HeckNav: Interpretable *He*uristic *K*nowledge *Nav*igation via Bayesian Probabilistic Inference

Anonymous CVPR submission

Paper ID *****

Abstract

Object goal navigation (ObjectNav) is a well defined and fundamental problem for robotic to navigate, that starting in a random point at unseen environment to a given object. Recent advanced study in ObjectNav uses Vision-Language Models (VLMs) or multimodal large language models(MLLMs) for object detection and decision-making. However, due to the "black-box" nature and implicit reasoning of VLMs, current ObjectNav systems often rely on the opaque internal capabilities of the models or exhaustive prompt engineering, lacking transparency and explicit planning. In this paper, we propose HeckNav, a zero-shot, heuristic-based knowledge navigation module that enhance VLMs with Knowledge Graph (KG) and Bayesian probabilistic inference. To enhance planning efficiency while maintaining a minimal footprint, we introduce a lightweight KG module that employ Bayesian inference to derive navigational plans based on the robot's observed environment. Furthermore, our framework provides unique level of interpretability by visualizing the agent's environmental awareness and its path-finding process through an abstract graph based on heuristic value preferences. Experimental results demonstrate that HeckNav achieves significant improvement on the HM3D and MP3D datasets especially on smaller VLM.

1. Introduction

"Knowledge is crude." - David Papineau

ObjectNav [4] requires an embodied agent to navigate toward a specific object category within an unseen environment. This task tests the agent's capacity for object detection, spatial awareness, sequential planning, and autonomous execution. The workflow of objectNav can be divided into:

- **Localization:** agent is aware of surrounding environment with semantic meaning.
- **Mapping:** agent remembers the area that has been explored and unexplored.
- **Planning:** agent makes plan based on the known information and area of interest for long-term.
- **Decision making:** agent use judgement to make the decision based on plan and surrounding area for the following step.
- **Goal detection:** the agent stops when the goal is reached or when it believes that the goal is reached.

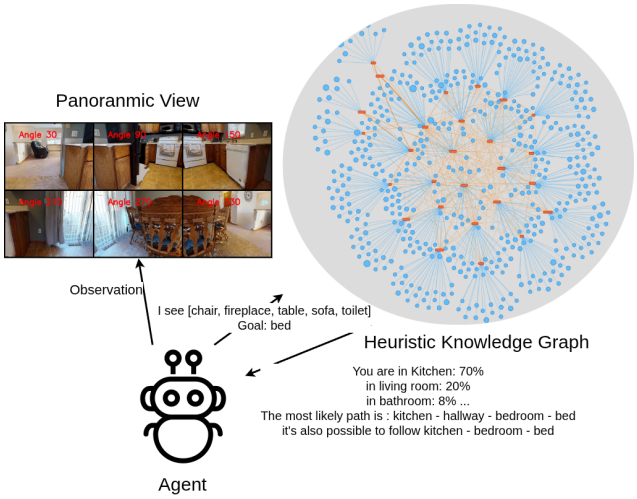

Figure 1. Overview of HeckNav module. It capture the statistical probability between regions and objects, make Bayesian inference over linguistic knowledge graph possible

In the very early stage navigation problem, the pioneer like DD-PPO [26] trains policies using reinforcement learning algorithms for decision making. Frameworks in the era doesn't come with planning process, they just merely predict next step based on the trained policy. Such method is demanding for training data, prolonged training times, and high computational resources.

Early milestones in objectNav, such as ZSON[19], use of multimodal embeddings, CLIP[23], to transfer policies trained on image navigation to open-vocabulary targets. However, ZSON relies heavily on implicit end-to-end rein-

forcement learning policies, which often struggle with long-horizon exploration and lack explicit spatial memory.

To address these exploration inefficiencies, subsequent works like VLFM[33] integrated simple VLM-driven semantic scoring with classical frontier-based mapping. Based on the map, they capture images and use BLIP-2[13] to get cosine-similarity score. The similarity score incorporates with the field of view to build value map. Using these data, adapt a pre-trained PointNav policy. However, this framework depends on BLIP to evaluate the similarity, which would cause disaster, as BLIP is not powerful enough to infer from an unrelated item to the goal.

