RWMNAV: RELATIONAL WORLD MODEL FOR COMPOSITIONAL ZERO SHOT OBJECT NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero shot object navigation requires an agent to leverage open vocabulary semantic knowledge to search for unseen target categories in unfamiliar environments. However, existing approaches typically represent navigation goals as a single object category, while natural language instructions in real world settings often specify the intended target instance through its relations with surrounding entities. Such relational constraints require the agent not only to recognize the target category, but also to reason about the compositional structure formed by the target, reference entities, and their relations, even before these entities become directly observable. To address this challenge, we propose RWMNav, a Relational World Model for compositional zero shot object navigation. RWMNav parses the language instruction into a relational goal graph and incrementally constructs an environmental relation graph from RGB-D observations. It then treats candidate frontiers as high level exploration actions and uses a training free relational world model instantiated with a frozen vision language model to predict the future relational states that may emerge along different exploration directions, including potential regions, entities, and their relations. The predicted relation graphs are matched with the relational goal graph to estimate goal compatibility, which is further combined with exploration utility and navigation cost for frontier selection. As exploration proceeds, predicted relational hypotheses are verified and corrected using newly acquired observations, forming a closed loop process of relational prediction, goal matching, exploration, and state update. Experiments on HM3D and MP3D demonstrate that RWMNav consistently improves relational navigation performance in terms of Relational Success Rate and Relational Success weighted by Path Length.

1 INTRODUCTION

Object Navigation (ObjectNav) requires an embodied agent to explore an unfamiliar environment, search for a specified target, and plan a feasible path from egocentric observations without access to a prior map, making it a fundamental task in embodied intelligence (Savva et al., 2019; Batra et al., 2020). Conventional approaches typically rely on large scale interactive training and learn navigation policies over a predefined closed set of object categories (Chaplot et al., 2020b;a; Ramakrishnan et al., 2022). With the rapid development of vision language models and open vocabulary perception, Zero Shot Object Navigation has emerged as a promising paradigm that leverages image text semantic alignment to recognize target categories unseen during training and incorporates open vocabulary semantic knowledge into spatial exploration (Radford et al., 2021; Li et al., 2022; Huang et al., 2023).

Existing methods primarily employ vision language models, large language models, and scene level commonsense knowledge to estimate where a target is likely to appear. ZSON and CoW exploit multimodal representations to match previously unseen targets (Majumdar et al., 2022; Gadre et al., 2023), while ESC and L3MVN leverage object co-occurrence patterns and functional knowledge of indoor spaces to guide exploration (Zhou et al., 2023; Yu et al., 2023). VLFM, VoroNav, and SG-Nav further enhance spatial reasoning by incorporating language conditioned value maps, topological structures, or scene graphs (Yokoyama et al., 2024; Wu et al., 2024; Yin et al., 2024). Despite these advances, most existing formulations still represent the navigation objective as a single object

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

category, under which the task is considered complete once the agent reaches any instance belonging to that category.

In real world environments, however, the intended target is often further constrained by its relations to surrounding entities. For example, “find the sofa” and “find the sofa next to the table” specify fundamentally different navigation goals. Similarly, instructions such as “the cabinet beside the bed” or “the chair in front of the window” require the agent to jointly reason about the target entity, reference entities, and the relations among them. In such cases, relational information is no longer merely auxiliary scene prior for facilitating search, but instead constitutes an essential part of the target specification itself. This challenge becomes particularly pronounced in partially observable environments, where neither the target nor its reference entities may be visible from the current viewpoint. The agent must therefore infer, from previously observed regions, objects, and spatial topology, what relational structures may emerge in different unexplored areas. Although open vocabulary scene graphs provide structured representations of observed entities and their relations (Gu et al., 2024; Koch et al., 2024; Werby et al., 2024), reasoning solely over the current state remains insufficient for determining which exploration direction is most likely to satisfy the relational goal.

World models provide a natural predictive framework for addressing this challenge (Hafner et al., 2019; Koh et al., 2021). Rather than predicting complete future visual observations, relational navigation is more concerned with task relevant future relational states namely, the regions the agent may enter, the entities it may encounter, and the potential relations among them after continuing exploration in a particular direction. Motivated by this perspective, we propose RWMNav. RWMNav first parses the language instruction into a relational goal graph and constructs a spatial semantic map together with an environmental relational state from RGB-D observations. It then treats candidate frontiers as high level exploration actions and employs a frontier conditioned relational world model to predict the future relational states associated with different exploration directions. The predicted states are subsequently matched against the relational goal graph to assess their structural compatibility. Finally, RWMNav jointly considers relational compatibility, exploration utility, and path cost to select the next navigation frontier, while continuously updating the relational state using newly acquired observations, thereby forming a closed loop process of prediction and decision making.

