

FSR-VLN: Fast and Slow Reasoning for Vision-Language Navigation with Hierarchical Multi-modal Scene Graph

Xiaolin Zhou^{1*}, Tingyang Xiao^{1*}, Liu Liu¹, Yucheng Wang¹, Maiyue Chen¹,
Xinrui Meng², Xinjie Wang¹, Wei Feng¹, Wei Sui², and Zhizhong Su¹

Abstract—Visual-Language Navigation (VLN) is a fundamental challenge in robotic systems, with broad applications for the deployment of embodied agents in real-world environments. Despite recent advances, existing approaches are limited in long-range spatial reasoning, often exhibiting low success rates and high inference latency, particularly in long-range navigation tasks. To address these limitations, we propose FSR-VLN, a vision-language navigation system that combines a Hierarchical Multi-modal Scene Graph (HMSG) with Fast-to-Slow Navigation Reasoning (FSR). The HMSG provides a multi-modal map representation supporting progressive retrieval, from coarse room-level localization to fine-grained goal view and object identification. Building on HMSG, FSR first performs fast matching to efficiently select candidate rooms, views, and objects, then applies VLM-driven refinement for final goal selection. We evaluated FSR-VLN across four comprehensive indoor datasets collected by humanoid robots, utilizing 87 instructions that encompass a diverse range of object categories. FSR-VLN achieves state-of-the-art (SOTA) performance in all datasets, measured by the retrieval success rate (RSR), while reducing the response time by 82% compared to VLM-based methods on tour videos by activating slow reasoning only when fast intuition fails. Furthermore, we integrate FSR-VLN with speech interaction, planning, and control modules on a Unitree-G1 humanoid robot, enabling natural language interaction and real-time navigation. The code and video can be available at <https://horizonrobotics.github.io/robotlab/fsr-vln/>

I. INTRODUCTION

Visual-Language Navigation (VLN) is a fundamental task in embodied AI, enabling robots to operate effectively in complex real-world environments [1]. Despite significant progress in map-free VLN research and growing efforts to equip robots with visual-language reasoning [2]–[5], existing methods remain limited in long-range spatial cognition [6], particularly for long-range navigation. A key bottleneck is the lack of persistent long-range spatial memory, which encodes, organizes, and retrieves environmental knowledge. Such memory allows robots to capture comprehensive spatial relationships and adapt to complex indoor environments [7], [8]. Without it, robots struggle to understand and reason about long-range spaces. To address this limitation, recent spatial memory-based approaches have explored embedding 2D and 3D geometric maps with semantic features [9]–[12], integrating dense geometric reconstructions with the pre-trained zero-shot Vision Foundation Model (VFM) such as CLIP [13], OWL-ViT [14], and LSeg [15]. These approaches support open-vocabulary object-level retrieval while

maintaining high geometric fidelity. Building on geometric semantic maps, methods such as Clio [16], ConceptGraph [17]–[19], RoboExp [18], and OpenIN [10] leverage open-vocabulary 3D scene graphs to represent long-range environments and provide semantic interfaces for prompting large language model (LLM). Parallel work, including HOVSG [20] and IRS [21], abstracts these dense maps into hierarchical structures spanning floors, rooms, and objects, supporting efficient semantic object retrieval for long-range navigation.

Although 2D and 3D semantic maps and 3D scene graphs provide geometry-consistent and hierarchical spatial memory, they rely on pre-extracted visual features [3], [22], lack direct interaction with a vision-language model (VLM), and are sensitive to geometric noise. These limitations, in turn, hinder their adaptability to diverse user instructions and complex real-world navigation [23], [24].

Inspired by human navigation, where individuals reason over previously observed images or objects without detailed 3D maps, image-based topological navigation has emerged as a promising alternative [25]–[28]. By retaining raw image data, image-based topological maps facilitate interaction with LLM and VLM, such as GPT-4o and Gemini, which excel at reasoning over both visual and textual inputs [29]. Despite their high success rate in image-level navigation, existing methods lack explicit 3D geometric structure and usually rely on video captioning, which is inefficient for reasoning over long sequences [30].

To address these challenges, we introduce **FSR-VLN**, a novel **Fast and Slow Reasoning** system for vision-language navigation. The contributions of our paper are as follows.

- 1) To the best of our knowledge, we are the first to introduce **Hierarchical Multi-modal Scene Graph (HMSG)** map representation. Our approach elegantly synthesizes the strengths of hierarchical scene graphs for long-range navigation with image topological graphs for fine-grained reasoning, which is essential for LLM/VLM-based systems. This unique hybrid representation facilitates progressive retrieval, from coarse-grained navigational cues to precise goal localization, all while preserving multi-modal information to ensure robustness and high success rates in real-world environments.
- 2) Inspired by the dual-process theory of human cognition, we introduce a novel **Fast-to-Slow Navigation Reasoning** system. The system operates in two distinct stages: a fast, intuitive matching step retrieves candidate views and objects, followed by a slow, deliberate