Following the era of CLIP-based scoring, the rapid evolution of MLLMs such as ChatGPT and Gemini [6, 21], or complex question-answering VLMs like LLaVa and Qwen[16, 22], a profound paradigm shift in Embodied AI that makes vision-processing, planning, and decision making integrated into single model. Unlike simple VLMs limited to embedding alignments, MLLMs exhibit emergent reasoning, rich contextual understanding. On the other hand, simple VLMs with smaller parameters can be deployed on robots locally. However, the limited context window of models makes each step as an isolated query as the navigation horizon extends. VLM requires robust external memory mechanism.

With the MLLM advancement, WMNav[20] pushed the boundaries by framing the MLLM as a 'World Model'. By maintaining a Curiosity Value Map, WMNav enables the agent to simulate future states and evaluate potential outcomes before acting, thereby reducing trial-and-error. However, such world-model-based approaches implicitly rely on the MLLM's internal simulation capabilities, which incurs high computational overhead, suffers from hallucination risks, and lacks a transparent, symbolic representation of common sense. In the meanwhile, the performance of WMNav decrease dramatically when switch to small VLM, makes it impossible to deploy on local robot.

To bridge these gaps, our HeckNav framework introduces a highly optimized, lightweight knowledge graph synergized with Bayesian probabilistic inference. Instead of relying on expensive MLLM-based forward simulations, HeckNav leverages explicit, structured heuristic priors to logically deduce target locations (e.g., inferring a microwave's location based on a visible sink). This approach not only reduces computational burden but also provides an unique, interpretable visualization of the navigation reasoning process.

Extensive experiments on the standard HM3Dv1 and MP3D datasets demonstrate the superiority of HeckNav. Notably, our framework achieves an absolute improvement of 1.4% in Success Rate (SR) on HM3D, 3.1% SR on MP3D over the benchmark using Gemini-2.0-Flash and over 3.1% SR using Qwen7b/8b, while maintaining an ultra-low memory footprint of $< 2.2MB$

The main contributions of our work are summarized as follows:

- **A Novel Interpretable Module:** We introduce an interpretable visualization mechanism that maps the agent's spatial awareness and decision-making logic for long-term.
- **Extreme Memory Efficiency:** We introduce a highly optimized knowledge graph without the massive computational overhead.
- **SOTA Performance:** HeckNav improves success rates standard benchmarks (HM3D and MP3D) compared with SOTA framework especially on small VLM."

2. Related Work

ObjectNav Datasets. The development of Embodied AI is deeply rooted in the availability of high-fidelity simulation environments. A significant lineage of datasets, including **Gibson** [28], **Matterport3D (MP3D)** [3], **Replica** [2], **HM3D** [24], **HM3Dv2** [30] and **HM3D-OVON** [34], are reconstructed from real-world physical scans. Among these, **MP3D** is particularly distinguished by its exhaustive, human-annotated semantic metadata, providing granular ground-truth associations between object and room. Here, we conduct a systematic statistical representation of these human-annotated semantic distributions, transforming metadata into a probabilistic knowledge graph that guides the agent for long-term planning.

Foundation Models for Semantic Navigation. The integration of Large Language Models (LLMs) and MLLMs has revolutionized the decision-making pipeline in ObjectNav. Beyond the predictive world models like **WMNav** [20], several works explore the reasoning capabilities of foundation models. For instance, **L3MVN** [35] utilize LLMs to score frontiers by evaluating the semantic relevance of room-object associations. Another parallel research track focuses on persistent spatial representations. **VLMaps** [9] and **CoW** [5] store pixel-level VLM embeddings in a 2D grid, enabling open-vocabulary spatial querying. **InstructNav**[18], **NavCoT** [15] introduces a Chain-of-Thought prompting strategy for MLLMs to enhance step-by-step navigation reasoning. **VISTAv2**[10] uses diffusion models to evaluate the path quality by "imagine" the future scene. While these methods demonstrate impressive zero-shot capabilities, they often encounter a "performance deterioration" using smaller model, making it hard to deploy on local robot. **HeckNav** addresses these limitations by distilling these complex reasoning processes into a lightweight Bayesian Knowledge Graph, effectively bridging the gap between high-level MLLM perception and efficient, low-level symbolic execution for long-term planning.