The main contributions of this work are as follows:

- We propose RWMNav, a unified relational world model framework for compositional zero shot object navigation, which integrates relational goal understanding, spatial semantic scene representation, and future state reasoning within a closed loop navigation pipeline.
- We design a frontier conditioned relational world model that predicts frontier specific future relational states, including potential regions, entities, and their relations, enabling anticipatory reasoning over unexplored space before the target relation is directly observed.
- We develop a goal conditioned future state matching strategy that evaluates candidate frontiers by aligning predicted relation graphs with the relational goal graph, and combines relational compatibility with exploration utility and path cost for navigation decision making.

2 Related Work

2.1 Zero shot Object Goal Navigation

Zero shot object navigation aims to enable an agent to search for previously unseen objects without relying on navigation training data for the target categories, by leveraging open vocabulary visual models and general semantic knowledge. In recent years, research in this area has gradually progressed from simple image text matching toward more sophisticated target understanding and multi source semantic planning. LGX exploits scene level commonsense knowledge encoded in large language models to infer promising directions in which a target may appear, and combines this reasoning with vision language models for open vocabulary target localization (Dorbala et al., 2024a). LOC-ZSON improves matching under complex target descriptions through language driven target retrieval and object centric visual representations (Guan et al., 2024). Pixel Navigator transforms the outputs of foundation models into pixel level navigation cues (Cai et al., 2024), while InstructNav integrates multiple value maps with dynamic navigation reasoning to support more general natural language guided navigation (Long et al., 2024). More recent studies have further enhanced naviga-

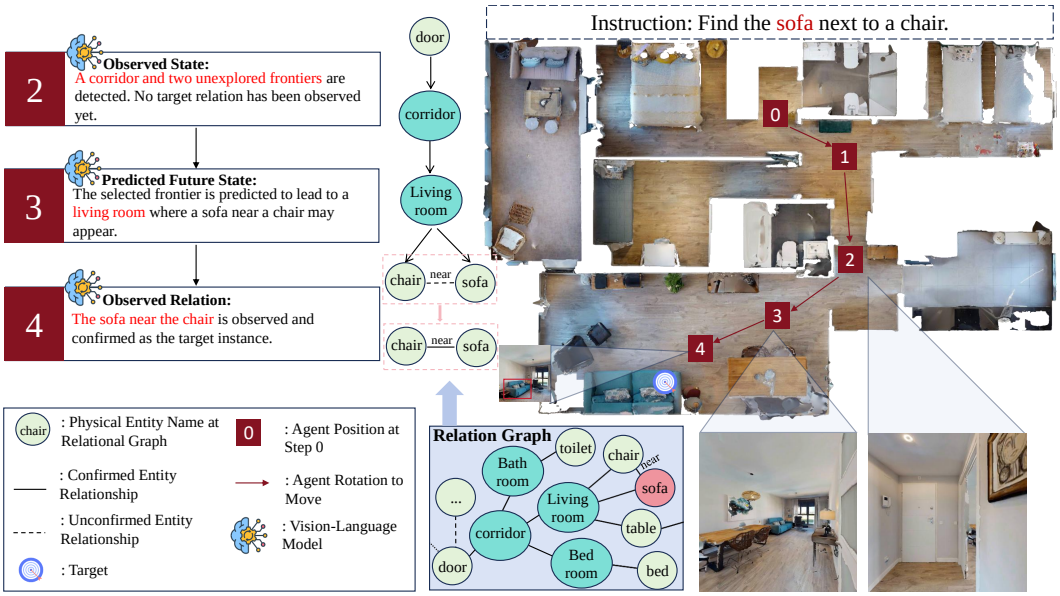


Figure 1: **Relational World Model Navigation with VLM**. In relational object navigation, our model first predicts the potential relational state behind unexplored frontiers (e.g., a living room may contain a sofa near a chair), then selects the frontier most likely to satisfy the relational goal and updates the relation graph with newly observed evidence during navigation.

tion in open environments from the perspectives of dynamic scene adaptation, semantic verification, and long term memory (Ma et al., 2024; Unlu et al., 2024; Dorbala et al., 2024b; Yu et al., 2025). Despite these advances in target understanding and exploration for zero shot navigation, existing ObjectNav formulations still typically adopt a single object category or target semantic concept as the primary navigation condition, while inter object relations are largely treated as auxiliary scene knowledge for estimating likely target locations. In real world environments, however, the intended target instance often can only be specified accurately through its relations with surrounding entities, as in "the cabinet beside the bed" or "the sofa next to the table." This setting requires a navigation agent not only to recognize the target category, but also to jointly reason about the target entity, reference entities, and their compositional relations, while exploiting these relational constraints to guide subsequent exploration even before the relevant entities have been observed.

2.2 World Model

World models construct internal predictive representations of the environment, enabling an agent to anticipate potential future states before executing actions in the physical world and thereby supporting prospective planning and decision making. Classical world models typically learn environmental state transitions and perform imagination in a latent space (Hafner et al., 2019), while PathDreamer extends this paradigm to indoor navigation by predicting plausible future scenes to facilitate long horizon decision making (Koh et al., 2021). For zero shot object navigation, however, predicting abstract environment states that are directly relevant to target search is often more valuable than generating complete future visual observations. Such states may include the types of rooms and entities that are likely to exist in an unexplored region, as well as the potential relations among them.

