EMKG: Embodied Multimodal Knowledge Graph Integration for Open-World Object-Goal Navigation in Embodied Systems

Abstract— **This paper addresses the challenge of 3D spatial understanding and navigation using quadruped robots, focusing on integrating multimodal Reasoning-Action Graphs (RAG) for spatial dependency-based navigation strategies. Unlike traditional methods that rely heavily on vision and language input for direct action predictions,our approach builds a knowledge base graph that connects multimodal information to improve the robot's spatial reasoning. We propose the Embodied Multimodal Knowledge Graph Vision-Language Model in Embedded System (EMKG-VLM), which integrates visual encoders, language models, and knowledge graph networks to enhance multimodal reasoning and decision making. EMKG-VLM achieves state-of-the-art performance in both simulated and real-world navigation tasks, showing a 15.4% improvement in fine-grained retrieval accuracy and a 12.1% increase in mean average precision (mAP) in zero-shot and few-shot scenarios. The framework emphasizes the integration of multimodal knowledge and dynamic spatial reasoning, showing superior performance in tasks that require sophisticated spatial understanding and cross-modal retrieval. This highlights EMKG-VLM's proficiency in 3D scene comprehension and multimodal data utilization for navigation. The deployment of EMKG-VLM on embedded platforms with Jetson Orin further showcases its potential for real-time reasoning and practical applications in embodied intelligent systems. By continuously updating the knowledge graph with feedback from the Vision-Language Model (VLM), EMKG-VLM adapts to environmental changes in real-time, significantly enhancing the navigation and decision-making capabilities of intelligent robots in dynamic environments. The code and models are included in the supplementary materials to ensure anonymity.**

I. Introduction

Object Goal Navigation (ObjectNav) tasks aim to enable robots to locate and interact with objects in unexplored environments [1]. Existing methodologies frequently depend on metric maps that amalgamate visual and semantic data to represent free space. These approaches are susceptible to errors in open-world scenarios, especially when precise localization is critical. Misalignment in pose estimation compromises the accuracy of the map, affecting global navigation and action planning [2]. This issue is particularly problematic in indoor environments, where GPS and compass sensors are unavailable, leading to repeated estimation errors [3]. Although end-to-end methods bypass metric maps, they struggle to effectively represent global structure, thereby limiting their efficiency when the target is far from the agent's starting point [4]. Recent works have attempted to enhance navigation reasoning by integrating vision-language models (VLM) and language models (LLMs). Mobility VLA [5] employs long-context VLMs with topological graphs, while Open Scene Graphs (OSG) [6] improve LLM-based spatial reasoning for natural language-based target search. PixNav [7] utilizes LLM-based commonsense reasoning to refine waypoint selection. However, these approaches are limited by inference latency and long-term spatial consistency.

To address these limitations, we propose the Embodied Multimodal Knowledge Graph Vision-Language model (EMKG-VLM). This model offers an end-to-end approach for multimodal retrieval and navigation in 3D environments, specifically tailored for embodied intelligence. EMKG-VLM facilitates real-time navigation in unknown environments by dynamically updating a multimodal knowledge vector database and fine-tuning an embedded MobileVLM model.

The principal contributions of this work are as follows.

- **High-Precision Multimodal Retrieval with Dynamic Knowledge Graphs:** We introduce an EMKG system that employs Dynamic Graph Neural Networks (DGNN) to improve the multimodal retrieval accuracy and mean Average Precision (mAP), achieving a 12.3% improvement in retrieval accuracy and a 4.92% increase in mAP across various 3D scene datasets. This system demonstrates effective generalization in zero-shot tasks within complex 3D environments.
- **EMKG-VLM Framework for Object Goal Navigation:** The EMKG-VLM framework enhances object goal navigation by integrating real-world data into the Habitat simulator, thus improving decision-making and retrieval precision in unknown environments and exhibiting high adaptability in simulated navigation.
- **Real-Time Autonomous Actions through Optimized Vision-Language Commands and Reinforcement Learning:** We combine mobile-optimized InternLM2 visual-language commands with PPO-Clip reinforcement learning to enable real-time decision making and autonomous navigation in dynamic open-world environments, thus improving adaptability and effectiveness for real-world applications.

II. Related Work

A. Object Goal Navigation

Object Goal Navigation (ObjectNav) enables robots to locate and interact with objects in unexplored environments [1]. Recent advancements integrate vision-language models (VLMs) and language models (LLMs) to enhance navigation reasoning. Mobility VLA [5] employs long-context VLMs with topological graphs, while Open Scene Graphs (OSG) [6] improve LLM-based spatial reasoning for natural language target search. PixNav [7] refines waypoint selection using common sense reasoning. However, these approaches