Graph and Graph-Based Navigation. There are already plenty of study about the 3D-structure graph [1,

7, 12] that extract the semantic representation into abstract graph as prior. Utilizing structured prior to encode environmental common sense has long been a pursuit in Embodied AI. Frameworks such as AKGVP [29], NavRAG[25], and HOZPlus [37] have pioneered the integration of **offline** Knowledge Graphs (KGs) to provide the agent with a global semantic context. On the other hand, there are many **online** graph frameworks like TopoNav[17], TSGM[11], UniGoal[32], DGN[14]and SG-Nav[31] that served as memory mechanism. In this paper, the graph we discuss will be severed as ***offline prior***. There is a fundamental limitation of these existing KG-based frameworks is their reliance on binary or unweighted relational structures. While they successfully establish that certain objects "can" coexist (e.g., a stove is in a kitchen), they fail to explicitly model the *statistical strength* or *probabilistic tendency* of these relationships. In these graphs, all edges are treated with equal significance, neglecting the fact that some spatial correlations are far more predictive than others in real-world distributions. **HeckNav** addresses this gap by introducing a Bayesian probabilistic layer atop the knowledge graph. By deriving edge weights from the statistical analysis of human-annotated datasets like MP3D, our framework quantifies the "affinity" between entities, allowing the agent to prioritize exploration targets based on the rigorous mathematical likelihood rather than simple connectivity.

3. Method

3.1. Statistical Knowledge Grounding

To empower the agent with structured common sense, we construct a lightweight, multi-relational Knowledge Graph (KG) derived from the semantic metadata of the MP3D dataset. Unlike implicit end-to-end priors, our KG is explicitly built to encode both topological layouts and probabilistic semantic distributions.

3.1.1. Multi-relational Graph Construction

Formally, we define our Knowledge Graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V} = \mathcal{V}_R \cup \mathcal{V}_O$ consists of room nodes $\mathcal{V}_R$ (e.g., *bedroom, kitchen*) and object nodes $\mathcal{V}_O$ (e.g., *bed, cabinet*). The edge set $\mathcal{E} = \mathcal{E}_{nav} \cup \mathcal{E}_{prob}$ encompasses two distinct types of relationships:

- **Navigation Edges ($\mathcal{E}_{nav}$):** Undirected edges connecting adjacent room nodes based on spatial annotations (e.g., portals/doors). These define the topological navigability of the environment.
- **Probabilistic Semantic Edges ($\mathcal{E}_{prob}$):** Directed edges connecting object nodes to room nodes. The weight of these edges represents the statistical correlation probability, bridging semantic concepts with spatial containers.

3.1.2. Probabilistic Edge Formulation

To quantify the semantic edges $\mathcal{E}_{prob}$, we aggregate the spatial distributions from the raw metadata. Let $C(o, r)$ denote the total occurrences of object $o \in \mathcal{V}_O$ residing within room $r \in \mathcal{V}_R$. We formulate the conditional probability $P(r|o)$—the likelihood that an observed object o belongs to a room type r—using a normalized frequency distribution.

3.2. Heuristic Navigation via Graph Reasoning

During navigation, HeckNav dynamically queries this lightweight KG to form long-term reasoning chains. The decision-making process is divided into location belief updating and semantic path generation.

3.2.1. Zero-Shot Perception and Room Belief Update

At each navigation step t, the agent faces cognitive tasks: localizing its current semantic context. HeckNav accomplishes this by deriving two explicit probability distributions over all room categories $\mathcal{V}_R$, conditioned on real-time observations and the specified goal.