3 Problem Formulation

We further consider the setting in which a specific target instance is identified through entity relations expressed in natural language. The agent is deployed in an unfamiliar indoor environment, where it continuously explores using egocentric RGB-D observations together with pose information relative

to the starting position. From the language instruction, it constructs a relational goal composed of the target entity, reference entities, and their spatial or semantic relations. Before the target and its associated entities have been fully observed, the agent must infer the entity relation structures that may exist in different unexplored regions and ultimately navigate to the correct target instance that simultaneously satisfies both the target category requirement and the specified relational constraints, while minimizing unnecessary exploration and travel distance.

4 RWMNAV METHODOLOGY

4.1 OVERVIEW

The overall architecture of RWMNav is illustrated in Fig 2. At each time step t, the agent receives an egocentric RGB-D observation, its current pose, and a relational goal instruction. The language instruction is first parsed into a relational goal graph composed of the target entity, reference entities, and the relations among them. The system then projects visual semantic information onto a top-down spatial semantic map, extracts candidate frontiers along the boundary between explored and unexplored regions, and simultaneously maintains an environmental relational state consisting of observed entities, regions, and their interrelations. Building on these representations, RWMNav treats each candidate frontier as a high level exploration action and employs a frontier conditioned relational world model to predict the future relational states that may emerge after exploring along different frontiers, including potential rooms, entities, and their relational structures. The model subsequently evaluates the correspondence between each predicted relational state and the relational goal graph, and combines this relational compatibility with path cost and exploration utility to derive the final frontier score. The agent then navigates toward the highest scoring candidate frontier and continuously updates both the spatial semantic map and the environmental relational state using newly acquired observations.

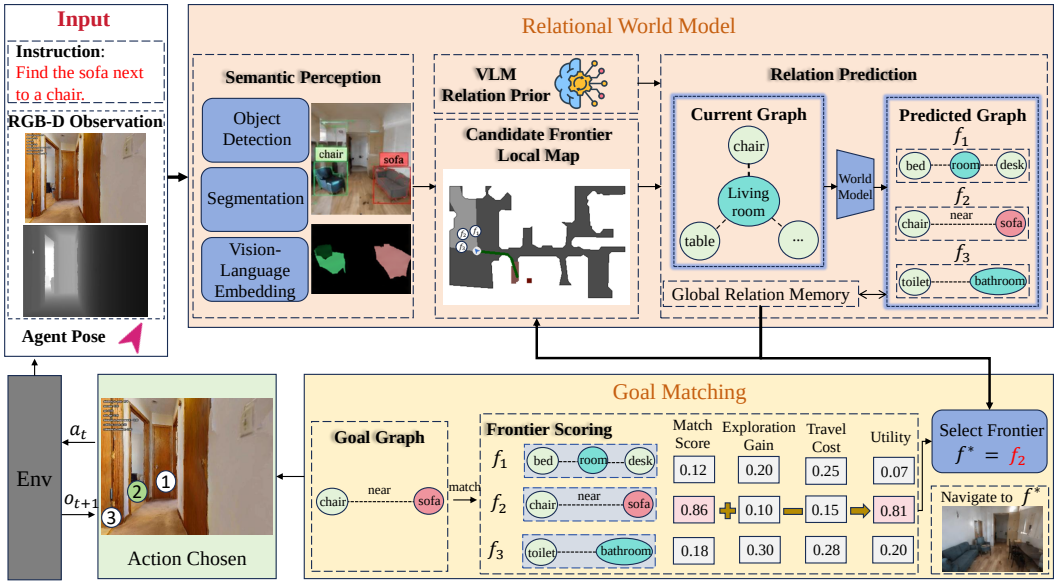


Figure 2: **The RWMNav framework**. RWMNav first constructs semantic and relational representations from the RGB-D observations, and predicts a future relation graph for each candidate frontier using the relational world model. The predicted graphs are matched with the goal graph, and the agent selects the frontier with the highest utility by jointly considering relational relevance, exploration gain, and travel cost.

