Embodied AI in Unstructured 3D Spaces: Fusing Mid- and Long-Range Sensing for Instruction-Aware Construction Robotics

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

***Abstract*— Robust navigation in 3D construction environments requires an embodied agent to perceive and reason across diverse spatial and dynamic cues. In this paper, we present a multisensory navigation framework for simulated construction sites using an object-centric embodied agent equipped with LiDAR, depth, IMU, and proximity sensors. Built in NVIDIA Isaac Sim, our system fuses geometric and inertial information to interpret high-level instructions and execute efficient, collision-free trajectories. We encode construction environments using abstracted object-level scene representations and leverage an instruction-tuned language model to iteratively generate action sequences. Experimental evaluations demonstrate that our agent significantly outperforms visual-only and traditional LiDAR-based navigation baselines in goal success rate, path efficiency, collision avoidance, and trajectory smoothness. Through ablation studies, we validate the complementary roles of each sensor modality in supporting long-range planning, local maneuvering, and safe obstacle avoidance. This work highlights the importance of multisensor integration for enabling intelligent, context-aware navigation in safety-critical construction scenarios.**

I. INTRODUCTION

Embodied AI represents an emerging field that focuses on creating intelligent agents capable of perceiving, navigating, and operating within three-dimensional environments [1]. At the core of embodied AI research is the development of navigation capabilities, where agents learn to perform tasks in simulated environments using raw pixel data as input [2]. These tasks span a spectrum of complexity, from target-driven navigation, where agents must find specific locations [3], to instruction-based visual navigation, where agents follow natural language directions [4], and even embodied question answering (EQA), where agents must explore an environment to find information needed to answer questions [5].

Construction sites present unique challenges for navigation due to their unstructured and dynamic nature. Objects ranging from tall structures (ladders, columns) to low-lying items (tools, material piles) are distributed unpredictably, creating a much more complex environment than structured spaces like warehouses or factories [6]. To address these challenges, researchers have employed LiDAR Odometry and Mapping (LOAM) methods to generate comprehensive 3D point clouds of construction sites, which can then be projected into 2D grid maps for efficient navigation [6], [7]. Although. multisensory perception represents a critical advancement in embodied AI navigation, as it allows agents to develop richer representations of their 3D environments beyond what can be achieved with visual data alone. Despite its importance, multisensory 3D scene representation learning has historically received less attention compared to unimodal approaches [8].

Recent research has explored even more innovative sensory combinations, such as using echoes with RGB imagery to estimate wide field-of-view depth information for 3D navigation. The PointGoal echo navigation approach directly leverages echo data to perceive spatial cues, using an echo encoder that maps binaural echoes into vector representations which are then processed alongside GPS signals to determine agent actions [9]. For construction site navigation specifically, multimodal systems often integrate camera imagery with LiDAR sensing. This combination provides complementary information where the camera captures visual data as pixel images while strategically placed LiDAR sensors provide precise depth measurements, enabling agents to perceive their environment in three dimensions [10]. However, most existing embodied navigation frameworks either rely solely on visual inputs or lack integration of diverse spatial sensors suited for complex, dynamic environments like construction sites. There remains a critical need for instruction-driven agents that can fuse mid- and long-range sensing modalities to reason about structure, avoid obstacles, and adaptively plan in unstructured 3D spaces [11], [3].

The key contributions of this research include:

- We propose an object-centric embodied navigation system that integrates LiDAR, depth, IMU, and proximity sensors to enable efficient and safe instruction-driven navigation in cluttered 3D construction environments using Isaac Sim.
- We develop a closed-loop control pipeline where a LLM interprets high-level tasks and sequentially generates navigation actions by reasoning over fused multisensory representations and dynamic observations.
- We conduct comprehensive experiments showing significant improvements in goal success rate, path efficiency, and collision avoidance over strong baselines.

II. METHODOLOGY

In this section, we illustrate our framework for multisensory navigation within 3D simulated construction environments, which leverages an object-centric embodied

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 mbonya1@lsu.edu msoley1@lsu.edu

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

agent architecture. Our method integrates visual, LiDAR, IMU, and Proximity sensors into a coherent decision-making pipeline, guided by an instruction-tuned Large Language Model (LLM). The framework enables an agent to perceive, explore, and reason about complex scenes in a goal-driven manner. The method consists of four primary components: (1) scene encoding using object-centric representations, (2) multimodal action design, (3) multisensory feedback integration, and (4) LLM-based policy generation for embodied interaction.