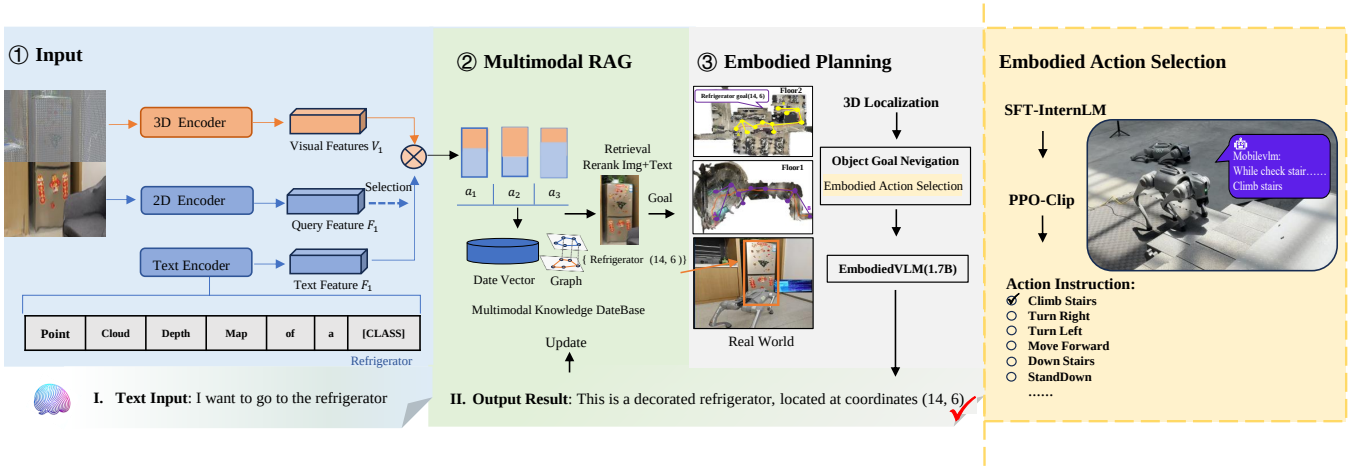

Fig. 1. The EMKG-VLM system integrates 3D point cloud processing with multimodal knowledge graphs and a vision-language model, enabling autonomous robotic navigation and task execution through real-time decision-making and spatial reasoning. The system processes input data, generates an embodied action plan, and selects actions for real-world interaction.

face challenges in inference latency and long-term spatial consistency. EMKG-VLM addresses these by integrating a structured multimodal knowledge graph with real-time visual-language retrieval, enabling efficient path planning and adaptive navigation strategies.

B. Knowledge-Based Visual Question Answering

Knowledge-driven Visual Question Answering (VQA) systems, such as KAT [8] and REVIVE [9], leverage external knowledge for complex query reasoning [10]. Retrieval-Augmented Generation (RAG) further enhances zero/few-shot learning by dynamically retrieving relevant content [11]. However, these methods rely on static textual augmentation and cannot dynamically update knowledge representations in response to environmental changes. EMKG-VLM incorporates an Embodied knowledge graph update mechanism, allowing real-time multimodal scene refinement in robotic navigation.

C. 3D Scene Understanding and Multimodal Retrieval

Aligning 3D scene language data remains a challenge. Datasets like ShapeNet and Objaverse focus on object-centric understanding [12], while scene-level datasets extend toward vision-language navigation (VLN) [13]. 3D scene graphs (3DSGs) [14] provide hierarchical spatial-semantic representations but lack adaptability in dynamic environments. The complexity of 3D scenes, including intricate spatial arrangements and inter-object relationships, complicates effective knowledge graph construction [15], [16].

Multimodal RAG systems primarily integrate text and visual data, but effective utilization of 3D scene data for indoor retrieval remains a challenge. The reliance on 2D datasets limits scalability for 3D environments [17], and existing 3D datasets remain small compared to their 2D counterparts [18], [19], hindering deep learning applications [20]. Additionally, current knowledge graphs primarily process 2D data and struggle with integrating multimodal 3D spatial relationships [21], limiting their performance in complex tasks [22].

EMKG-VLM enhances 3D multimodal retrieval via contrastive knowledge alignment, enabling real-time 2D-3D mapping for object-based reasoning. It also mitigates hallucinations in VLMs by incorporating structured multimodal knowledge graphs to refine object relations and reduce noise [23].

III. Methodology

The Embodied Multimodal Knowledge Graph Vision-Language Model (EMKG-VLM) is designed for autonomous robot navigation in complex environments. The system consists of four key modules: the High-Level Cognition Module for environment perception and goal interpretation, the Dynamic Multimodal Knowledge Graph (DMKG) Module for constructing relations between the environment and textual inputs, the Mobile Vision-Language Model (MobileVLMv2) for real-time updates and dynamic graph refinement, and the Low-Level Control Module for navigation execution. EMKG-VLM integrates multimodal retrieval and vision-language processing to provide continuous feedback for Object Goal Navigation (ObjectNav) while establishing hierarchical object relationships to facilitate scene-aware actions.