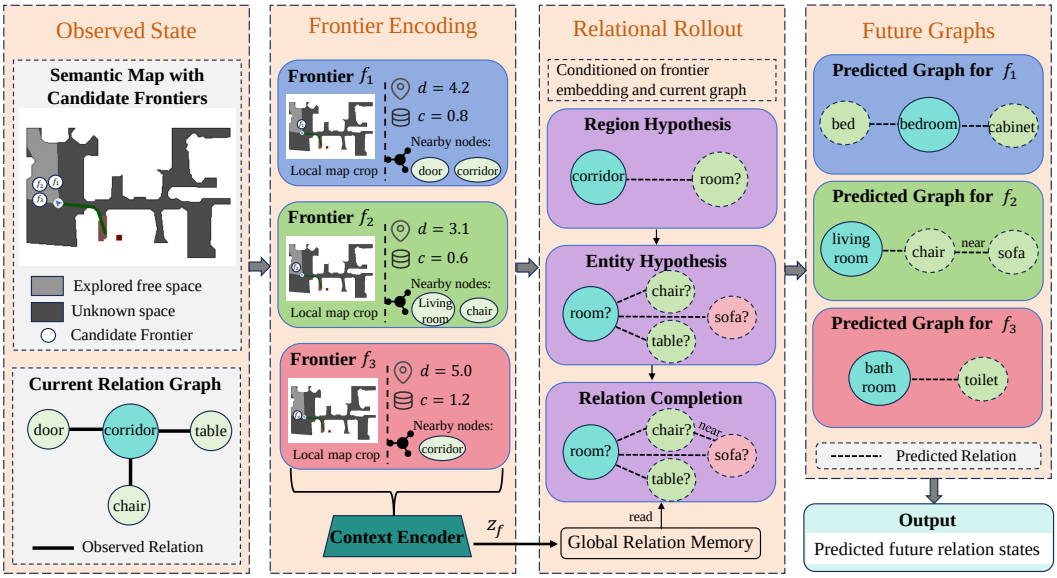


Figure 3: **Frontier conditioned relation prediction**. RWMNav encodes the local context of each candidate frontier and performs relational rollout over the current relation graph. The model progressively hypothesizes potential regions, entities, and their relations, producing a future relation graph for each frontier.

4.2 Relational Goal and Spatial Representation

4.2.1 Relational Goal Graph

Given a natural language navigation instruction q, RWMNav first identifies the target entity, reference entities, and their relational constraints, and constructs a relational goal graph $\mathcal{G}^q = (V^q, E^q)$. The node set V^q represents the target and reference entities involved in the instruction, while the edge set E^q encodes the spatial or semantic relations among them. For example, for the instruction "find the cabinet beside the bed," the goal graph contains two entity nodes, cabinet and bed, connected by a beside relation. For compositional descriptions involving multiple reference entities, the model can further construct a multi node, multi relation goal structure. This relational graph serves as the task condition throughout the navigation process, extending the agent's objective from matching a single object category to satisfying the complete entity relation structure specified by the instruction.

4.2.2 Relational Goal Graph

During environment exploration, RWMNav extracts open vocabulary semantic features from egocentric RGB observations and combines them with depth information to construct local 3D point clouds. Multi view observations are projected into a unified world coordinate system using the agent pose, forming a top-down spatial semantic map that jointly represents occupancy and semantic information. Detected entities serve as semantic anchors for subsequent relational modeling.

Meanwhile, candidate frontiers are extracted and clustered along the boundary between explored free space and unexplored regions:

$$\mathcal{F}_t = \{f_t^1, f_t^2, \ldots, f_t^N\}. \tag{1}$$

Each frontier contains its spatial location, navigation distance, and nearby semantic context. We therefore regard candidate frontiers as high level exploration actions, based on which the relational world model predicts the possible relational structures behind unexplored regions.

4.3 Relational World Model

Conventional zero shot navigation methods typically evaluate candidate regions directly from current observations or semantic maps, making it difficult to explicitly reason about environmental structures that may emerge after exploring a particular direction. To address this limitation, RWMNav maintains a dynamic relational representation of the observed environment and predicts the future relational states associated with different candidate frontiers. Rather than generating future RGB observations, the model focuses on abstract regions, entities, and relations that are directly relevant to navigation.

4.3.1 Modeling of Environmental Relational States

At time step t, RWMNav represents the currently observed environment as a dynamic relation graph:

$$\mathcal{G}_t = (V_t, E_t), \tag{2}$$

where V_t contains observed object and region nodes, and E_t describes the spatial and semantic relations among them. For each node v_i, its state is represented as

$$h_i = [e_i, p_i, c_i], \tag{3}$$

where e_i denotes its open vocabulary semantic representation, p_i encodes spatial or regional information, and c_i denotes the observation confidence. Relational edges describe spatial and semantic relations such as *near*, *beside*, *inside*, *contain*, and *connected-to*.

The relation graph is incrementally updated by integrating newly acquired RGB-D observations. Importantly, RWMNav only stores relational states supported by actual observations in $\mathcal{G}_t$. Objects, rooms, and relations that have not yet been observed are instead treated as latent future states. This separation prevents uncertain predictions from directly affecting the persistent relational memory and provides a reliable basis for prospective reasoning over unexplored regions.

4.3.2 Frontier Conditioned Future Relational Rollout

Given the current relation graph $\mathcal{G}_t$, spatial semantic map M_t, and a candidate frontier f, the relational world model predicts the future relational state that may emerge after exploring along that frontier:

$$\hat{\mathcal{G}}^{f}_{t+\tau} = W_\theta\left(\mathcal{G}_t, M_t, f\right), \tag{4}$$

where τ denotes the navigation horizon associated with the frontier level exploration action. The model combines the semantic context surrounding the frontier with the current relational state to infer the regions, entities, and relations that may become observable after exploration.