A. Object-Centric Scene Representation

In our simulated 3D construction environment, we begin by abstracting the scene into a structured set of object-centric representations. The agent initially performs a panoramic visual sweep, capturing RGB-D frames from multiple viewpoints. From these views, object proposals are extracted and encoded using a pretrained CLIP-based vision encoder. Object-level features are then fused across views using multi-view association, and spatial localization is achieved through positional embeddings. Each object is annotated with its semantic label, 3D bounding box, estimated material type, and contextual metadata (e.g., "scaffold joint", "metal pipe", "heated surface"), forming a compact scene graph.

In parallel, environmental proximity cues are collected via simulated microphones and processed with a CLAP-based proximity encoder. These features are associated with emitting objects and appended to the scene graph. The result is a multimodal, object-centric embedding of the construction site that preserves both global structure and local detail relevant to navigation and safety inspection.

B. Embodied Interaction via Action Tokens

To facilitate active exploration, we define a set of discrete action tokens that parameterize the agent's interactions with the environment. These include:

- `<SELECT>`: Choose a target object based on linguistic and sensory context.
- `<NAVIGATE>`: Move towards the selected object using a geometric path planner.
- `<OBSERVE>`: Capture close-range RGB-D data and update the object's visual representation.
- `<LiDAR>`: Simulate 3D environment for collecting surface deformation and object data.
- `<Proximity>`: Apply a distance force to generate impact distance used for safety inference.
- `<IMU>`: capture motion of robots for rearrangement or functional interaction.
- `<LOOK-AROUND>`: Acquire contextual information about nearby objects in the scene.

These tokens enable compositional behaviors that mimic real-world inspection procedures, such as identifying loose fittings, detecting overheated machinery, or distinguishing material types under visual ambiguity.

C. Multisensory State Feedback Encoding

Following each interaction, the sensory outcomes are encoded and appended to the agent's perceptual state via specialized state tokens. IMU data are simulated using a differentiable motion robot and transformed into 2D heatmaps. These are projected into the LLM's embedding space via a dedicated IMU adapter trained for modality alignment. Similarly, distance readings are visualized as heatmap dsitance patterns and processed using a distance adapter. Impact distance captured by the `<Proximity>` token is encoded using a CLAP. Each of these encoded modalities is structured into tokenized feedback blocks (e.g., `<LiDAR>...</LiDAR>`, `<IMU>...</IMU>`) and fed into the LLM, enabling it to reason over time-varying, multimodal sensor states.

D. LLM-Based Instruction and Policy Generation

At the heart of the framework lies a multimodal instruction-tuned LLM, adapted from the LLaVA architecture. The model is trained to generate context-aware action tokens conditioned on the current task prompt and the evolving multisensory state. During training, the LLM is optimized using a hybrid loss: (1) standard autoregressive cross-entropy loss for action and language token prediction, and (2) a contrastive attention-based loss to ensure correct object selection during `<SELECT>` actions.

Our LLM is tuned on our simulated dataset of construction-specific multisensory interactions generated in simulation. Each datapoint includes a prompt (e.g., "Find the red pipe near the scaffolding"), scene representation, a sequence of actions and states, and reasoning explanations. The LLM is trained in a recurrent fashion, where each newly acquired sensory token informs the next step of policy generation.

During inference, the agent executes in a closed-loop manner. Upon receiving a prompt and scene input, the LLM generates an initial action. The agent then performs the action in the 3D simulator, retrieves the resulting sensory data, and updates the input state. This iterative process continues until a task-specific completion signal is detected (e.g., successful object retrieval or hazard localization).

E. Simulation Platform and Integration

Our framework is deployed in a high-fidelity construction simulation environment built on top of Isaac Sim and extended with construction-relevant assets. Interactive objects are augmented with physically realistic materials, distance profiles, and IMU properties. The embodied agent is equipped with simulated RGB-D sensors, LiDAR, and IMU, enabling rich, physics-based interaction with the environment. Overall, the proposed methodology bridges object-level abstraction, rich sensory feedback, and large-scale language-based reasoning to enable intelligent navigation and inspection within simulated construction sites.