Current Room Distribution ($P(R_{curr})$): First, the VLM parses the current panoramic RGB observation to identify a set of salient anchor objects $\mathcal{O}_{obs}$. Instead of relying on geometric coordinates, the agent infers its current semantic location by aggregating the probabilistic evidence from these observed anchors. First, the model initialize the current room distribution by calculating:

$$P(R_i)_{initial} = \frac{f(R_i, \mathcal{O}_{obs})}{\sum_{j \in \mathcal{R}} f(R_j, \mathcal{O}_{obs})} \tag{1}$$

The function $f(R_i, o_{obs})$ represntes the likelihood of observing o_{obs} given that the agent is in room R_i. Based on the current belief distribution over all rooms, the scores are L1-normalized across all room nodes. After movement, the belief is updated as follows:

$$P(R_i \mid \mathcal{O}_{obs}) = \frac{P(\mathcal{O}_{obs} \mid R_i) P(R_i)}{P(\mathcal{O}_{obs})} \tag{2}$$

Here, the denominator is a dynamic normalization constant. It fluctuates based on the semantic density of the current observation.

$$P(\mathcal{O}_{obs}) = \sum_{j \in \mathcal{R}} P(\mathcal{O}_{obs} \mid R_j) P(R_j) \tag{3}$$

Goal Room Distribution ($P(R_{goal})$): Simultaneously, the agent must estimate the destination. Using the given target object o_{target}, we query the probabilistic semantic edges $\mathcal{E}_{prob}$ in our Knowledge Graph. The unnormalized belief for the target residing in room r is derived directly from the statistical prior: $S_{goal}(r) = P(r|o_{target})$. This is similarly normalized to yield the goal room distribution, $P(R_{goal})$.

By transforming both the observations and the text-specified goal into unified room-level probability distributions, HeckNav establishes the source and destination priors required for the subsequent semantic path generation.

3.2.2. Semantic Navigation Chain Generation

With the dual candidate sets $\mathcal{R}^*_{curr}$ and $\mathcal{R}^*_{goal}$ established, HeckNav shifts from perception to cognitive planning. We formulate this as a semantic path search problem over the topological navigation edges $\mathcal{E}_{nav}$ within our Knowledge Graph.

For any presumed source room $r_s \in \mathcal{R}^*_{curr}$ and goal room $r_g \in \mathcal{R}^*_{goal}$, a candidate semantic path $\Gamma = (r_s, r_1, \ldots, r_g)$ is defined as a sequence of topologically connected rooms. Let $N = |\Gamma|$ denote the number of hops (i.e., edges) in this path. A naive approach to evaluate the traversability of Γ would be the cumulative product of all edge probabilities. However, since transition probabilities strictly reside in $(0, 1]$, a pure multiplicative formulation inherently suffers from the *multi-hop attenuation problem*—it penalizes longer, multi-hop trajectories, biasing the agent toward strictly myopic, short-distance explorations.

To inherently decouple the semantic confidence from the path length, we introduce a length-normalized transition scoring function. By taking the N-th root of the cumulative product, we effectively compute the **geometric mean** of the transition probabilities. The transition score of a path Γ is formulated as:

$$S_{trans}(\Gamma) = \left(\prod_{i=1}^{N} P_{nav}(r_i | r_{i-1}) \right)^{\frac{1}{N}} \tag{4}$$

where $P_{nav}(r_i|r_{i-1})$ represents the navigation edge probability between adjacent rooms derived from the prior graph.

The overall confidence of a complete semantic chain $\mathcal{C}_{s \to g}$ is then synergized by integrating the initial location belief, the length-normalized path score, and the goal estimation:

$$S_{chain}(\mathcal{C}_{s \to g}) = P(r_s) \cdot S_{trans}(\Gamma(r_s \to r_g)) \cdot P(r_g) \tag{5}$$

The framework evaluates all possible paths between the Top-K source and goal candidates, selecting the chain with the highest overall score $\arg\max S_{chain}$. The agent then can decomposite the task and decide next step.

3.3. Transparent and Interpretable Reasoning

Unlike end-to-end black-box models or implicit world models, HeckNav's decision-making process is entirely transparent. Because both the environmental prior $\mathcal{G}$ and the dynamic semantic chains are explicitly formulated, the agent's cognitive state can be visualized in real-time. By overlaying the active room beliefs $P(R_{curr})$, $P(R_{goal})$, and the highest-scoring semantic path onto the topological graph, humans can intuitively trace the exact probabilistic reasoning that led to the agent's navigational choices. This explicit symbolic structure enhances the trust and debuggability of the Embodied AI system.