The predicted future relational state is represented as

$$\hat{\mathcal{G}}^{f}_{t+\tau} = \left(\hat{V}^{\mathrm{reg}}_f, \hat{V}^{\mathrm{obj}}_f, \hat{E}^{\mathrm{rel}}_f\right), \tag{5}$$

where $\hat{V}^{\mathrm{reg}}_f$ and $\hat{V}^{\mathrm{obj}}_f$ denote predicted region and object nodes, respectively, while $\hat{E}^{\mathrm{rel}}_f$ represents their potential relational structure. Different frontiers therefore induce different hypotheses about the future environment, allowing the agent to compare possible relational outcomes before physically exploring them. For example, one frontier may be predicted to lead to a bedroom containing a cabinet beside a bed, while another may lead to a living room containing a sofa near a table.

4.4 Relational Goal Matching and Frontier Decision Making

After predicting future relational states for candidate frontiers, RWMNav evaluates how well each prediction satisfies the relational goal specified by the language instruction. Unlike conventional

ObjectNav methods that mainly evaluate semantic relevance to a single target category, our method jointly considers the target entity, reference entities, and their relational constraints.

4.4.1 Future Relational State Matching

Given the relational goal graph $\mathcal{G}^q$ and the predicted future relation graph $\hat{\mathcal{G}}^f_{t+\tau}$, RWMNav computes a relational compatibility score:

$$S_{\text{rel}}(f) = \text{Match}\left(\mathcal{G}^q, \hat{\mathcal{G}}^f_{t+\tau}\right). \tag{6}$$

The matching process jointly considers entity level semantic compatibility and relation level structural consistency:

$$S_{\text{rel}}(f) = \alpha S_{\text{ent}}(f) + \beta S_{\text{relation}}(f), \tag{7}$$

where $S_{\text{ent}}(f)$ measures the semantic compatibility of the target and reference entities, while $S_{\text{relation}}(f)$ evaluates whether their predicted relations satisfy the relational constraints in the instruction.

This design prevents a frontier from receiving a high score merely because it is likely to contain an instance of the target category. Instead, a frontier is preferred when both the predicted entities and their relational configuration are consistent with the complete relational goal.

4.4.2 Frontier Evaluation and Online Relational Update

A candidate region with high relational compatibility is not necessarily the optimal exploration direction, since it may require a long navigation path or provide limited exploration benefit. RWMNav therefore combines relational goal satisfaction, exploration utility, and navigation cost when evaluating candidate frontiers:

$$S(f) = \lambda_r S_{\text{rel}}(f) + \lambda_e S_{\text{exp}}(f) - \lambda_c C(f), \tag{8}$$

where $S_{\text{rel}}(f)$ denotes relational goal compatibility, $S_{\text{exp}}(f)$ represents the expected exploration utility of the candidate region, and $C(f)$ denotes the path cost from the current agent position to frontier f. The frontier with the highest score is selected as the next high level exploration target.

After reaching the selected region, the agent acquires new RGB-D observations and updates both the spatial semantic map and the environmental relation graph with newly observed regions, entities, and relations. Previously predicted relational hypotheses are subsequently verified or corrected according to actual observations rather than being directly written into the persistent graph. Through this iterative process, RWMNav forms a closed loop of future relation prediction, relational goal matching, physical exploration, and relational state update.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate the relevant methods on three large scale benchmarks in the Habitat simulator: HM3Dv1, based on HM3D-Semantics-v0.1 (Yadav et al., 2022), with 1,462 relational episodes across 20 scenes and 6 target categories; HM3Dv2, based on HM3D-Semantics-v0.2 (Yadav et al., 2023), with 782 relational episodes across 36 scenes and 6 target categories; and MP3D, based on Matterport3D (Chang et al., 2017), with 1,476 relational episodes across 11 scenes and 21 target categories. All retained episodes contain valid relational constraints that uniquely identify the target instance.

Evaluation Metrics. We report Relational Success Rate (RSR) and Relational Success weighted by Path Length (RSPL) as the primary evaluation metrics. RSR measures the proportion of episodes in which the agent successfully reaches the target instance that satisfies all relational constraints

Table 1: Comparison with state-of-the-art methods on relational zero shot object navigation. RSR and RSPL denote relational success rate and relational success weighted by path length, respectively.

Method	Zero shot	HM3Dv1		HM3Dv2		MP3D	
		RSR↑	RSPL↑	RSR↑	RSPL↑	RSR↑	RSPL↑
ESC	✓	34.7	16.2	40.8	18.6	10.8	4.1
L3MVN	✓	36.9	17.8	42.6	19.7	11.6	4.5
OpenFMNav	✓	39.5	19.7	45.2	21.5	12.7	4.9
VoroNav	✓	40.2	20.7	46.1	22.3	13.0	5.1
InstructNav	✓	42.5	20.3	48.7	22.0	14.2	5.3
VLFM	✓	41.3	21.0	47.5	22.8	13.8	5.4
ApexNav	✓	42.9	21.8	48.4	23.4	14.3	5.5
WMNav	✓	44.8	22.8	51.5	24.9	15.7	6.1
SG-Nav	✓	<u>46.7</u>	<u>23.9</u>	<u>54.0</u>	<u>26.2</u>	<u>16.8</u>	<u>6.6</u>
RWMNav (Ours)	✓	**51.8**	**27.1**	**61.0**	**30.3**	**19.5**	**7.7**

specified by the instruction. RSPL further evaluates navigation efficiency by weighting relational success with the ratio between the shortest path distance and the actual trajectory length. An episode receives zero RSPL if the relational goal is not successfully satisfied. Higher RSR and RSPL indicate better relational navigation accuracy and path efficiency, respectively.