A. Multimodal Knowledge Graph Module

This module processes multi-modal data to construct a structured scene representation. Given a textual input Q (e.g. navigation instructions or object descriptions), a large-scale pre-trained language model extracts textual features F_t:

$$F_t = \text{LanguageEncoder}(Q), \tag{1}$$

To enhance the integration of 2D and 3D spatial features, we extend Efficient Point Cloud Learning (EPCL). Farthest Point Sampling (FPS) selects M center points, and K-Nearest Neighbors (KNN) groups K surrounding points into

patches. These patches are processed through a Multi-Layer Perceptron (MLP) to generate query embeddings:

$$Q_p = \mathbf{Q}(P) \in R^{M \times D_q} \tag{2}$$

Here, Q_p represents the query embeddings derived from the point cloud, where $\mathbf{Q}(P)$ is the transformation applied to the point cloud P, and $R^{M \times D_q}$ indicates that the resulting embeddings are real-valued matrices with dimensions M and D_q.

To align the 3D point cloud features with the pre-trained language model, we introduce a learnable task query, which passes through a fully connected layer (FC) and is then concatenated with the point cloud query embeddings Q_p before being fed into the frozen CLIP Transformer:

$$Z_0 = [Q_{\text{task}}, Q_p] + E_{\text{pos}}. \tag{3}$$

In this equation, Q_{task} represents the learnable task query, Q_p corresponds to the query embeddings derived from the point cloud, and E_{pos} denotes the positional encodings, which are added to retain spatial information.

The fusion of multimodal features takes place through Transformer layers:

$$Z_l = \mathbf{F}(Z_{l-1}) + Z_{l-1}. \tag{4}$$

Here, $\mathbf{F}$ represents the operations within the Transformer layers (including self-attention and feedforward networks), while the residual connection $+Z_{l-1}$ ensures efficient information flow across layers.

Finally, to achieve cross-modal alignment, we computed the cosine similarity to measure the semantic correlation between the 2D image features and the 3D point cloud query embeddings:

$$\text{cosine similarity}(Q_{\text{image}}, Q_{\text{point cloud}}). \tag{5}$$

To achieve cross-modal alignment, cosine similarity is used to measure semantic correlations between the 2D image and 3D point cloud features.

$$\text{corr}\left(F_{\text{img}}, F_{\text{pc}}\right) = \frac{F_{\text{img}} \cdot F_{\text{pc}}}{\|F_{\text{img}}\| \, \|F_{\text{pc}}\|}. \tag{6}$$

The alignment is optimized using contrastive loss:

$$\mathcal{L}_{CL} = -\sum_{a,b \in N(a)} \log\left(\frac{\exp\left(s(a^*, b^*)\right)}{\sum_{z \in N(a)} \exp(s(b, z))}\right). \tag{7}$$

B. Multimodal Knowledge Graph Representation

In addition to aligning vision-language representations, EMKG-VLM incorporates multimodal knowledge graphs (MKG) to improve perception and decision making. MKGs encode structured relationships between visual, textual, and knowledge-based elements, which improves the interpretation of navigation goals.

- **Subgraph** G'': Retrieved from the MKG, containing visual and semantic relationships.

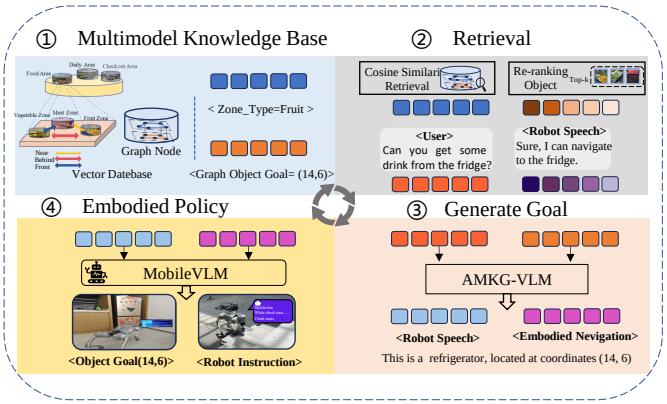

Fig. 2. The system processes navigation tasks by first categorizing the environment with the MRAG Graph Vector (Step 1), followed by Retrieval (Step 2) to gather relevant data. The Generate Goal step (Step 3) then creates navigation goals, such as locating objects at specific coordinates. Finally, Embodied Planning (Step 4) uses MobileVLM for real-time decision-making, enabling efficient navigation and task execution.

- **Node Embedding** Z_q: Encoded using Relational Graph Attention Networks (RGAT), capturing entity relations in the graph.

The cluster center $\phi(w_i)$ for each region is computed as

$$\phi(w_i) = \frac{\sum_{(x_j, y_j) \in W_i} g(x_j, y_j)}{|W_i|}. \tag{8}$$

The connection strength between two nodes $e(w_i, w_j)$ is defined as:

$$e(w_i, w_j) = \frac{\sum_{(x_i, y_i) \in W_i} \sum_{(x_j, y_j) \in W_j} q_{jk}}{|W_i||W_j|}, \tag{9}$$

where q_{jk} is an indicator function for spatial proximity.

The knowledge graph is dynamically updated as new observations are processed:

$$\phi(V_t) = \lambda I_{W_c} h^T(p_t, y_t) + \left(I - \lambda I_{W_c} I_{W_c}^T\right) \phi(V_{t-1}). \tag{10}$$

The Embodied knowledge graph is updated incrementally:

$$K_{t+1} = K_t + \Delta K_t, \tag{11}$$

where K_t represents the current knowledge graph and ΔK_t represents newly acquired information.

