time. Each node $n_i \in \mathcal{N}_t$ corresponds to a navigable location, and is associated with an embedding vector $\mathbf{h}_i \in R^C$ computed from a local neighborhood $\mathcal{B}_i$ in the BEV map:

$$\mathbf{h}_i = \frac{1}{|\mathcal{B}_i|} \sum_{(u,v)\in\mathcal{B}_i} \mathbf{F}_{BEV}[u,v]. \tag{3}$$

Graph edges encode relative reachability between locations based on navigation topology. To ensure temporal consistency, we align overlapping regions of BEV maps between steps t and $t+1$ and update node features via local fusion:

$$\tilde{\mathbf{F}}_{BEV}^{(t+1)}[u,v] = \text{Fusion}\left(\mathbf{F}_{BEV}^{(t)}[u,v], \mathbf{F}_{BEV}^{(t+1)}[u,v]\right), \quad (u,v)\in\mathcal{O}, \tag{4}$$

where $\mathcal{O}$ is the overlapping region, and Fusion denotes a feature-level update cross-attention function [9].

C. Instruction-Guided Decision Scoring

The agent receives an instruction sequence $\mathcal{L} = \{\ell_1,\ldots,\ell_T\}$, which is embedded via a text encoder into $\mathbf{E}_{\mathcal{L}} \in R^{T\times C}$. We perform hierarchical decision scoring based on both the global scene graph and the local BEV layout.

1) Node-Level Scoring: We apply a multi-head cross-modal transformer that takes node features $\{\mathbf{h}_i\}$ and language embeddings $\mathbf{E}_{\mathcal{L}}$ as input, producing updated node embeddings $\{\tilde{\mathbf{h}}_i\}$:

$$\tilde{\mathbf{h}}_i = \text{Transformer}_{\text{global}}(\mathbf{h}_i, \mathbf{E}_{\mathcal{L}}). \tag{5}$$

These are passed through a lightweight MLP to generate global decision scores $s_i^{(g)}$ over all candidate nodes:

$$s_i^{(g)} = \text{MLP}_g(\tilde{\mathbf{h}}_i). \tag{6}$$

2) Grid-Level Scoring: Similarly, the BEV grid features $\mathbf{F}_{BEV}$ are fused with the instruction via a transformer module:

$$\tilde{\mathbf{F}}_{BEV}[u,v] = \text{Transformer}_{\text{local}}(\mathbf{F}_{BEV}[u,v], \mathbf{E}_{\mathcal{L}}), \tag{7}$$

producing a grid-wise score map $s^{(l)}[u,v]$:

$$s^{(l)}[u,v] = \text{MLP}_l(\tilde{\mathbf{F}}_{BEV}[u,v]). \tag{8}$$

For each navigable candidate location n_k, we define its neighborhood $\mathcal{B}_k$ and compute a pooled local score using a Gaussian-weighted average:

$$s_k^{(l)} = \sum_{(u,v)\in\mathcal{B}_k} \mathcal{G}_\sigma(\Delta x_{uv}, \Delta y_{uv}) \cdot s^{(l)}[u,v], \tag{9}$$

where $(\Delta x_{uv}, \Delta y_{uv})$ are relative displacements from the grid to the node center, and $\mathcal{G}_\sigma$ is a normalized bivariate Gaussian kernel.

3) Final Decision and Action: The final score for each candidate location is a convex combination of global and local scores:

$$s_k = \alpha \cdot s_k^{(l)} + (1-\alpha) \cdot s_k^{(g)}, \tag{10}$$

where $\alpha \in [0,1]$ is a tunable weighting factor. The agent selects the action corresponding to $\arg\max_k s_k$.

D. Geometric Supervision via 3D Detection

To enhance BEV features with object-level geometry, we train a 3D object detector using simulated construction data with oriented bounding box annotations. The detector outputs $\{\hat{\mathbf{o}}_j\}$ (object category, center, and orientation) from BEV features and is supervised using a combination of classification and regression losses:

$$\mathcal{L}_{\text{det}} = \lambda_1 \cdot \mathcal{L}_{\text{cls}} + \lambda_2 \cdot \mathcal{L}_{\text{reg}}, \tag{11}$$

where $\mathcal{L}_{\text{reg}}$ includes both position and rotation terms. This supervision enables the agent to recognize spatial layouts (e.g., cranes, barricades, trailers) and supports instruction grounding in occluded or partially visible conditions.