Implementation Details. All experiments are conducted in Habitat with a discrete action space using a $0.25\,\mathrm{m}$ forward step and 30° rotation/pitch increments, with at most 500 steps per episode. To construct relational episodes, we enumerate reachable target instances and annotated reference entities in each scene, and derive valid relational tuples from the 3D semantic annotations using predefined geometric rules. Relational instructions are generated from fixed templates and retained only when the specified relations uniquely identify one target instance; ambiguous or invalid episodes are discarded. An episode is successful only when the agent stops within $0.2\,\mathrm{m}$ of this designated target instance. GroundingDINO and Mobile SAM are used for open vocabulary perception and map construction, while Qwen2.5-VL-Instruct-7B parses instructions into relational goal graphs. The relational world model predicts future relational states for candidate frontiers and selects the next frontier through goal graph matching.

5.2 Comparison with SOTA Methods

We compare RWMNav with representative zero shot object navigation methods on HM3Dv1, HM3Dv2, and MP3D. As shown in Table 1, RWMNav consistently achieves the best performance across all three benchmarks. Compared with SG-Nav, the strongest baseline, RWMNav improves RSR by 5.1, 7.0, and 2.7 percentage points on HM3Dv1, HM3Dv2, and MP3D, respectively, while improving RSPL by 3.2, 4.1, and 1.1 percentage points.

Most existing methods rely on semantic priors, value maps, or currently observed scene structures to guide exploration. Although these approaches provide effective semantic guidance, they mainly reason about the current environment and do not explicitly model the relational structures that may appear after exploring an unseen region. In contrast, RWMNav predicts frontier conditioned future relational states and evaluates their consistency with the relational goal graph before execution. This enables the agent to select exploration directions that are more likely to lead to the target instance satisfying the complete relational constraints, resulting in better navigation success and path efficiency.

5.3 Ablation Study

Effect of Different Modules. We conduct ablation experiments on HM3Dv2 to evaluate the contribution of the main components of RWMNav. As shown in Table 2, removing the Relational Goal Graph results in the largest performance drop, decreasing RSR from 61.0 to 46.2 and RSPL from 30.3 to 22.0. Removing the Relational World Model also causes a substantial degradation, demon-

Table 2: Ablation study of RWMNav on HM3Dv2.

Method	RSR↑	RSPL↑
w/o Relational Goal Graph	46.2	22.0
w/o Relational World Model	49.8	23.8
w/o Global Relation Memory	56.2	27.1
w/o Goal Graph Matching	52.5	24.9
w/o Exploration Utility	59.0	28.4
RWMNav	**61.0**	**30.3**

Table 3: Performance of RWMNav under different relational goal types on HM3Dv2.

Relation Type	RSR↑	RSPL↑
Next to / Beside	66.4	33.7
Left / Right of	63.1	31.9
Inside / Contain	61.8	31.1
In front of / Behind	59.7	29.4
Between	54.0	25.4
Overall	**61.0**	**30.3**

strating the importance of predicting future relational states in unexplored regions. Goal Graph Matching further contributes 8.5 points in RSR and 5.4 points in RSPL, confirming the effectiveness of explicitly matching predicted relational structures with the navigation goal.

Global Relation Memory further improves RSR and RSPL by preserving relational information accumulated during exploration, while Exploration Utility mainly improves navigation efficiency by balancing relational relevance with exploration benefit and travel cost. Overall, the full RWMNav achieves the best performance, demonstrating that these components provide complementary contributions to relational navigation.

5.4 Analysis of Different Relation Types

To further analyze the relational reasoning capability of RWMNav, we evaluate its performance under different types of relational goals on HM3Dv2. We group the instructions according to their primary spatial relations, including *next to/beside*, *left/right of*, *inside/contain*, *in front of/behind*, and *between*. As shown in Table 3, RWMNav achieves consistent performance across different relational constraints. The model performs particularly well on *next to/beside* and containment relations, where local entity configurations provide relatively strong relational cues. In contrast, *between* is more challenging because successful navigation requires jointly reasoning over multiple reference entities and their spatial configuration. Nevertheless, RWMNav maintains competitive performance across all relation types, demonstrating its ability to generalize relational world modeling to diverse compositional navigation goals.