4. Experiment

4.1. Data and Evaluation Metrics

Knowledge Graph Optimization and Specifications We meticulously optimized the raw semantic categories extracted from the MP3D dataset. Architectural elements that are ubiquitous across all indoor environments—such as *walls, ceilings, floors, and doors*—are discarded for goal-oriented semantic reasoning. Consequently, we filtered out these non-discriminative entities into different categories during the graph construction phase.

The resulting refined Knowledge Graph is highly compact, comprising exactly 52 nodes: 26 room category nodes ($\mathcal{V}_R$) and 26 salient object category nodes ($\mathcal{V}_O$). These nodes are interconnected by 1,421 directed edges ($\mathcal{E}_{nav}$ and $\mathcal{E}_{prob}$), encapsulating both the topological navigation links and the probabilistic semantic affinities. In our implementation, the module that contains graph and retrieval function only occupies a mere $< 2.2MB$ of runtime memory and each inference step takes $< 2ms$ (module test on $i5 - 12600K$ with $32GB$ memory). This extremely low memory footprint validates our claim of HeckNav being an lightweight, plug-and-play module. The category we used for object nodes is based on **mpcat40**.

Dataset The first version of HM3D[24] comes with 2000 validation episodes and 6 goal object categories such as sofa, bed on 20 environments. MP3D[3] has 2195 validation episodes and 21 goal object categories on 11 scenes.

Metrics Like other frameworks, we adopt Success Rate (SR) and Success Rate Weighted by Inverse Path Length (SPL). SR is a robustness test and SPL is efficiency test

4.2. Experimental Setup

MLLM Perception Engine To execute the zero-shot semantic perception and anchor extraction, we intentionally deploy **Gemini 2.0-Flash** as our primary Vision-Language Model (VLM) backend as we explicitly avoid performance improvement by merely utilizing heavier, more powerful models such as Gemini 1.5-Pro or Gemini 3.0-Flash.

Implementation Details To evaluate the versatility and effectiveness of our proposed Bayesian Knowledge Graph, we implement **HeckNav** as a lightweight, plug-and-play cognitive enhancer built upon the state-of-the-art **WMNav** architecture.

Crucially, we do **not** alter or bypass WMNav's original computational pipeline, **nor** change any config from the WMNav. The core sensory modules and the complex curiosity-driven world-model simulations are fully re-

Method	VLM	Reason/Plan	MP3D		HM3D	
			SR (%) ↑	SPL (%) ↑	SR (%) ↑	SPL (%) ↑
CoW [5]	CLIP	FM	-	-	9.2	4.9
VLFM [33]	BLIP	FM + DDPPO	36.4	17.5	52.5	30.4
ESC [38]	GLIP	DeBERTa v3	28.7	14.2	39.2	22.3
L3MVN [35]	Mask R-CNN	RoBERTa-large	-	-	50.4	23.1
VoroNav [27]	BLIP	GPT-3.5	-	-	42.0	26.0
HOZ++ [37]	CNN+GCN	GPT-R	37.0	15.2	-	-
SG-Nav [31]	LLaVA-1.6 (7B)	GPT-4	40.2	16.0	54.0	24.9
WMNav [20]	Gemini-1.5-Pro	Gemini-1.5-Pro	**45.4**	**17.2**	**58.1**	**31.2**
WMNav [20]	Gemini-2.0-Flash	Gemini-2.0-Flash	43.3	15.3	54.1	28.5
HeckNav	Gemini-2.0-Flash	Gemini-2.0-Flash	**44.7**	**15.4**	**57.2**	**29.9**

Table 1. Comparison of performance on the MP3D and HM3D datasets. We seperate the models used for detection and planning.

tained to ensure a strictly fair baseline comparison. Instead, our HeckNav module is integrated orthogonally: while WMNav simulates implicit future states, our pre-computed topological and probabilistic graph acts as an explicit cognitive guide. By injecting our *Semantic Navigation Chains* into the agent's reasoning loop, HeckNav provides statistical priors that assist and constrain the heavy MLLM computations, effectively preventing the agent from hallucinating highly improbable spatial layouts.