¹Horizon Robotics, Beijing, China

²D-Robotics Robotics, Beijing, China

*Equal contribution

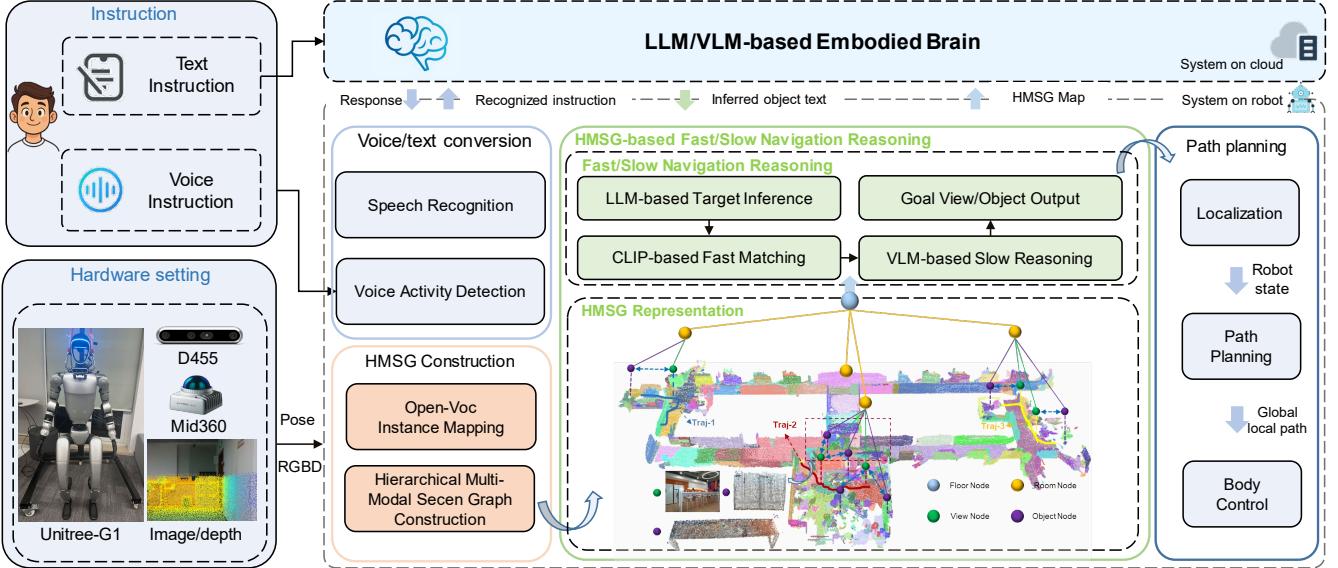


Fig. 1. System Overview. The proposed humanoid robotics navigation system integrates HMSG with FSR to achieve view/object-level real-world long-range navigation. Specifically, RGBD and pose data are first utilized to construct HMSG, which provides a hierarchical and multimodal feature-based representation of the environment. During online interaction, the user's text or voice input is converted into instructions via voice activity detection and speech recognition, and the LLM infers the target object. Based on the HMSG, fast-matching and slow VLM reasoning jointly identify the optimal goal view/object. The identified goals are subsequently used by the global path planning.

reasoning stage where an VLM verifies and refines the final goal. This multi-stage architecture seamlessly integrates efficient feature-space matching with robust VLM-driven visual verification, leading to a significant improvement in success rates for real-world, long-range navigation.

- 3) Evaluated on 87 robot-collected instructions across four diverse categories in long-range real-world indoor environments, FSR-VLN achieves state-of-the-art (SOTA). It achieves 77% higher success rates than HOVSG and 167% higher than MobilityVLA through superior object localization, while reducing average response time by 82% compared to MobilityVLA.
- 4) By integrating FSR-VLN with speech interaction, planning, and control modules on the Unitree-G1 humanoid robot, we present a comprehensive humanoid robotics navigation system that is capable of operation guided by natural language.

II. RELATED WORKS

A. Object Navigation with Geometric Semantic Map

Recent works have extended geometric point cloud maps by incorporating semantic features to enable object navigation (ObjNav). Systems such as OK-Robot [31], VLMap [23], OVL-MAP [32], and BeliefMapNav [33] embed features from Vision Foundation Models (VFs, e.g., CLIP [13], OWL-ViT [14], LSeg [15]) into voxel grids. While this approach preserves geometric consistency and supports open-vocabulary queries, it suffers from several critical limitations. Despite these advances, the performance of the systems remains limited by the inherent weaknesses of VFs and their sensitivity to 3D reconstruction noise and odometry drift. Reliance on a single map representation further restricts

integration with LLMs and VLMs, which excel at cross-modal reasoning and generalization.

Building on semantic point cloud maps, scene graph-based methods represent objects and spatial concepts as nodes, with their relations encoded as edges, providing compact and expressive models of long-range environments [34]–[36]. This object-centric decomposition supports higher-level reasoning for navigation and manipulation [20]. While many works employ 3D scene graphs [35]–[37] to efficiently model large environments, most rely on closed-set semantics. To enable open-vocabulary queries, approaches such as ConceptGraphs [38], HOVSG [20], DOVSG [39], and OpenIN [40] construct open-vocabulary 3D scene graphs, which can also interface with LLM.

However, these approaches typically depend on rigid instruction formats (e.g., “object A in region B on floor C”) and semantic similarity-based retrieval, which limit flexibility and hinder full exploitation of VLM reasoning. In contrast, our method leverages LLM and VLM to interpret diverse user instructions and to refine semantic retrieval, thereby improving navigation success in real-world environments.

B. ObjNav with Image-based Topological Graph

Several recent works represent environments as image-based topological graphs to facilitate VLN. MobilityVLA [27] introduces a vision-language action (VLA) framework that leverages long-context VLM for goal frame retrieval and uses topological graphs for waypoint planning. Uni-NaVid [41] adopts a video-based VLA approach, retrieving the most visually similar image to a query object and generating robot actions in an end-to-end manner. ReMEMBR [42] emphasizes spatio-temporal memory by storing caption embeddings with pose and time metadata for retrieval-augmented reasoning. More recently, MapGPT [43] incorporates map-guided

prompting and adaptive path planning, significantly improving zero-shot generalization. Astra [44] extends image-based topological maps with landmark information, while RoboHop [45] employs a zero-shot “segment servoing” strategy to reach object subgoals. Building on RoboHop’s open-set navigation pipeline, TANGO [46] generates sub-object goals through a global path planner grounded in an object-level topological graph.