6 Conclusion

We propose RWMNav, a relational world model framework for compositional zero shot object navigation. RWMNav represents language instructions as relational goal graphs and predicts future relational states for candidate frontiers, enabling the agent to reason about potential entities and spatial relations before direct observation. The predicted relation graphs are matched with the goal graph and combined with exploration utility and travel cost for navigation decision making. Experiments on HM3Dv1, HM3Dv2, and MP3D demonstrate the effectiveness of RWMNav, while ablation studies further verify the contributions of relational goal modeling, future relation prediction, and goal matching. These results show that reasoning over future relational structures effectively extends zero shot object navigation beyond category level search toward compositional relational goals.

AI USE STATEMENT

Generative AI tools were used to assist with refining the conceptual framework, translation and language editing, manuscript organization, LaTeX formatting, and the implementation of part of the experimental code. All AI-assisted content was manually reviewed and verified by the authors. All reported quantitative results were obtained from actual experiments. The authors take full responsibility for the final content of this work, including all text, claims, results, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the implementation and evaluation details necessary to reproduce our experiments throughout the paper and supplementary material. Section 5 describes the datasets, evaluation metrics, navigation settings, perception models, and experimental protocol. Appendix A provides additional details on relational benchmark construction, geometric relation definitions, the training free relational world model, relational goal matching, and the complete RWMNav inference procedure. Additional analyses and navigation visualizations are provided in Appendix B. All compared methods are evaluated using the same relational episode splits and evaluation criteria.

REFERENCES

Dhruv Batra, Aaron Gokaslan, and etc Aniruddha Kembhavi. Objectnav revisited: On evaluation of embodied agents navigating to objects, 2020. URL https://arxiv.org/abs/2006.13171.

Wenzhe Cai, Siyuan Huang, Guangran Cheng, Yuxing Long, Peng Gao, Changyin Sun, and Hao Dong. Bridging zero-shot object navigation and foundation models through pixel-guided navigation skill. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5228–5234. IEEE, 2024.

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.

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)*, 2020a.

Devendra Singh Chaplot, Dhiraj Gandhi, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Learning to explore using active neural slam. In *International Conference on Learning Representations (ICLR)*, 2020b.

Vishnu Sashank Dorbala, James F. Mullen, and Dinesh Manocha. ¡i¿can an embodied agent find your "cat-shaped mug"?¡/i¿ llm-based zero-shot object navigation. *IEEE Robotics and Automation Letters*, 9(5):4083–4090, May 2024a. ISSN 2377-3774. doi: 10.1109/lra.2023.3346800. URL http://dx.doi.org/10.1109/LRA.2023.3346800.

Vishnu Sashank Dorbala, Vishnu Dutt Sharma, Pratap Tokekar, and Dinesh Manocha. Improving zero-shot objectnav with generative communication. *arXiv preprint arXiv:2408.01877*, 2024b.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23171–23181, June 2023.

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)*, pp. 5021–5028. IEEE, 2024.

Tianrui Guan, Yurou Yang, Harry Cheng, Muyuan Lin, Richard Kim, Rajasimman Madhivanan, Arnie Sen, and Dinesh Manocha. Loc-zson: Language-driven object-centric zero-shot object retrieval and navigation, 2024. URL https://arxiv.org/abs/2405.05363.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ArXiv*, abs/1912.01603, 2019. URL https://api.semanticscholar.org/CorpusID:208547755.

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.

Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14738–14748, October 2021.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RriDjddCLN.

Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment, 2024. URL https://arxiv.org/abs/2406.04882.

Ji Ma, Hongming Dai, Yao Mu, Pengying Wu, Hao Wang, Xiaowei Chi, Yang Fei, Shanghang Zhang, and Chang Liu. Doze: A dataset for open-vocabulary zero-shot object navigation in dynamic environments. *IEEE Robotics and Automation Letters*, 9(9):7389–7396, 2024. doi: 10.1109/LRA.2024.3426381.

Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 32340–32352. Curran Associates, Inc., 2022. doi: 10.52202/068431-2343.

Alec Radford, Jong Wook Kim, and etc Hallacy, Chris. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Santhosh K. Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. Poni: Potential functions for objectgoal navigation with interaction-free learning. In *Computer Vision and Pattern Recognition (CVPR), 2022 IEEE Conference on*. IEEE, 2022.

Manolis Savva, Abhishek Kadian, and etc Maksymets, Oleksandr. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Halil Utku Unlu, Shuaihang Yuan, Congcong Wen, Hao Huang, Anthony Tzes, and Yi Fang. Reliable semantic understanding for real world zero-shot object goal navigation, 2024. URL https://arxiv.org/abs/2410.21926.

Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical Open-Vocabulary 3D Scene Graphs for Language-Grounded Robot Navigation. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024. doi: 10.15607/RSS.2024.XX.077.

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. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53757–53775, 2024.

Karmesh Yadav, Santhosh Kumar Ramakrishnan, John Turner, and etc Aaron Gokaslan. Habitat challenge 2022. https://aihabitat.org/challenge/2022/, 2022.