4.3. Comparison with SOTA Methods

As HeckNav's KG extract from MP3D, we will compare the performance of HeckNav method on MP3D and HM3Dv1 dataset for generalizability in Table 1. Because Gemini-1.5-pro api is not access at the time of experiment, we pick a less powerful model to compare with WMNav. **FM** means **Frontier Map**, DDPPO means **Decentralized Distributed Proximal Policy Optimization**

The result shows that even with model degradation from 1.5-pro to 2.0-flash, HeckNav has a good improvement over WMNav, and still outperform most SOTA models. This results shows that explicit reasoning module would benefit the VLM's performance.

In addition to ultra-lightweight and fast inference, the module, based on Bayesian inference, can track the probability of possible paths and generate a colored graph reasoning process for explainability.

4.4. Robustness Study

4.4.1. Performance on Different Models and Datasets

As the ultimate goal is to deploy the agent on local, resource-constrained devices, we evaluate **HeckNav** against the **WMNav** baseline using Qwen-VL series in Table 2, across the HM3Dv1 and HM3Dv2 datasets. Traditional navigation metrics fail to analysis the model hallucination. Therefore, in this comparison, we introduce three additional perception-centric metrics: **Hallucination Rate (Hal.)**, **Precision (Prec.)**, and **F1-score**.

- **Hallucination Rate (Hal.):** Every VLMs have potential of visual hallucinations, often claiming to see a target object that is not present. As a result, the success rate won't achieve 100% as the robot already stop and claim a false positive stop.
- **Precision (Prec.):** Addition to Hallucination Rate, this metric measures that when robot claim it reaches a goal, what's the actual rate it has reach the goal.
- **F1-score:** By providing the harmonic mean of precision and recall, the F1-score offers a holistic evaluation of the agent's semantic perception capabilities during exploration.

By incorporating these metrics alongside SR and SPL, we prove that **HeckNav** not only improves physical navigation efficiency but enhances the semantic and cognitive reliability of open-source local models. We also find that use a more powerful VLM model will not only lead to higher success rate but also higher hallucination rate. [1].

4.4.2. HM3Dv2 Comparison

Due to curiosity about the performance of **HeckNav** on latest Qwen3-VL-8B-Instruct, we compare it with other SOTA models that has HM3Dv2 dataset tested in table 3. We find that even with local deployed Qwen3 8B model, the performance of HeckNav is still competitive to those GPT or Gemini based API-calling models. It outperforms those GPT-4 based framewoks on the success rate, and the SPL also achieves second place in the table. What's more, the entire framework takes about $23GB$ of VRAM including

[1] Due to computational constraints, these experiments are not exhaustive. However, both HeckNav and the baseline are tested on the exact same 500 episodes to ensure fairness.

Model	Method	HM3Dv1 [24]					HM3Dv2 [30]				
		SR↑	SPL↑	Hal.↓	Prec.↑	F1↑	SR↑	SPL↑	Hal.↓	Prec.↑	F1↑
Qwen-2.5(7B)	WMNav	39.2	18.4	**13.4**	74.5	51.4	51.6	21.1	8.0	86.6	64.7
	HeckNav	**45.2**	**20.5**	13.8	**76.6**	**56.9**	**53.6**	**22.9**	**6.0**	**89.9**	**67.2**
Qwen-3(8B)	WMNav	55.1	28.0	17.2	76.2	64.0	66.0	28.1	11.1	85.6	74.5
	HeckNav	**56.7**	**29.7**	**16.0**	**78.0**	**65.7**	**69.3**	**29.9**	**10.5**	**86.9**	**77.1**

Table 2. Performance evaluation using Qwen on HM3D. All metrics are reported as percentages (%). By using explicit reasoning with the HeckNav graph, the agent consistently achieves higher SR and semantical F1-scores are also notably outperform across different datasets.