Although these methods preserve rich semantic cues and achieve strong performance in image-level reasoning, they rely solely on video captioning, which is inefficient in long video contexts and often leads to mismatches between semantic and geometric information [44]. In real-world environments, however, abundant spatial information—such as room numbers, equipment labels, and other annotations can significantly improve navigation performance [42]. To leverage this, we propose a hierarchical multi-modal scene graph representation that encodes both spatial and semantic features across floors, rooms, views, and objects. Building on these multi-modal features, we introduce a fast-to-slow reasoning mechanism to enhance navigation success rates.

III. METHOD

Our primary contribution is centered on map representation and navigation reasoning. To demonstrate the practical utility of our FSR-VLN system, we integrate it with speech interaction, planning, and control modules on the Unitree-G1 humanoid robot, forming a complete real-world navigation system.

As illustrated in Fig. 1, we first build HMSG with real-world LiDAR–camera datasets collected by Unitree-G1. Specifically, the SOTA SLAM system FAST-LIVO2 [47] is used to extract RGBD data and poses to create an instance-level open-vocabulary map. Building on an instance-level open-vocabulary map [20], we construct the Hierarchical Multi-modal Scene Graph (HMSG), which divides the environment into four levels—floor, room, view, and object. Each node is enriched with multi-modal features, including geometric attributes, semantic information, and topological connections.

During online interaction, the speech recognition module [48] converts user speech into text, which is then parsed by an LLM to extract the corresponding query. In the FSR pipeline, the query text is first grounded to candidate regions and objects through CLIP-based similarity matching and then refined via VLM-based visual verification to ensure robust and accurate navigation. By combining the HMSG representation with fast-to-slow reasoning (FSR), FSR-VLN enables efficient and interpretable navigation, particularly in long-range and cluttered real-world environments. Further details of this mechanism are provided in Section III-C. After obtaining the target object’s coordinates, the system performs path planning and whole-body control to reach the goal.

A. Hierarchical Multi-modal Scene Graph Representation

As shown in Fig. 2, the proposed HMSG is organized into four levels: floor, room, view, and object nodes. Each node

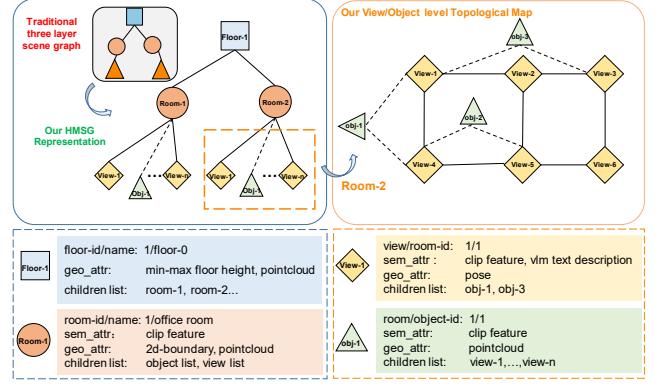


Fig. 2. HMSG representation. Our proposed HMSG is a four-level hierarchy: floor, room, view, and object nodes. Each node contains multi-modal features, including geometric attributes, semantic attributes, and topological relationships.

encodes multi-modal features, including geometric attributes, semantic attributes, and explicit topological relationships. This design supports layer-wise retrieval and enables fast-to-slow navigation reasoning with LLM and VLM.

Floor nodes record unique floor identifiers, names, geometric attributes (e.g., min/max heights, PLY point clouds), and references to contained room nodes.

Room nodes store ID, 2D polygon boundaries, point clouds, semantic attributes (name, CLIP embeddings), and links to associated view and object nodes. They support long-range object-level navigation across rooms.

Object nodes represent discrete instances within a room. Each node has geometric properties (3D bounding boxes, point clouds), semantic embeddings, and links to parent room nodes and image views.

Prior scene graph methods typically use floor, room, and object nodes, relying solely on CLIP matching, which constrains visual–spatial understanding. While image-based topological graphs with VLM reasoning improve navigation success, they lack 3D structure and incur high computational cost. To address these limitations, we introduce **view nodes** as a dedicated layer to capture spatial and perceptual relationships between image views and objects. This layer enables reasoning over image views with VLM, allowing the robot to select contextually relevant views for image-level navigation while enhancing object-level localization.

View nodes represent specific visual perspectives within a room. Each node stores CLIP embeddings and VLM-generated descriptions as semantic features, and camera poses as geometric attributes. They are linked to visible object nodes, encoding visibility relationships that support multi-view perception and image- or object-level grounding. Undirected edges between view nodes are defined by relative poses, enabling global path planning.

B. Hierarchical Multi-modal Scene Graph Construction

Following HOVSG [20], we first construct an open-vocabulary map at the instance-level. Objects and views are associated with rooms based on geometric overlap in top-down maps, resulting in a scene graph enriched with both

Algorithm 1 HMSG Construction

Require: Floor–room–object layout, views with poses
Ensure: Hierarchical multi-modal scene graph G

- 1: Initialize empty graph G
- 2: **for** each floor f **do**
- 3: Add node(f) to G
- 4: **for** each room r in f **do**
- 5: Add node(r) and edge(f, r)
- 6: **for** each image view v in r **do**
- 7: Add node(v) and edge(r, v)
- 8: **for** each object o in r **do**
- 9: Add node(o) and edge(r, o)
- 10: best_view \leftarrow None, min_depth $\leftarrow \infty$
- 11: **for** each view v in r where o is visible **do**
- 12: Add edge(v, o)
- 13: **if** mean_depth(o, v) $<$ min_depth **then**
- 14: best_view $\leftarrow v$, min_depth \leftarrow mean_depth(o, v)
- 15: Assign best_view to o
- 16: **return** G

geometric and semantic features.