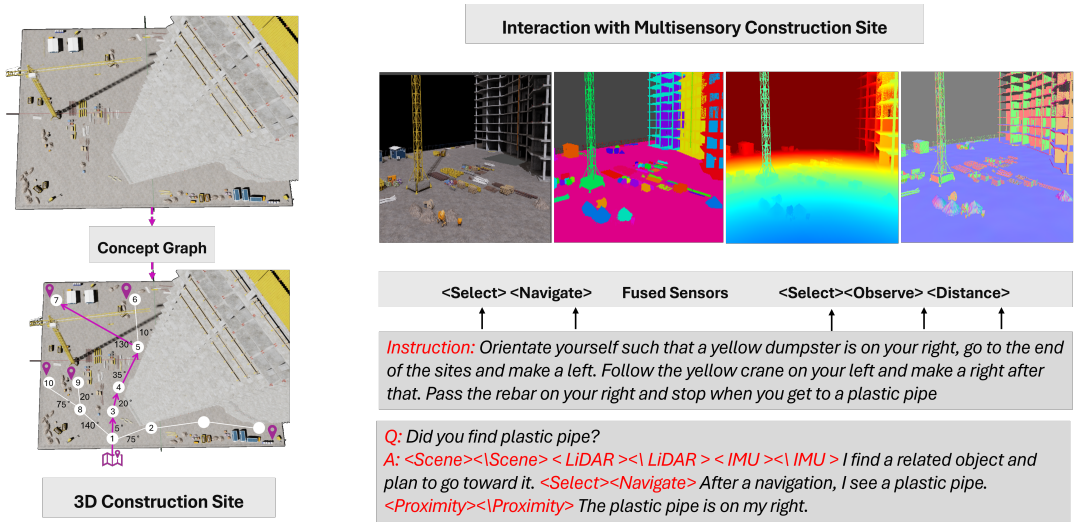

Fig. 1. An overview of the proposed method. The scene is initially represented in a graph concept, object-focused format, with detailed multisensory object properties only becoming accessible when the agent actively engages with them through interactions. We've created a collection of action tokens that represent the different ways agents can interact with the environment. When interactions occur, their outcomes are communicated back to the Large Language Model (LLM) through state tokens.

F. 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 [12]. Our simulated environment, which we call 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 multisensory navigation framework in 3D simulated construction environments using the NVIDIA Isaac Sim platform. The agent is equipped with LiDAR, depth sensing, IMU, and proximity sensors, and operates under diverse conditions involving clutter, occlusions, and partial observability. Our experiments focus on assessing the agent's ability to interpret navigation goals from high-level instructions and execute safe, efficient paths in complex, object-dense spaces. The evaluation benchmarks four key metrics: (1) **Goal Success Rate** (GSR), (2) **Path Efficiency** (PE), (3) **Collision Rate** (CR), and (4) **Trajectory Smoothness** (TS).

A. Quantitative Evaluation

Table I summarizes the performance of our agent compared to baseline navigation strategies, including a visual-only agent (VO), a classical LiDAR-based path planner (CLP), and our ablated model without IMU-proximity integration (Ours–IMU+Prox).

TABLE I

Navigation performance comparison across 100 episodes in diverse construction scenes.

Model	GSR (%)	PE (%) ↑	CR (%) ↓	TS (m/s²) ↓
VO (RGB-D Only)	61.3	71.4	18.7	0.93
CLP (LiDAR Planner)	69.8	76.2	12.3	0.88
Ours–IMU+Prox (Ablated)	74.5	82.6	10.5	0.79
Ours (Full)	**85.7**	**90.1**	**6.2**	**0.65**

Our full model achieves a Goal Success Rate (GSR) of 85.7%, a significant improvement over traditional LiDAR-only and RGB-D models. The Path Efficiency (PE), computed as the ratio of optimal to actual path length, indicates that our model generates near-optimal paths while reducing redundant or oscillatory motion. Notably, the Collision Rate (CR) is reduced by over 50% compared to the RGB-D baseline, attributed to the use of short-range proximity data for reactive obstacle avoidance and the IMU data for dynamic stability during movement. Trajectory Smoothness (TS), calculated as the average linear acceleration along the path, reflects the agent's ability to maintain stable motion through uneven or cluttered spaces.

B. Ablation Study

To assess the individual contribution of each sensing modality, we conducted ablation experiments where one sensor was removed at a time. Table II summarizes the impact of each removal on key navigation metrics.

TABLE II

Ablation study: effect of sensor modalities on navigation metrics.

Configuration	GSR (%)	PE (%)	CR (%)	TS (m/s²)
Full (LiDAR + Depth + IMU + Prox)	**85.7**	**90.1**	**6.2**	**0.65**
w/o Proximity	78.6	84.9	11.8	0.79
w/o IMU	80.2	86.1	9.5	0.91
w/o LiDAR	68.9	74.3	13.7	0.94

Removing LiDAR resulted in the most significant degra-

dation in both path efficiency and goal success rate. This is expected, as LiDAR provides long-range geometric awareness of the 3D structure of the environment, enabling reliable global path planning and spatial localization. In construction sites, where large structural elements such as scaffolds, columns, and equipment can occlude vision, LiDAR ensures the agent can navigate around large-scale obstructions and maintain situational awareness.