deploy Qwen3-VL-8B, which makes it possible to transfer to local device later on.[2]

4.5. Failure Case Analysis

Despite the substantial performance gains, we conducted a thorough investigation into the failure cases of **HeckNav** to identify remaining challenges in semantic navigation. Our analysis reveals three primary modes of failure:

1. **Dataset Artifacts and Initialization Constraints:** In certain MP3D episodes, the agent spawns exterior to the building structure, devoid of any meaningful indoor semantic reference objects. This initialization fundamentally contradicts the core premise of the indoor ObjectNav task. Without valid visual anchors ($\mathcal{O}_{obs} = \emptyset$), our Bayesian graph struggles to formulate an initial semantic belief $P(R_{curr})$, inadvertently degrading the overall success rate for these outlier episodes.
2. **Local Trapping in Expansive Regions (Memory Mechanism Deficits):** When navigating massive, open-plan spaces (e.g., churches or grand halls), the agent may exhibits localized roaming behaviors, repeatedly visiting the same sub-regions before eventually breaking out of the loop. Although the agent often ultimately reaches the target, this redundant exploration severely penalizes the SPL metric.
3. **Out-of-Distribution (OOD) Semantic Contexts:** Certain episodes require navigating toward highly uncommon target objects (e.g., *gym equipment*) starting from atypical locations (e.g., a *lobby*). These rare semantic pairs possess near-zero co-occurrence probabilities in our statistical prior. Consequently, the agent may become trapped in the starting region, unable to establish a confident semantic chain. Notably, these OOD scenarios are also nearly impossible for baseline frameworks relying purely on zero-shot LLM/MLLM reasoning, highlighting a shared bottleneck in current Embodied AI.
4. **Complexity of Environment Detection:** The idea of knowledge graph is to separate rooms by the mostly correlated items detected. However boundaries between rooms are not clear (e.g. livingroom and kitchen may connected without walls), or the object behind doors or windows will be detected to undermine the result of localization.

This analysis suggests that future iterations could heavily benefit from an explicit **spatial memory mechanism**, where the agent dynamically updates its topological graph by masking previously visited nodes and penalizing the traversal weights of explored regions. Such an episodic memory would not only prevent redundant looping to drastically improve SPL, but also force the agent to explore novel frontiers when trapped in OOD semantic contexts.

[2] $23GB$ of VRAM is tested on RTX 6000 pro with 22% gpu_memory_utilization plus running simulation. It's not the exhaustive experiments that are conducted by running 50 instance at the same time.

Method	VLM (Largest)	SR (%) ↑	SPL (%) ↑
L3MVN [35]	Mask R-CNN	36.3	15.7
SG-Nav [31]	GPT-4	54.0	24.9
InstructNav [18]	GPT-4V	58.0	20.9
SGM [36]	GPT-4	60.2	**30.8**
DORAEMON [8]	Gemini-1.5-Pro	**66.5**	20.6
HeckNav (Ours)	Qwen3-VL (8B)	64.6	28.5

Table 3. Comparison with SOTA methods on the HM3Dv2.

5. Conclusion

In this paper, we introduce **HeckNav**, a zero-shot, neuro-symbolic navigation module that integrates explicit Bayesian probabilistic inference with MLLMs. By extracting priors into an ultra-lightweight Knowledge Graph module with , HeckNav effectively offloads complex spatial reasoning from the MLLM to generate deterministic, length-normalized semantic chains. Operating as a plug-and-play module over the state-of-the-art WMNav, our framework achieves up to a 3.1% absolute improvement in Success Rate using Gemini 2.0-Flash, and consistently outperform baseline (1.6% to 6.0% SR, 1.8% SPL) using local models using Qwen. Furthermore, our explicit graph formulation enables real-time visualization of path-finding decisions, leveraging the interpretability gap in current black-box Embodied AI systems.