To align graph embeddings with textual features, we apply a knowledge adaptation function.

$$F_q' = \text{Softmax}\left(\frac{RF_q^T}{\sqrt{d_n}}\right) F_q. \tag{12}$$

Multimodal alignment is further optimized using a triplet loss function.

$$L_a = \sum_{j=1}^{M} \max(d(p_j, p_k) - d(p_j, p_m) + \beta, 0). \tag{13}$$

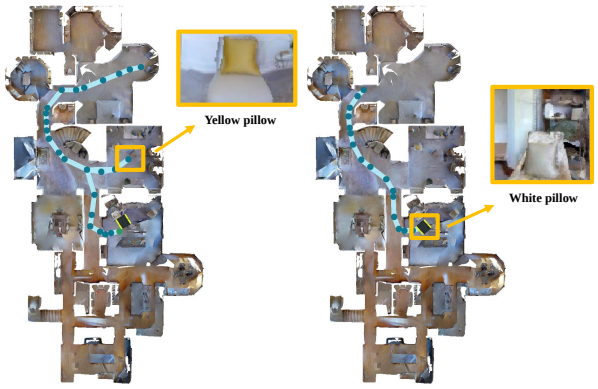

Fig. 3. This is a navigation experiment in the Habitat simulator. In the left image, the original algorithm first locates a yellow object before adjusting to find the white pillow based on feedback. In the right image, EMKG accurately locates the White pillow through image-text retrieval, demonstrating more precise navigation capabilities.

C. Adaptive Feedback Update Mechanism

To ensure adaptability, EMKG-VLM integrates an adaptive feedback update mechanism. Given an initial knowledge graph state K_t, new multimodal observations O_t refine the graph:

$$\Delta K_t = f_{\text{update}}(K_t, O_t), \tag{14}$$

$$K_{t+1} = K_t + \Delta K_t. \tag{15}$$

To determine whether an update is required, we compute an alignment score S_m:

$$S_m = \frac{Z_q \cdot F_q}{\|Z_q\| \|F_q\|}. \tag{16}$$

If $S_m < \tau$, the system refines the relationships between the nodes to improve the comprehension of the scene.

Reinforcement learning (PPO-Clip) is incorporated to optimize real-time navigation refinement:

$$\mathcal{L}_{\text{RL}} = E\left[\min(r_t A_t, \text{clip}(r_t, 1-\epsilon, 1+\epsilon)A_t)\right]. \tag{17}$$

D. 2D-3D Multimodal Mapping

To improve spatial reasoning, EMKG-VLM integrates a 2D-3D multimodal mapping strategy, aligning 2D image features with 3D scene representations. This enables:

- Context-aware navigation linking 2D semantic features to 3D spatial locations.
- Enhanced target identification by dynamically correlating visual and textual embeddings with 3D point clouds.

E. Additional Methodological Details

Additional details regarding the Mobile Vision-Language Model and implementation specifics are provided in the Appendix.

TABLE I

COMPARISON OF OBJECTNAV METHODS IN HABITAT (2D VS. 3D METHODS)

Method	SR (↑)	SPL (↑)	DTG (↓)
RIM [24]	0.503	0.170	-
OVRL-V2 [25]	0.281	0.647	-
PEANUT [26]	0.339	0.606	-
Skip-SCAR [27]	0.362	0.604	-
TSGM-IL	0.578	0.031	4.313
TSGM-RL	0.674	0.028	3.538
DD-PPO [28]	0.150	0.107	-
SemExp [29]	0.717	0.396	-
PONI [30]	0.736	0.410	-
LROGNav [31]	0.783	0.446	-
EMKG (Ours)	**0.812**	**0.506**	**2.134**

IV. EXPERIMENT

This section addresses the following research questions:

- **Q1. How does EMKG perform in Object Navigation (ObjectNav) within simulated environments?**
- **Q2. How does EMKG perform in constructing and recognizing multimodal knowledge across various environments?**
- **Q3. How does EMKG-VLM perform in processing open-ended questions?**
- **Q4. How does EMKG perform in real-world scenarios, specifically on the Unitree quadruped robot platform, integrating vision-language and reinforcement learning models for autonomous behavior?**

A. Simulated Experiments

1) EMKG Performance in Habitat Simulated Object Navigation: We evaluate EMKG for Object Navigation (ObjectNav) in the Habitat simulation platform using the Gibson and HM3D-Semantics v0.1 datasets, with Success Rate (SR), Success-weighted Path Length (SPL), and Distance-to-Goal (DTG) as key metrics. EMKG incorporates multimodal knowledge graphs and dynamic spatial reasoning, outperforming conventional methods. Our prompt is: "Imagine you are a robot programmed for navigation tasks. You interpret coordinates in the format (x, y), representing positions in a visual or image-based world. Use the provided data to analyze and respond based on the spatial context." This formulation enables EMKG to leverage structured spatial reasoning and contextual information effectively.