E. Dataset

To facilitate the development and evaluation of our vision-language navigation agent, we created a realistic simulated construction site environment using NVIDIA Isaac Sim [10]. Our simulated environment, "SynthConstruct", comprises 15 master construction sites, each representing a unique phase of construction with varying structural elements. The sites vary in size from 500 to 2,300 square meters, totaling a traversable space of over 120 Km² for robotic navigation within the simulator. SynthConstruct includes a vast array of construction-related objects to ensure a comprehensive representation of real-world construction elements.

III. Results and Discussion

We evaluate our proposed BEV-based vision-language navigation framework (BEV-VLN) on our simulated 3D construction environment, focusing on tasks that require complex spatial reasoning, occlusion awareness, and instruction grounding. Our model is compared against established baselines that rely on panoramic or egocentric observations and lack explicit geometric reasoning.

A. Quantitative Evaluation

1) Evaluation Metrics: Following standard practice in embodied navigation, we report the following metrics:

- **Success Rate (SR)**: Percentage of episodes where the agent reaches the goal within a given threshold.
- **Success weighted by Path Length (SPL)**: SR weighted by the efficiency of the agent's path.
- **Navigation Error (NE)**: Final distance from the agent's stopping point to the goal.
- **Object Grounding Accuracy (OGA)**: Success in identifying and aligning with referenced objects (e.g., "stand beside the orange barrel").

2) Performance Comparison: Table I presents the performance of our BEV-VLN model against three baselines: (1) a language-conditioned panoramic navigation model, (2) an egocentric model with local mapping, and (3) a transformer-based navigation agent without explicit geometry.

TABLE I
NAVIGATION PERFORMANCE ON 3D CONSTRUCTION SIMULATION TASKS.

Model	SR (%)	SPL (%)	NE (m)	OGA (%)
Panoramic VLN Baseline	54.3	42.1	2.93	48.5
Transformer w/o BEV	58.7	47.4	2.61	52.9
Egocentric + Local Map	61.5	50.2	2.35	56.1
BEV-VLN (Ours)	**68.9**	**58.4**	**1.79**	**63.7**

a) Improved Geometric Reasoning.: Our model significantly reduces the average navigation error (1.79m), demonstrating better spatial localization. The inclusion of oriented 3D object detection enhances the agent's awareness of occluded objects, such as machinery behind temporary barriers or scaffolding.

b) Enhanced Instruction Grounding.: BEV-VLN outperforms all baselines in object grounding accuracy (+7.6% over the next best model), confirming its ability to align linguistic references with visual cues. For instance, instructions referencing spatial relationships like "walk past the concrete mixer and stop at the steel beam on your left" are resolved more accurately due to the fine-grained BEV grid attention and object detection supervision.

B. Ablation Study

As illustrated in Table II, we conduct ablation experiments to explore effect of key components:

- Removing the 3D detection head reduces SR by 5.4%, highlighting the role of geometric supervision.
- Disabling BEV temporal alignment leads to drift and lower SPL, confirming the importance of spatial memory.
- Replacing Gaussian pooling with uniform averaging in grid-to-node scoring degrades OGA by 3.2%.

C. Discussion

These results demonstrate that BEV-based spatial reasoning, when fused with language grounding, significantly enhances navigation robustness in complex, cluttered, and

TABLE II

ABLATION STUDY ON THE IMPACT OF DIFFERENT COMPONENTS IN THE BEV-VLN MODEL.