As shown in Algorithm 1, the Hierarchical Multi-modal Scene Graph (HMSG) takes a floor–room–object layout and posed image views as input, and outputs a directed graph G , where nodes represent semantic entities (floors, rooms, views, objects) and edges encode structural relations (e.g., room-in-floor, object-in-room). Floors and rooms are sequentially added with CLIP features; unlike HOVSG, FSR-VLN leverages GPT-4o to infer room names from image views. Each view is associated with CLIP embeddings, VLM-generated captions, and camera poses, while the edges connect visible objects to their corresponding views. Additionally, we compute the mean depth of each object across visible views and select the view with the minimum depth (closest appearance) as its representative. The resulting HMSG encodes hierarchical topology and multi-modal features—including semantic, geometric, and visibility information, providing the foundation for fast-to-slow reasoning to improve navigation performance.

C. LLM/VLM and Multi-modal feature-based Fast-to-slow Navigation Reasoning

As illustrated in Fig. 3, our navigation reasoning comprises three steps: LLM-based user instruction understanding, Fast Matching, and Slow Reasoning.

LLM-based user instruction understanding. To interpret user instructions for object navigation, we introduce an LLM-based object query inference module that handles both spatially explicit and implicit natural language commands. For spatial instructions (e.g., “Take me to the blue cylindrical stool in the office”), the LLM acts as a hierarchical concept parser, decomposing the input into structured components such as floor, region, and object. The output maps directly to the nodes in the hierarchical scene graph, enabling precise localization.

For non-spatial instructions (e.g., “I’m tired, where can I find a blue cylindrical stool?” or “I’m thirsty”), the LLM functions as a goal inference agent, identifying the most relevant object or region based on user intent. The inferred object-level semantics are then resolved into spatial targets via the scene graph. This prompting framework supports generalization across diverse user expressions, both explicit and implicit, facilitating robust human-robot interaction in real-world environments.

Fast Matching: Multi-modal feature-based Goal Room/View/Object Chosen. After instruction understanding, if a room name is provided in the instruction, the system first matches the room and subsequently performs view and object matching within that room. Using the image-view nodes in Fig. 2 and the LLM-interpreted text query, CLIP-based fast matching is performed between the query text and HMSG view-layer embeddings to identify the goal view. In parallel, object-level locations are determined by matching CLIP features between the query text and object embeddings, with the object exhibiting the highest similarity considered the potential target. Although fast grounding of the goal view and object instances, the navigation is limited by the matching ability of the CLIP model, which may provide the wrong object. Thus, based on CLIP-based fast matching, VLM-based slow reasoning is further introduced to obtain a more precise goal view and object.

Slow Reasoning: LLM/VLM-based View/Object Chosen Refinement. After fast matching, the goal view, goal object, and other candidate views are obtained. Using the reasoning capabilities of VLM, we employ GPT-4o to verify whether the object is present in the best view corresponding to the matched object from the fast-matching step. Since objects in the HMSG map are guaranteed to appear within their corresponding best views, if GPT-4o determines that the interpreted object is not present in the best view corresponding to the matched object, the matched object is considered unreliable. Although fast matching may fail, the correct goal view may still exist among the matched views or unmatched candidates. To address this, we reapply both LLM and VLM reasoning to identify the optimal goal image. Specifically, we first use the LLM to reason over textual descriptions of unmatched views in the HMSG, selecting the most semantically consistent image as view-1. Next, the fast-matched view is compared with view-1, and VLM inference is applied to determine the final optimal goal image. Once the goal image is finalized, the goal object is updated by traversing its object list in the HMSG and recalculating the CLIP similarities between the query text and each object.

IV. EXPERIMENT

A. Experiment Setup

Dataset. To validate our system in the real world, we use a Unitree-G1 humanoid robot equipped with a calibrated Intel RealSense D455 RGBD camera and a Mid360 LiDAR (Figure 1) to collect LiDAR–camera data across long-range office environments with long corridors and multiple rooms, denoted as Room1, Room2, Room3, and Room4. We also