EMKG achieves an SR of 0.812%, improving upon LROGNav [31] and PONI [30] by 2.9% and 7.6%, respectively. SPL increases by 13.5% and 23.4% over these baselines, while DTG is reduced to 2.134%, representing a 39.7% improvement over TSGM-RL. Compared to reinforcement learning approaches such as DD-PPO [28] and RIM [24], EMKG significantly enhances path efficiency and overall success rates. Modular methods including PEANUT [26] and Skip-SCAR [27] exhibit lower SR and SPL, further demonstrating EMKG's effectiveness in structured navigation.

To establish a more comprehensive evaluation approach, we emphasize the integration of multimodal knowledge and real-time spatial adaptation. Unlike static models, our

TABLE II
ABLATION STUDY ON ZERO-SHOT RETRIEVA

Task		Train Datasets	Test Datasets	Rank 1(↑)	Rank 5(↑)	mAP(↑)
Q+D→T+D	Global	ARKitScenes	MultiScan	**84.78**	98.91	90.40
			3RScan	71.42	95.91	82.32
			ScanNet	77.56	85.23	80.45
	Ours	ARKitScenes	MultiScan	**97.82**	98.91	98.19
			3RScan	81.63	98.32	89.86
			ScanNet	80.12	92.12	**79.01**
	Global	MultiScan	ARKitScenes	94.56	98.91	96.89
			HM3D	83.39	95.66	88.61
			ScanNet	79.34	87.12	83.45
	Ours	MultiScan	ARKitScenes	95.65	98.91	97.82
			HM3D	89.53	98.55	**93.39**
			ScanNet	84.78	92.56	82.67
	Global	3RScan	ARKitScenes	92.34	97.89	95.12
			MultiScan	87.12	99.01	91.23
			ScanNet	**81.45**	88.34	85.01
	Ours	3RScan	ARKitScenes	96.89	99.01	**98.23**
			MultiScan	98.45	97.21	98.67
			ScanNet	**83.23**	93.01	84.12
	Global	ScanNet	ARKitScenes	91.45	97.23	94.89
			MultiScan	80.45	96.89	88.34
			SUN RGB-D	**65.78**	92.89	76.45
	Ours	ScanNet	ARKitScenes	94.12	99.01	**96.34**
			MultiScan	92.45	96.89	94.78
			SUN RGB-D	**77.12**	97.56	85.67

system enables dynamic navigation strategies, enhancing adaptability in non-deterministic environments. Experimental results show that EMKG-VLM achieves state-of-the-art performance in ObjectNav, demonstrating superior efficiency, robustness, and generalization compared to existing baselines

2) EMKG in Multimodal Fine-Grained Text-Image Retrieval and 3D Scene Construction: To address Q2, we evaluate EMKG in simulated environments using the SenseVerse dataset, which includes 68,000 3D scenes and 2.5 million scene-language pairs, along with established datasets such as ScanNet, MultiScan, ARKitScenes, and 3RScan. Precision and recall metrics were employed to assess the model's effectiveness in multimodal knowledge construction and retrieval tasks.

EMKG utilizes dedicated image and text encoders to generate high-dimensional embeddings, which are synthesized into a unified representation. The model re-encodes object attributes from SenseVerse to capture complex spatial relationships and geometric characteristics. For retrieval, EMKG employs a hybrid search methodology, integrating 3D coordinates and semantic information to refine predictions.

As shown in Table II, EMKG outperforms existing benchmarks in both global and fine-grained retrieval tasks. It achieves a 15.4% improvement in Rank 1 accuracy, along with significant enhancements in Rank 5 accuracy and mean Average Precision (mAP), comparing EMKG (Ours) to the Global method on the MultiScan dataset. Notably, EMKG reaches a Rank 1 accuracy of 94.56% in global retrieval and 95.65% in fine-grained retrieval on the MultiScan dataset, surpassing baseline models in large-scale multimodal learning.

These results highlight EMKG's superior performance in multimodal knowledge integration, cross-modal retrieval, and 3D scene understanding. Through advanced vision-language fusion and spatial reasoning, EMKG addresses key challenges in multimodal retrieval, such as object ambiguity and scene complexity, providing robust and precise retrieval capabilities in real-world environments.

B. Real-World Experiments

1) EMKG-VLM Performance in Open-World End-to-End Tasks: We evaluated EMKG-VLM on the NVIDIA Jetson AGX Orin platform, optimizing for CUDA acceleration to ensure real-time efficiency. The evaluation focused on two key metrics: Average Accuracy (Avg. Accuracy) and Inference Latency (Evalavg), which measures processing efficiency in tokens per second. As shown in Table III, EMKG-VLM achieves an Avg. Accuracy of 63.1% with an Evalavg of 45.30 tokens/s and a total inference time of 5.55s. These results significantly outperform MobileVLM and LLaVA-1.5, with EMKG-VLM 1.7B surpassing MobileVLM 1.7B by 8.1% in accuracy while maintaining the highest processing rate. The 3B variant further improves accuracy to 64.2%, outperforming LLaVA-1.5 (3.3B) by 7.3% despite a higher Evalavg.