Model Variant	SR (%) ↑	SPL (%) ↑	NE (m) ↓	OGA (%) ↑
Full Model (BEV-VLN)	**68.9**	**58.4**	**1.79**	**63.7**
w/o 3D Object Detection	63.5	52.1	2.41	56.4
w/o Temporal BEV Fusion	61.2	49.8	2.53	54.9
w/o Scene Graph Reasoning	60.3	46.5	2.74	52.6
w/o Grid-Level Scoring	57.8	44.7	2.89	49.3
w/ Uniform Grid Pooling	65.0	53.6	2.18	60.5

occluded environments such as construction sites. The explicit encoding of object geometry allows the agent to disambiguate between visually similar but spatially distinct targets. Moreover, the dual-level scoring strategy ensures that both fine-grained and global context are considered during action selection.

We believe this method lays the foundation for safer and more interpretable navigation agents in real-world industrial and field robotics applications.

IV. CONCLUSION

In this paper, we presented a novel BEV-based vision-language navigation system specifically designed for simulated 3D construction environments. The framework integrates spatially aligned BEV features, a dynamically updated scene graph, and a dual-level instruction-conditioned decision module to enable precise and context-aware navigation. By incorporating object-level 3D detection, the agent gains the ability to perceive occluded elements and better interpret instruction-grounded spatial relationships. Extensive experiments demonstrate that our method outperforms traditional panoramic and egocentric baselines in terms of navigation accuracy, efficiency, and object grounding performance. The results underscore the effectiveness of combining top-down geometric representations with language-driven reasoning in complex, real-world-inspired domains. This approach opens new opportunities for deploying intelligent embodied agents in safety-critical environments such as construction, manufacturing, and industrial inspection, where spatial understanding and semantic comprehension are both essential. As these technologies mature, they could democratize robotic automation for small-scale builders, improve inspection and maintenance in hazardous settings, and pave the way for fully autonomous construction sites, transforming how infrastructure is built worldwide. Future work may explore generalization to outdoor construction scenarios and real-world transfer via sim-to-real adaptation.

REFERENCES

[1] S. Karimi, R. G. Braga, I. Iordanova, and D. St-Onge, "Semantic navigation using building information on construction sites," *ArXiv*, vol. abs/2104.10296, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233324236

[2] Z. Xu, H. Jin, X. Han, H. Shen, and K. Shimada, "Intent prediction-driven model predictive control for uav planning and navigation in dynamic environments," *IEEE Robotics and Automation Letters*, vol. 10, pp. 4946–4953, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272831797

[3] H. Oyediran, W. Turner, K. Kim, and M. Barrows, "Integration of 4d bim and robot task planning: Creation and flow of construction-related information for action-level simulation of indoor wall frame installation," *ArXiv*, vol. abs/2402.03602, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267499992

[4] K. Dörfler, G. Dielemans, S. Leutenegger, S. E. Jenny, J. Pankert, J. Sustarevas, L. Lachmayer, A. Raatz, and D. Lowke, "Advancing construction in existing contexts: Prospects and barriers of 3d printing with mobile robots for building maintenance and repair," *Cement and Concrete Research*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272661559

[5] S. Halder and K. Afsari, "Robots in inspection and monitoring of buildings and infrastructure: A systematic review," *Applied Sciences*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256866763

[6] J. Zhang, K. Wang, R. Xu, G. Zhou, Y. Hong, X. Fang, Q. Wu, Z. Zhang, and W. He, "Navid: Video-based vlm plans the next step for vision-and-language navigation," *ArXiv*, vol. abs/2402.15852, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:267938569

[7] X. Liang, L. Ma, S. Guo, J. Han, H. Xu, S. Ma, and X. Liang, "Cornav: Autonomous agent with self-corrected planning for zero-shot vision-and-language navigation," in *Annual Meeting of the Association for Computational Linguistics*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259202797

[8] L. Yue, D. Zhou, L. Xie, F. Zhang, Y. Yan, and E. Yin, "Safe-vln: Collision avoidance for vision-and-language navigation of autonomous robots operating in continuous environments," *IEEE Robotics and Automation Letters*, vol. 9, pp. 4918–4925, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:265033616

[9] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.

[10] NVIDIA, "Iaac sim: robotics simulation and synthetic data genera- tion," [Accessed March, 8, 2025]. [Online]. Available: https://developer.nvidia.com/isaac/sim