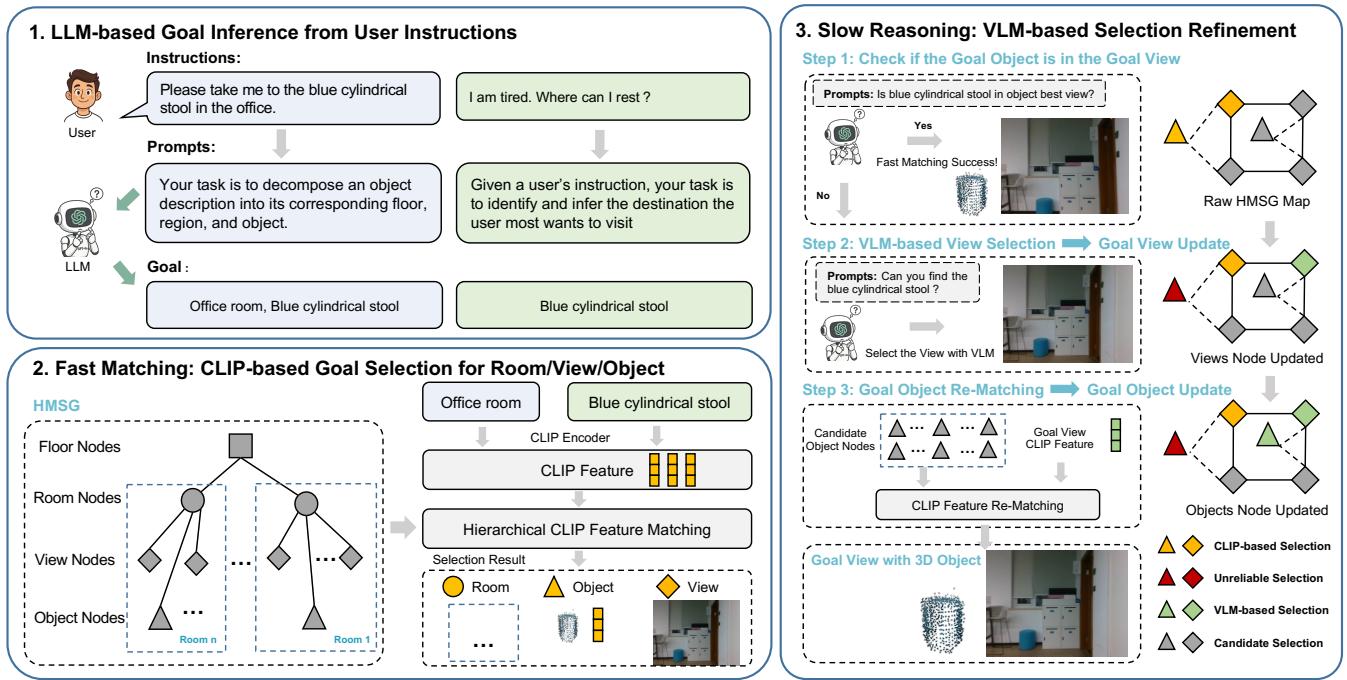


Fig. 3. The navigation reasoning follows a coarse-to-fine process: 1). LLM interprets user instructions into structured object/room queries; 2). CLIP-based fast matching, as intuition retrieves candidate goal rooms, views, and objects. 3). VLM-based slow reasoning refines the candidate results to ensure accurate goal view and object selection.

evaluated our approach on eight scenes from the HM3D-SEM dataset [49].

User Instructions. To evaluate the generalization of our system across diverse instructions, we follow the experimental setup of MobilityVLA [27] and crowd-source 87 user instructions in four categories: Reasoning-Free (RF) with explicit targets, Reasoning-Required (RR) requiring implicit goal inference, Small Objects (SO), which are challenging due to limited visual footprints, and Spatial Target (ST) designed to test long-range memory using room-type information. The instructions span 23, 18, 15, and 14 object categories, respectively.

Metrics. In this paper, we focus on evaluating whether the target has been successfully retrieved. Following the metric in [27], [50], we report the success rate (SR) and retrieval success rate ($RSR_{top-n@k}$), where the RSR is defined as the percentage of queries for which at least one of the top- n predictions ($n \in 1, 5$) lies within k meters (Euclidean distance) of the ground truth. We evaluate performance with $k \in 1, 2, 3, 4, 5, m$ to account for positional variance introduced by 2D image-derived nodes and 3D point clouds. Notably, when $n = 1$ and $k = 1$, SR is equivalent to RSR.

Baselines. We evaluated FSR-VLN against several SOTA ObjNav methods using different map representations: CLIP-based 3D voxel maps (OK-Robot [31]), CLIP-based 3D scene graphs (HOVSG [20]), and image-based topological graphs (MobilityVLA [27]). OK-Robot performs object retrieval using CLIP-based OWL-ViT [14] features. HOVSG retrieves floors, rooms, and objects based on CLIP features, with its 3D scene graph structure supporting long-range retrieval. MobilityVLA retrieves goal images from all views; as it is not open-sourced, we implemented a

version using GPT-4o, first selecting the top 50 candidate frames via CLIP similarity before inference due to GPT-4o's context limitations. OK-Robot and MobilityVLA lack room-level information and cannot interpret Spatial Target (ST) instructions referencing a specific room. For fair comparison, we consider only the object query derived from LLM-based instruction understanding for these methods and HOVSG. Retrieval is also counted as successful if objects of the same type from other rooms are returned for ST instructions. In contrast, FSR-VLN leverages room-level spatial information, and retrieval is considered successful only when the target object is in the specified room. By integrating the HMSG map with navigation reasoning, FSR-VLN jointly predicts the goal view and object, and a retrieval is successful only if both are correctly identified.

In the HM3D-SEM dataset, as osmAG-LLM, we select HOV-SG and osmAG-LLM as our baseline because they share fundamental design principles with our approach [50].

B. Benchmark Results

Quantitative Analysis.

The performance of FSR-VLN compared to other methods on real-world datasets is summarized in Table I, reporting the average success rate (SR) over 87 instructions across four evaluation sets. FSR-VLN achieves the highest SR of 92% (80/87), substantially outperforming baselines: MobilityVLA: 34.5% (30/87), OK-Robot: 60.9% (53/87), HOVSG: 51.7% (45/87). This corresponds to relative improvements of 167%, 51%, and 77%, respectively, demonstrating the effectiveness of our approach. A similar trend is observed for RSR@Top1: FSR-VLN consistently achieves the best performance across distance thresholds, reaching 96.6%

TABLE I

OBJECT RETRIEVAL COMPARISON WITH SR, RSR@TOP1, AND RSR@TOP5 METRICS. THE METRICS ARE VISUALIZED **GREEN** (BEST) TO **RED** (WORST).

Method	Map Representation&Navigation	Time	SR	RSR@Top1					RSR@Top5				
				1m	2m	3m	4m	5m	1m	2m	3m	4m	5m
OK-Robot [10]	CLIP based 3D Voxel Map	0.2s	0.609	0.609	0.609	0.609	0.609	0.609	0.632	0.632	0.632	0.632	0.632
HOVSG [20]	CLIP based 3D Scene Graph	0.2s	0.517	0.517	0.573	0.586	0.596	0.596	0.770	0.816	0.828	0.828	0.828
MobilityVLA [27]	Image based Topological Graph + VLM	30s	0.345	0.345	0.598	0.759	0.805	0.954	-	-	-	-	-
FSR-VLN (Ours)	Multimodal 3D Scene Graph + VLM	5.5s	0.920	0.920	0.943	0.943	0.966	0.966	-	-	-	-	-

(84/87) at 4–5 meters, indicating robust long-range retrieval. Importantly, retrieval in FSR-VLN is considered successful only if the target object is located within the specified room; even under this stricter criterion, it achieves the highest SR and RSR, highlighting the advantages of HMSG-based reasoning.