Several optimizations contribute to this performance. Dynamic knowledge graph updates enhance real-time scene representation, while a lightweight processing architecture tailored for Jetson Orin reduces latency without compromising accuracy. Additionally, contrastive knowledge alignment between 2D and 3D data improves multimodal processing efficiency. The system's 3D capabilities—enabled by multi-

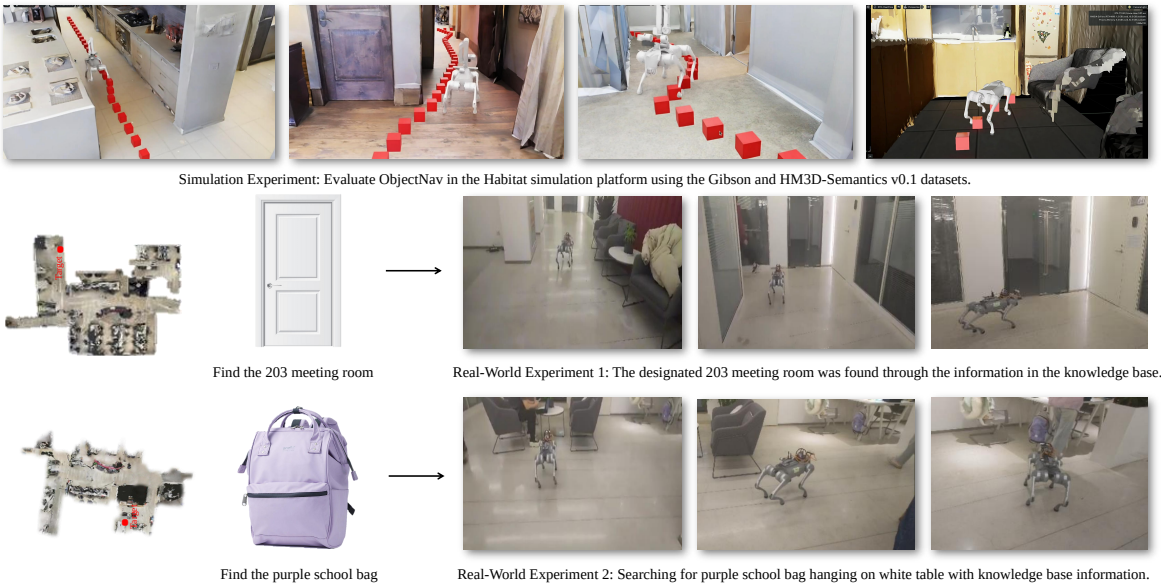

Fig. 4. Simulation Experiment to Real-World Experiment

modal knowledge graph embeddings and real-time spatial reasoning—provide significant advantages over traditional 2D-based models in processing complex environments.

These results validate EMKG-VLM's effectiveness for real-time robotic applications in dynamic environments. Its rapid inference (5.55s for EMKG-VLM 1.7B) positions it as a robust solution for robotic perception and decision-making, where 3D spatial understanding is crucial for accurate and efficient navigation.

2) Model Latency Comparison in Indoor Scene Detection: In addition to real-time navigation, we compared EMKG-VLM's efficiency with other vision-language models in indoor scene detection tasks. As shown in Table III, EMKG-VLM consistently achieves high accuracy while minimizing inference latency. EMKG-VLM 1.7B outperforms MobileVLM 1.7B by 8.1% in accuracy and further enhances the 3B variant, surpassing LLaVA-1.5 (3.3B) by 7.3%. Importantly, EMKG-VLM reduces latency compared to these models, reinforcing its suitability for real-time applications that require both fast processing and high accuracy.

These findings underline EMKG-VLM's ability to optimize real-time vision-language processing for embedded systems. Its multimodal retrieval mechanisms enable effective decision-making and make it an ideal solution for real-world robotics in both 2D and 3D environments.

3) EMKG's Real-World Autonomy: EMKG's real-world autonomy was tested on a Unitree quadruped robot with a mid360 depth sensor, an RGB camera, and Jetson Orin. The robot, trained via deep reinforcement learning in a curriculum-based Teacher-Student framework, navigated varied terrains, including flat surfaces and stairs.

As shown in Figure 7, EMKG successfully reached a meeting room and located a purple school bag on a white table, demonstrating strong spatial reasoning and adaptive path-following. When encountering obstacles, it dynamically adjusted its trajectory using multimodal perception to enhance navigation efficiency.

This evaluation employed an omnivector action interface via Lightweight Communications and Marshalling (LCM), enabling real-time interaction between the vision-language model (VLM) and the robot's locomotion policy. This significantly improved task completion rates and navigation success.

These results confirm EMKG's ability to transition from simulation to real-world deployment, effectively processing multimodal inputs and adapting to dynamic environments. Its strong 3D spatial awareness and navigation precision highlight its potential for real-world embodied AI applications.