While not individually ablated in this table due to fusion with LiDAR in our base setup, depth data complements LiDAR by capturing dense near-field spatial information with rich resolution. It is especially useful for detecting floor-level hazards (e.g., cables, tools) and navigating through tight spaces or thresholds. Removing the IMU increased trajectory instability, as reflected by the increase in average acceleration (TS) and reduced success rate. IMU data allows the agent to infer its motion state (velocity and orientation) in real-time, which is crucial for maintaining balance and executing smooth navigation—especially on uneven terrain, ramps, or during sudden changes in direction, all of which are common in dynamic construction environments.

The absence of the proximity sensor caused a notable rise in collision rate. Proximity sensing acts as a safety buffer, enabling reactive avoidance of objects that are too close to the agent—such as workers, hanging cables, or handheld tools—where LiDAR's resolution or depth perception may be insufficient due to occlusion or sensor blind spots. Its inclusion is critical for ensuring the agent can operate safely in dense or human-populated areas. Overall, each sensor modality supports a unique and complementary function in enhancing navigation robustness. The ablation study validates that construction sites demand a fusion of long-range structural awareness, local obstacle detection, motion state estimation, and close-range safety assurance for effective and safe autonomous navigation.

C. Limitations and Future Work

While our framework performs robustly in simulated settings, deployment to real construction robots would require calibration of sensor fusion models under real-world noise and drift. Additionally, current policies are pre-trained and do not adapt online; incorporating reinforcement learning or model-predictive control (MPC) would improve adaptability in highly dynamic or partially mapped environments. Finally, introducing memory-based models could enhance long-horizon task planning across multiple goals. Also, the results demonstrate the advantages of combining LiDAR, depth, IMU, and proximity sensing in a structured, object-aware policy framework. Our approach enables robust, safe, and efficient navigation across cluttered, unstructured construction environments, offering a practical path toward deployable autonomous site robots.

IV. Conclusion

We presented a multisensory object-centric framework for autonomous navigation in simulated 3D construction sites. Leveraging LiDAR, depth, IMU, and proximity sensors, our embodied agent operates under the supervision of an instruction-tuned language model that interprets spatial semantics and generates action plans. The fusion of long-range and short-range sensing enables the agent to navigate safely and efficiently in cluttered, partially observable environments common in real-world construction. Our evaluations in Isaac Sim show that the proposed method achieves higher goal success rates, better path efficiency, and significantly reduced collisions compared to RGB-D and classical LiDAR-based baselines. Future work will focus on transferring the framework to real hardware, incorporating online learning and memory for persistent multi-goal planning, and integrating with dynamic scene understanding for operation in changing construction environments.

References

[1] F. Landi, R. Bigazzi, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, "Spot the difference: A novel task for embodied agents in changing environments," *2022 26th International Conference on Pattern Recognition (ICPR)*, pp. 4182–4188, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248239837

[2] M. Hahn, K. T. Carlberg, R. Desai, and J. M. Hillis, "Learning a visually grounded memory assistant," *ArXiv*, vol. abs/2210.03787, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247225659

[3] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. K. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3357–3364, 2016. [Online]. Available: https://api.semanticscholar.org/CorpusID:2305273

[4] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. D. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3674–3683, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:4673790

[5] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2135–213 509, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:35985986

[6] Y. Wu, J. Wei, J. Oh, and D. C. Llach, "Towards human-centered construction robotics: A reinforcement learning-driven companion robot for contextually assisting carpentry workers," *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 90–97, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268733331

[7] T. Shan, B. Englot, D. Meyers, W. Wang, C. Ratti, and D. Rus, "Lio-sam: Tightly-coupled lidar inertial odometry via smoothing and mapping," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5135–5142, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220280471

[8] J. H. Lim, P. H. O. Pinheiro, N. Rostamzadeh, C. J. Pal, and S. Ahn, "Neural multisensory scene inference," *ArXiv*, vol. abs/1910.02344, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202766460

[9] L. Zhu, E. Rahtu, and H. Zhao, "Beyond visual field of view: Perceiving 3d environment with echoes and vision," *ArXiv*, vol. abs/2207.01136, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250264349

[10] A. Gerstenslager, J. Lewis, L. McKenna, and P. Patel, "Autonomous navigation in complex environments," *ArXiv*, vol. abs/2401.03267, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:266844595

[11] H. Wang, W. Liang, L. V. Gool, and W. Wang, "Towards versatile embodied navigation," *ArXiv*, vol. abs/2210.16822, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253237724

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