MobilityVLA, constrained by semantic–geometric mismatches [44] due to weak 3D spatial cues, exhibits the lowest short-range $RSR_{top-1@k}$ ($k = 1, 2m$). However, at larger distance tolerances (3–5m), it ranks second to FSR-VLN, reflecting its strong image-sequence reasoning capabilities.

OK-Robot, constrained by OWL-ViT-based retrieval, achieves an average SR of 61%, while HOVSG, similarly limited by CLIP-like models, reaches 52%. Both methods lack spatial verification of matched objects, often resulting in navigation failures. In contrast, FSR-VLN leverages the HMSG representation, which encodes geometric, semantic, and topological relations across floors, rooms, views, and objects. Built on this representation, the VLM-based FSR mechanism first performs fast feature matching to retrieve candidate views and objects, followed by VLM reasoning for verification. When mismatches are detected, slow reasoning refines the final selection, thereby enhancing overall navigation performance.

TABLE II

OBJECT RETRIEVAL COMPARISON WITH RSR@TOP1 AND RSR@TOP5 ON HM3DSEM. THE METRICS ARE VISUALIZED **GREEN** (BEST) TO **RED** (WORST).

Method	Time	RSR@Top1			RSR@Top5		
		1m	2m	3m	1m	2m	3m
HOVSG [20]	0.2s	0.52	0.64	0.70	0.76	0.82	0.88
osmAG-LLM [49]	3.0s	0.28	0.50	0.69	0.47	0.83	0.90
FSR-VLN (Ours)	5.5s	0.87	0.88	0.88	0.92	0.92	0.92

As illustrated in Table II, the Top-1 retrieval success rate of osmAG-LLM is significantly lower than that of HOVSG. The main reason is that osmAG-LLM loses the hierarchically rich visual open-vocabulary feature representation present in HOVSG and instead relies solely on textual recognition results. Although the osmAG-LLM semantic map is also hierarchical, it only preserves XML-level textual information without retaining visual CLIP features. In contrast, our method not only preserves CLIP-based visual embeddings but also further interacts with the original image information through VLM, leading to a notable improvement in retrieval success.

TABLE III

COMPARISON OF RSR@TOP1 PERFORMANCE ACROSS DIFFERENT FSR-VLN SETTINGS. THE HIGHEST VALUE IS HIGHLIGHTED IN **BOLD**.

Method	Time	RSR@Top1				
		1m	2m	3m	4m	5m
Ours (wo ST / wo NR)	1.5s	0.724	0.747	0.770	0.805	0.805
Ours (w ST / wo NR)	1.5s	0.816	0.862	0.885	0.908	0.908
Ours (w ST / w NR)	5.5s	0.920	0.943	0.943	0.966	0.966

Qualitative analysis. Fig. 4 illustrates the goal view and object retrieval results of FSR-VLN for different instructions in Room4. FSR-VLN successfully retrieves the goal view and object across four instruction types in long-range indoor environments.

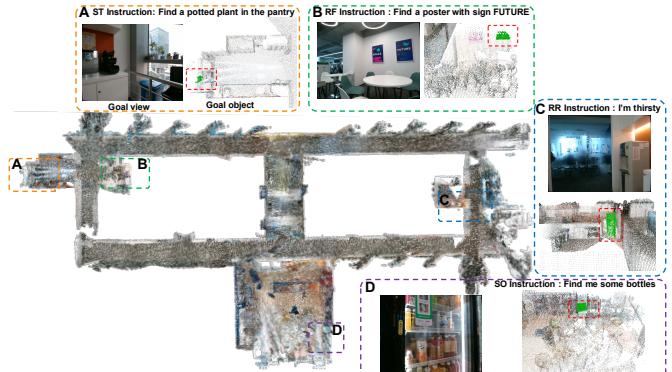


Fig. 4. The goal view and object retrieval results of FSR-VLN for four different instructions (Reasoning-Free, Reasoning-Required, Small Object, and Spatial Target) in Room4 (40mx20m).

Run-Time Analysis. MobilityVLA is the slowest, requiring approximately 30 s to reason over image view sequences. In contrast, OK-Robot and HOVSG compute text–object similarity using OWL-ViT and CLIP in only 0.2 s. FSR-VLN achieves an average response time of only 1.5 s when relying solely on goal inference and fast matching, and 5.5 s when incorporating slow reasoning. By invoking slow reasoning only when fast matching fails, and applying VLM refinement to candidate views rather than entire sequences, FSR-VLN reduces average response time by 82% compared with MobilityVLA while improving success rates. Within the integrated navigation system, FSR-VLN runs in parallel with user interactions, ensuring seamless and responsive operation.

C. Ablations Analysis

To further validate the effectiveness of HMSG representation and FSR, we examine the impact of spatial instructions and the FSR process on FSR-VLN performance, as summarized in Table III. Without **Navigation Reasoning (NR)** and **Spatial Target (ST)** instructions guide long-range navigation is guided by restricting object search to the target room, reducing global matching errors, and improving RSR. In long-range environments, this room-level guidance is particularly critical; when combined with the hierarchical scene graph, the search is further confined to the designated room, enhancing navigation success. With NR, extended VLM reasoning allows FSR to verify the correctness of fast matching, and VLM-based selection refinement further increases RSR, demonstrating the effectiveness of the approach.

V. CONCLUSION AND FUTURE WORK

We present FSR-VLN, a humanoid robotics VLN system that integrates a Hierarchical Multi-modal Scene Graph (HMSG) with Fast-to-Slow Navigation Reasoning (FSR). The HMSG encodes geometric, semantic, and topological relationships across floors, rooms, views, and objects, supporting FSR to fast retrieve candidate goals and refine them for more robust navigation. Experiments on real-world datasets demonstrate that FSR-VLN outperforms SOTA baselines in success rate, even under strict spatial constraints. These results highlight the effectiveness of HMSG and multi-stage FSR for robust navigation. Nevertheless, FSR-VLN has certain limitations. The construction of HMSG is time-consuming, making the approach unsuitable for real-time mapping, and it assumes static environments, limiting its applicability in dynamic settings. Our future work will address three key directions: enhancing the efficiency of scene graph construction, extending the system's robustness to dynamic environments, and integrating exploratory navigation capabilities to handle novel or ambiguous scenarios.