V. Conclusion

This work introduces the Embodied Multimodal Knowledge Graph Vision-Language Model (EMKG-VLM), a novel approach designed to enhance autonomous robotic navigation and task execution in complex, dynamic environments. By integrating multimodal information processing with continuous updates to the dynamic knowledge graph, EMKG-VLM achieves superior multimodal reasoning and adaptability across a wide range of simulated and real-world scenarios. The system excels in tasks demanding advanced spatial understanding and effective cross-modal retrieval, showcasing its strength in 3D scene comprehension and the integration of multimodal data for navigation. Furthermore, the system's real-time inference capabilities, demonstrated on embedded platforms like the Jetson Orin, validate its robustness and practical applicability. Overall, EMKG-VLM represents a significant advancement in the development of embodied intelligent systems, particularly for real-world navigation tasks in open, unstructured environments.

REFERENCES

[1] A. Wahid, A. Stone, K. Chen, B. Ichter, and A. Toshev, "Learning object-conditioned exploration using distributed soft actor critic," in *Proceedings of the 2020 Conference on Robot Learning*. Proceedings of Machine Learning Research, 2021, pp. 1–10.

[2] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Language conditioned spatial relation reasoning for 3d object grounding," *Advances in neural information processing systems*, vol. 35, pp. 20 522–20 535, 2022.

[3] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez Opazo, S. Gould *et al.*, "Vln (sic) bert: A recurrent vision-and-language bert for navigation," 2021.

[4] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3d instance segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, 2020, pp. 4867–4876.

[5] H.-T. L. Chiang *et al.*, "Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs," *ArXiv preprint arXiv:2407.07775*, 2024.

[6] J. Loo, Z. Wu, and D. Hsu, "Open scene graphs for open world object-goal navigation," *ArXiv preprint arXiv:2407.02473*, 2024.

[7] W. Cai *et al.*, "Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill," in *Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5228–5234.

[8] L. Gui, B. Wang, Q. Huang, A. Hauptmann, Y. Bisk, and J. Gao, "Kat: A knowledge augmented transformer for vision-and-language," 12 2021.

[9] Y. Lin, Y. Xie, D. Chen, Y. Xu, C. Zhu, and L. Yuan, "Revive: Regional visual representation matters in knowledge-based visual question answering," 2022. [Online]. Available: https://arxiv.org/abs/2206.01201

[10] D. Schwenk *et al.*, "Knowledge-augmented visual question answering," *arXiv preprint arXiv:2202.11986*, 2022.

[11] L. Gao *et al.*, "Multimodal retrieval and generation in vision-and-language navigation," *arXiv preprint arXiv:2201.10280*, 2022.

[12] B. Wu *et al.*, "Objaverse: Large-scale object dataset for learning object representations," *arXiv preprint arXiv:2104.13582*, 2021.

[13] H. Zhang *et al.*, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[14] K. Lin *et al.*, "Revive: A retrieval-augmented visual question answering system," *arXiv preprint arXiv:2206.01439*, 2022.

[15] G. Baruch, Z. Chen, A. Dehghan, T. Dimry, Y. Feigin, P. Fu, T. Gebauer, B. Joffe, D. Kurz, A. Schwartz *et al.*, "Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data," *arXiv preprint arXiv:2111.08897*, 2021.

[16] T. B. Brown, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[17] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*. Springer, 2020, pp. 422–440.

[18] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[19] I. Armeni, Z.-Y. He, J. Gwak, A. R. Zamir, M. Fischer, J. Malik, and S. Savarese, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 5664–5673.

[20] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, "Scanqa: 3d question answering for spatial scene understanding," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 19 129–19 139.

[21] D. Batra, A. Gokaslan, A. Kembhavi, O. Maksymets, R. Mottaghi, M. Savva, A. Toshev, and E. Wijmans, "Objectnav revisited: On evaluation of embodied agents navigating to objects," *arXiv preprint arXiv:2006.13171*, 2020.

[22] D. Z. Chen, Q. Wu, M. Nießner, and A. X. Chang, "D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans," 2021.

[23] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, J. Yang, X. Zheng, K. Li, X. Sun, Y. Wu, and R. Ji, "Mme: A comprehensive evaluation benchmark for multimodal large language models," *arXiv preprint arXiv:2306.13394*, 2023.

[24] S. Chen, T. Chabal, I. Laptev, and C. Schmid, "Object goal navigation with recursive implicit maps," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 7089–7096.

[25] K. Yadav, A. Majumdar, R. Ramrakhya, N. Yokoyama, A. Baevski, Z. Kira, O. Maksymets, and D. Batra, "Ovrl-v2: A simple state-of-art baseline for imagenav and objectnav," *arXiv preprint arXiv:2303.07798*, 2023.

[26] A. J. Zhai and S. Wang, "Peanut: Predicting and navigating to unseen targets," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2023, pp. 10 926–10 935.

[27] Y. Liu, Y. Cao, and J. Zhang, "Skip-scar: Hardware-friendly high-quality embodied visual navigation," *arXiv preprint arXiv:2405.14154*, 2024. [Online]. Available: https://arxiv.org/abs/2405.14154

[28] E. Wijmans, A. Kadian, A. Morcos, S. Lee, I. Essa, D. Parikh, M. Savva, and D. Batra, "Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames," *arXiv preprint arXiv:1911.00357*, 2019.