References

[1] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5664–5673, 2019. 2

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 2

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017. 2, 4

[4] Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *In Neural Information Processing Systems (NeurIPS)*, 2020. 1

[5] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. *CVPR*, 2023. 2, 5

[6] Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. 2

[7] Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5021–5028. IEEE, 2024. 3

[8] Tianjun Gu, Linfeng Li, Xuhong Wang, Chenghua Gong, Jingyu Gong, Zhizhong Zhang, Yuan Xie, Lizhuang Ma, and Xin Tan. Doraemon: Decentralized ontology-aware reliable agent with enhanced memory oriented navigation, 2025. 6

[9] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023. 2

[10] Yanjia Huang, Xianshun Jiang, Xiangbo Gao, Mingyang Wu, and Zhengzhong Tu. Vistav2: World imagination for indoor vision-and-language navigation, 2025. 2

[11] Nuri Kim, Obin Kwon, Hwiyeon Yoo, Yunho Choi, Jeongho Park, and Songhawi Oh. Topological Semantic Graph Memory for Image Goal Navigation. In *CoRL*, 2022. 3

[12] Ue-Hwan Kim, Jin-Man Park, Taek-jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE Transactions on Cybernetics*, 50(12): 4921–4933, 2020. 3

[13] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 2

[14] Shiyao Li, Ziyang Meng, Jiansong Pei, Jiahao Chen, Bingcheng Dong, Guangsheng Li, Shenglan Liu, and Feilong Wang. Knowledge-driven visual target navigation: Dual graph navigation. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7547–7554, 2025. 3

[15] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[16] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[17] Peiran Liu, Qiang Zhang, Daojie Peng, Lingfeng Zhang, Yihao Qin, Hang Zhou, Jun Ma, Renjing Xu, and Yiding Ji. Toponav: Topological graphs as a key enabler for advanced object navigation, 2025. 3

[18] Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment, 2024. 2, 6

[19] Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In *Neural Information Processing Systems (NeurIPS)*, 2022. 1

[20] Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2392–2399, 2025. 2, 5

[21] OpenAI. Introducing GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025. 2

[22] Qwen Team. Qwen3.5: Towards native multimodal agents, 2026. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1

[24] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 2, 4, 6

[25] Zihan Wang, Yaohui Zhu, Gim Hee Lee, and Yachun Fan. NavRAG: Generating user demand instructions for embodied navigation through retrieval-augmented LLM. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8430–8440, Vienna, Austria, 2025. Association for Computational Linguistics. 3

[26] Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra.

Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames, 2020. 1

[27] Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*, 2024. 5

[28] Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson Env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018. 2

[29] Nuo Xu, Wen Wang, Rong Yang, Mengjie Qin, Zheyuan Lin, Wei Song, Chunlong Zhang, Jason Gu, and Chao Li. Aligning knowledge graph with visual perception for object-goal navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5214–5220. IEEE, 2024. 3

[30] Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3d semantics dataset. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4927–4936, 2023. 2, 6

[31] Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *arXiv preprint arXiv:2410.08189*, 2024. 3, 5, 6

[32] Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. *arXiv preprint arXiv:2503.10630*, 2025. 3

[33] Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *International Conference on Robotics and Automation (ICRA)*, 2024. 2, 5

[34] Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024. 2

[35] Bangguo Yu, Hamidreza Kasaei, and Ming Cao. L3mvn: Leveraging large language models for visual target navigation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, page 3554–3560. IEEE, 2023. 2, 5, 6

[36] Sixian Zhang, Xinyao Yu, Xinhang Song, Xiaohan Wang, and Shuqiang Jiang. Imagine before go: Self-supervised generative map for object goal navigation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16414–16425, 2024. 6

[37] Sixian Zhang, Xinhang Song, Xinyao Yu, Yubing Bai, Xinlong Guo, Weijie Li, and Shuqiang Jiang. Hoz++: Versatile hierarchical object-to-zone graph for object navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5958–5975, 2025. 3, 5

[38] Kaiwen Zhou, Kaizhi Zheng, Connor Pryor, Yilin Shen, Hongxia Jin, Lise Getoor, and Xin Eric Wang. Esc: Exploration with soft commonsense constraints for zero-shot object navigation. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. JMLR.org, 2023. 5