REFERENCES

- [1] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, “Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3674–3683.
- [2] N. Yokoyama, S. Ha, D. Batra, J. Wang, and B. Bucher, “Vlfm: Vision-language frontier maps for zero-shot semantic navigation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 42–48.
- [3] Y. Zhang, Z. Ma, J. Li, Y. Qiao, Z. Wang, J. Chai, Q. Wu, M. Bansal, and P. Kordjamshidi, “Vision-and-language navigation today and tomorrow: A survey in the era of foundation models,” 2024. [Online]. Available: <https://arxiv.org/abs/2407.07035>
- [4] Y. Long, W. Cai, H. Wang, G. Zhan, and H. Dong, “Instructnav: Zero-shot system for generic instruction navigation in unexplored environment,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.04882>
- [5] G. Zhou, Y. Hong, and Q. Wu, “Navgpt: Explicit reasoning in vision-and-language navigation with large language models,” 2023. [Online]. Available: <https://arxiv.org/abs/2305.16986>
- [6] S. K. Ramakrishnan, E. Wijmans, P. Krahenbuehl, and V. Koltun, “Does spatial cognition emerge in frontier models?” *arXiv preprint arXiv:2410.06468*, 2024.
- [7] S. Ruan, L. Wang, C. Kang, Q. Zhu, S. Liu, X. Wei, and H. Su, “From reactive to cognitive: brain-inspired spatial intelligence for embodied agents,” 2025. [Online]. Available: <https://arxiv.org/abs/2508.17198>
- [8] J. Gu, E. Stefani, Q. Wu, J. Thomason, and X. Wang, “Vision-and-language navigation: A survey of tasks, methods, and future directions,” in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. [Online]. Available: <http://dx.doi.org/10.18653/v1/2022.acl-long.524>
- [9] S. Peng, K. Genova, C. M. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, “Openscene: 3d scene understanding with open vocabularies,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.
- [10] X. Jin, M. Frosi, and M. Matteucci, “Openfusion++: An open-vocabulary real-time scene understanding system,” 2025. [Online]. Available: <https://arxiv.org/abs/2504.19266>
- [11] K. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omaha, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. Tenenbaum, C. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, “Conceptfusion: Open-set multimodal 3d mapping,” *Robotics: Science and Systems (RSS)*, 2023.
- [12] C. Liu, K. Wang, J. Shi, Z. Qiao, and S. Shen, “Fm-fusion: Instance-aware semantic mapping boosted by vision-language foundation models,” *IEEE Robotics and Automation Letters (RA-L)*, vol. 9, no. 3, pp. 2232–2239, 2024.
- [13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” 2021. [Online]. Available: <https://arxiv.org/abs/2103.00020>
- [14] M. Minderer, A. Gritsenko, A. Stone, M. Neumann, D. Weissenborn, A. Dosovitskiy, A. Mahendran, A. Arnab, M. Dehghani, Z. Shen, X. Wang, X. Zhai, T. Kipf, and N. Houlsby, “Simple open-vocabulary object detection with vision transformers,” 2022. [Online]. Available: <https://arxiv.org/abs/2205.06230>
- [15] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, “Language-driven semantic segmentation,” in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=RriDjddCLN>
- [16] D. Maggio, Y. Chang, N. Hughes, M. Trang, D. Griffith, C. Dougherty, E. Cristofalo, L. Schmid, and L. Carlone, “Clio: Real-time task-driven open-set 3d scene graphs,” *IEEE Robotics and Automation Letters*, vol. 9, no. 10, pp. 8921–8928, 2024.
- [17] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa, C. Gan, C. M. de Melo, J. B. Tenenbaum, A. Torralba, F. Shkurti, and L. Paull, “Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning,” 2023. [Online]. Available: <https://arxiv.org/abs/2309.16650>
- [18] H. Jiang, B. Huang, R. Wu, Z. Li, S. Garg, H. Nayyeri, S. Wang, and Y. Li, “Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation,” 2024. [Online]. Available: <https://arxiv.org/abs/2402.15487>
- [19] Y. Tang, M. Wang, Y. Deng, Z. Zheng, J. Deng, and Y. Yue, “Openin: Open-vocabulary instance-oriented navigation in dynamic domestic environments,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.04279>
- [20] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, “Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation,” *Robotics: Science and Systems*, 2024.
- [21] H. Chen, Y. Lin, Z. Li, B. Ye, Y. Zhang, and X. Lyu, “Irs: Instance-level 3d scene graphs via room prior guided lidar-camera fusion,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.06804>
- [22] M. Pekkanen, T. Mihaylova, F. Verdoja, and V. Kyrki, “Do visual-language grid maps capture latent semantics?” 2025. [Online]. Available: <https://arxiv.org/abs/2403.10117>
- [23] C. Huang, O. Mees, A. Zeng, and W. Burgard, “Visual language maps for robot navigation,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.
- [24] S. Raychaudhuri and A. X. Chang, “Semantic mapping in indoor embodied ai – a survey on advances, challenges, and future directions,” 2025. [Online]. Available: <https://arxiv.org/abs/2501.05750>
- [25] N. Savinov, A. Dosovitskiy, and V. Koltun, “Semi-parametric topological memory for navigation,” in *International Conference on Learning Representations*, 2018. [Online]. Available: <https://openreview.net/forum?id=SygwwGbRW>