[29] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 4247–4258.

[30] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 890–18 900.

[31] L. Sun, A. Kanezaki, G. Caron, and Y. Yoshiyasu, "Leveraging large language model-based room-object relationships knowledge for enhancing multimodal-input object goal navigation," *arXiv preprint arXiv:2403.14163*, 2024. [Online]. Available: https://arxiv.org/abs/2403.14163

APPENDIX

A. Additional details of the method2D-3D Multimodal Mapping

EMKG-VLM bridges 2D visual data and 3D spatial information using a vision-language alignment mechanism. The process follows three steps:

1) **Feature Extraction**: - The 2D image encoder extracts visual features F_{img}. - The 3D point cloud encoder (e.g., PointNet++) extracts geometric features F_{pc}.

$$F_{\text{img}} = \text{ImageEncoder}(I), \tag{18}$$

$$F_{\text{pc}} = \text{PointCloudEncoder}(P), \tag{19}$$

2) **Semantic Alignment**: - Using cosine similarity, 2D embeddings are matched with 3D point cloud embeddings:

$$\text{corr}(F_{\text{img}}, F_{\text{pc}}) = \frac{F_{\text{img}} \cdot F_{\text{pc}}}{\|F_{\text{img}}\| \|F_{\text{pc}}\|}, \tag{20}$$

- If alignment confidence is low, the system refines mappings by leveraging language-based object relations.

3) **Knowledge Graph Integration**: - The aligned multimodal embeddings are integrated into the knowledge graph K_t, forming a spatially aware object relationship database.

By combining 2D-3D alignment and dynamic knowledge updates, EMKG-VLM enhances real-time scene understanding and improves the performance of object-goal navigation.

Fig. 5. Experimental supplement: Example of robot dog climbing stairs under vlm command 1

Fig. 6. Develop a system for object detection and information feedback in an open-world environment.

B. Additional details of the method Mobile Vision-Language Model

The MobileLLaMA series serves as the core LLM for enhancing the Multimodal Knowledge Graph Module, optimized for resource-limited devices and real-time deployment. Integrating MobileVLM improves indoor scene understanding by efficiently processing visual and language data. A lightweight downsampling projector (LDPv2) aligns visual data with the language model for seamless multimodal synchronization. Optimization techniques like pruning and quantization enhance MobileVLMv2's efficiency, supporting real-time updates to the Embodied Multimodal Knowledge Graph (EMKG). The objective function guiding this optimization is:

$$\mathcal{L} = \sum_{i=1}^{5} \lambda_i \mathcal{L}_i, \tag{21}$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ are image and text losses, $\mathcal{L}_3$ ensures cross-modal alignment, $\mathcal{L}_4$ provides regularization, and $\mathcal{L}_5$ stabilizes training via gradient penalties.

The navigation policy network determines the system's actions based on the current state, as described by the policy function:

$$\text{Policy}_{\text{target}} = f_{\text{policy}}(\text{state}, \text{action}), \tag{22}$$

where state refers to the current system state and action represents the action taken.

Transfer learning techniques are employed for behavior instruction transfer, enabling the model to adapt instructions to new environments. The loss function for this transfer process is:

TABLE III

COMPARISON OF INDOOR SCENE DETECTION MODEL LATENCY

Model	Avg. Accuracy (%)	Evalavg (token/s)	Total Time (s)
MGM-2B	61.8	19.10	14.30
EMKG-VLM 3B	64.2	28.95	8.67
LLaVA-1.5 3.3B	55.8	17.80	14.85
MobileVLM 3B	59.3	26.90	9.05
MGM-2B	61.8	19.10	14.30
LLaVA-1.5 1.4B	51.1	37.95	6.98
MobileVLM 1.7B	55.0	42.50	5.89
EMKG-VLM 1.7B	**63.1**	**45.30**	**5.55**

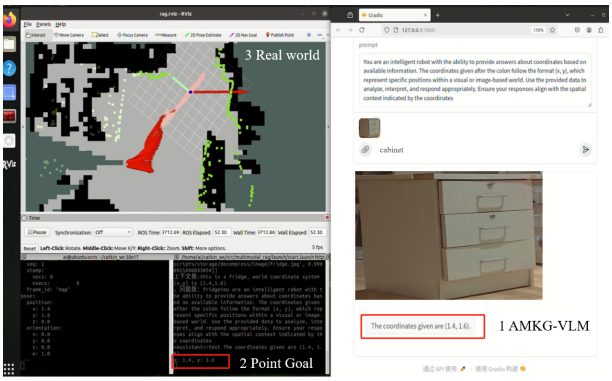

Fig. 7. Experimental supplement: Example of robot dog climbing stairs under vlm command 2

$$\text{Loss}_{\text{transfer}} = \text{Loss}_{\text{source}} - \text{Loss}_{\text{target}}, \tag{23}$$

where $\text{Loss}_{\text{source}}$ and $\text{Loss}_{\text{target}}$ represent the losses for the source and target tasks, respectively.