- [26] D. Shah, A. Sridhar, A. Bhorkar, N. Hirose, and S. Levine, “Gnm: A general navigation model to drive any robot,” 2023. [Online]. Available: <https://arxiv.org/abs/2210.03370>
- [27] Z. Xu, H.-T. L. Chiang, Z. Fu, M. G. Jacob, T. Zhang, T.-W. E. Lee, W. Yu, C. Schenck, D. Rendleman, D. Shah, F. Xia, J. Hsu, J. Hoech, P. Florence, S. Kirmani, S. Singh, V. Sindhwani, C. Parada, C. Finn, P. Xu, S. Levine, and J. Tan, “Mobility VLA: Multimodal instruction navigation with long-context VLMs and topological graphs,” in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: <https://openreview.net/forum?id=JScswMfEQ0>
- [28] S. Garg, K. Rana, M. Hosseinzadeh, L. Mares, N. Sünderhauf, F. Dayoub, and I. Reid, “Robohop: Segment-based topological map representation for open-world visual navigation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 4090–4097.
- [29] Z. Li, X. Wu, H. Du, F. Liu, H. Nghiem, and G. Shi, “A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges,” *arXiv preprint arXiv:2501.02189*, 2025.
- [30] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, “Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation,” 2024. [Online]. Available: <https://arxiv.org/abs/2409.13682>
- [31] P. Liu, Y. Orru, J. Vakil, C. Paxton, N. Shafiullah, and L. Pinto, “Demonstrating ok-robot: What really matters in integrating open-knowledge models for robotics,” in *Robotics: Science and Systems XX*, ser. RSS2024. Robotics: Science and Systems Foundation, Jul. 2024. [Online]. Available: <http://dx.doi.org/10.15607/RSS.2024.XX.091>
- [32] S. Wen, Z. Zhang, Y. Sun, and Z. Wang, “Ovl-map: An online visual language map approach for vision-and-language navigation in continuous environments,” *IEEE Robotics and Automation Letters*, vol. 10, no. 4, pp. 3294–3301, 2025.
- [33] Z. Zhou, Y. Hu, L. Zhang, Z. Li, and S. Chen, “Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation,” 2025. [Online]. Available: <https://arxiv.org/abs/2506.06487>
- [34] 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.
- [35] A. Rosinol, A. Gupta, M. Abate, J. Shi, and L. Carlone, “3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans,” *arXiv preprint arXiv:2002.06289*, 2020.
- [36] N. Hughes, Y. Chang, and L. Carlone, “Hydra: A real-time spatial perception system for 3d scene graph construction and optimization,” *arXiv preprint arXiv:2201.13360*, 2022.
- [37] E. Greve, M. Büchner, N. Vödisch, W. Burgard, and A. Valada, “Collaborative dynamic 3d scene graphs for automated driving,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 118–11 124.
- [38] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, “Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.
- [39] Z. Yan, S. Li, Z. Wang, L. Wu, H. Wang, J. Zhu, L. Chen, and J. Liu, “Dynamic open-vocabulary 3d scene graphs for long-term language-guided mobile manipulation,” 2025. [Online]. Available: <https://arxiv.org/abs/2410.11989>
- [40] Y. Tang, M. Wang, Y. Deng, Z. Zheng, J. Deng, S. Zuo, and Y. Yue, “Openin: Open-vocabulary instance-oriented navigation in dynamic domestic environments,” *IEEE Robotics and Automation Letters*, no. 99, pp. 1–8, 2025.
- [41] J. Zhang, K. Wang, S. Wang, M. Li, H. Liu, S. Wei, Z. Wang, Z. Zhang, and H. Wang, “Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks,” *arXiv preprint arXiv:2412.06224*, 2024.
- [42] A. Anwar, J. Welsh, J. Biswas, S. Pouya, and Y. Chang, “Remembr: Building and reasoning over long-horizon spatio-temporal memory for robot navigation,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 2838–2845.
- [43] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. K. Wong, “Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation,” *arXiv preprint arXiv:2401.07314*, 2024.
- [44] S. Chen, P. He, J. Hu, Z. Liu, Y. Wang, T. Xu, C. Zhang, C. Zhang, C. An, S. Cai *et al.*, “Astra: Toward general-purpose mobile robots via hierarchical multimodal learning,” *arXiv preprint arXiv:2506.06205*, 2025.
- [45] S. Garg, K. Rana, M. Hosseinzadeh, L. Mares, N. Sünderhauf, F. Dayoub, and I. Reid, “Robohop: Segment-based topological map representation for open-world visual navigation,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 4090–4097.
- [46] S. Podgorski, S. Garg, M. Hosseinzadeh, L. Mares, F. Dayoub, and I. Reid, “Tango: Traversability-aware navigation with local metric control for topological goals,” in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025.
- [47] C. Zheng, W. Xu, Z. Zou, T. Hua, C. Yuan, D. He, B. Zhou, Z. Liu, J. Lin, F. Zhu, Y. Ren, R. Wang, F. Meng, and F. Zhang, “Fastliv02: Fast, direct lidar–inertial–visual odometry,” *IEEE Transactions on Robotics*, vol. 41, pp. 326–346, 2025.
- [48] Z. Gao, Z. Li, J. Wang, H. Luo, X. Shi, M. Chen, Y. Li, L. Zuo, Z. Du, Z. Xiao, and S. Zhang, “Funasr: A fundamental end-to-end speech recognition toolkit,” in *INTERSPEECH*, 2023.
- [49] K. Yadav, R. Ramrakhya, S. K. Ramakrishnan, T. Gervet, J. Turner, A. Gokaslan, N. Maestre, A. X. Chang, D. Batra, M. Savva, A. W. Clegg, and D. S. Chaplot, “Habitat-matterport 3d semantics dataset,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 4927–4936.
- [50] F. Xie, S. Schwertfeger, and H. Blum, “osmag-llm: Zero-shot open-vocabulary object navigation via semantic maps and large language models reasoning,” *arXiv preprint arXiv:2507.12753*, 2025.