Karmesh Yadav, Jacob Krantz, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Jimmy Yang, Austin Wang, John Turner, Aaron Gokaslan, Vincent-Pierre Berges, Roozbeh Mootaghi, Oleksandr Maksymets, Angel X Chang, Manolis Savva, Alexander Clegg, Devendra Singh Chaplot, and Dhruv Batra. Habitat challenge 2023. https://aihabitat.org/challenge/2023/, 2023.

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. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 5285–5307. Curran Associates, Inc., 2024. doi: 10.52202/079017-0171. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/098491b37deebbe6c007e69815729e09-Paper-Conference.pdf.

Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48, 2024. doi: 10.1109/ICRA57147.2024.10610712.

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)*, pp. 3554–3560, 2023. doi: 10.1109/IROS55552.2023.10342512.

Ming-Ming Yu, Fei Zhu, Wenzhuo Liu, Yirong Yang, Qunbo Wang, Wenjun Wu, and Jing Liu. C-nav: Towards self-evolving continual object navigation in open world. *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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*, pp. 42829–42842, 2023.

A Implementation Details

In this section, we provide additional implementation details for the construction of the relational navigation benchmark and the inference procedure of RWMNav. The notation follows the problem formulation and method description in the main text. We first describe how relational targets and instructions are generated from the original ObjectNav benchmarks, and then elaborate on the training free relational world model and the complete navigation algorithm.

A.1 Relational Benchmark Construction

The original HM3D and MP3D ObjectNav benchmarks specify navigation goals using object categories. To evaluate compositional relational navigation, we construct relational episodes by augmenting the original scene annotations with geometrically verified relations between target and reference entities.

Target and Reference Entity Selection. For each scene, we first enumerate all annotated object instances and retain those belonging to the target categories considered by the corresponding ObjectNav benchmark. Let

$$\mathcal{O}_c = \{o_i \mid \mathrm{category}(o_i) = c\} \tag{9}$$

denote all instances of target category c. For each candidate target instance o_i, we search the annotated objects and regions in the same scene for valid reference entities. A target reference pair is retained only when their geometric configuration satisfies one of the predefined relational predicates.

For binary relations, a relational tuple is represented as

$$\mathcal{R} = (o_{\mathrm{tar}}, r, o_{\mathrm{ref}}), \tag{10}$$

where o_{tar} is the target instance, o_{ref} is the reference entity, and r denotes the corresponding spatial relation. Relations involving two reference entities, such as *between*, are represented as

$$\mathcal{R} = (o_{\mathrm{tar}}, r, o^1_{\mathrm{ref}}, o^2_{\mathrm{ref}}). \tag{11}$$

Geometric Relation Ground Truth. All relational labels are derived from the 3D semantic annotations rather than generated directly by a language model. Object positions and extents are obtained from their annotated 3D bounding boxes. For two objects o_i and o_j, we denote their horizontal centers by $\mathbf{p}_i$ and $\mathbf{p}_j$, respectively.

For proximity based relations such as *near* or *beside*, we compute the minimum horizontal distance between their bounding box footprints:

$$d_{\mathrm{box}}(o_i, o_j) = \min_{\mathbf{x}\in B_i, \mathbf{y}\in B_j} \|\mathbf{x} - \mathbf{y}\|_2. \tag{12}$$

A relation is considered valid when the two objects belong to a compatible spatial region and

$$d_{\mathrm{box}}(o_i, o_j) < \delta_{\mathrm{near}}, \tag{13}$$

where δ_{near} is a fixed geometric threshold used for benchmark generation.

For containment relations, the target is considered *inside* a reference region when its center lies inside the reference region and sufficient overlap exists between their spatial extents. For the *between* relation, let $\mathbf{p}_1$ and $\mathbf{p}_2$ denote the centers of two reference entities and $\mathbf{p}_t$ the target center. We compute the projection coefficient

$$\gamma = \frac{(\mathbf{p}_t - \mathbf{p}_1)^\top (\mathbf{p}_2 - \mathbf{p}_1)}{\|\mathbf{p}_2 - \mathbf{p}_1\|_2^2}. \tag{14}$$

The target is considered between the two reference entities when its projection lies on the interior of the connecting segment and its perpendicular distance to that segment is below a predefined threshold $\delta_{\mathrm{between}}$.

These predicates are evaluated directly from the annotated scene geometry, ensuring that the relational ground truth is independent of the navigation model.

Target Disambiguation. A generated relational description is retained only when the relation uniquely identifies one target instance of the specified category. Given a relational query q, we define the set of satisfying target instances as

$$\mathcal{O}(q) = \{o_i \in \mathcal{O}_c \mid o_i \text{ satisfies all relations in } q\}. \tag{15}$$

An episode is accepted only if

$$|\mathcal{O}(q)| = 1. \tag{16}$$

If several instances satisfy the same relation, an additional valid relational constraint is included when available. Otherwise, the candidate episode is discarded. Episodes without any valid relational target are also removed during benchmark construction. Therefore, every retained episode is associated with a unique ground truth target instance.

Instruction Generation. Natural language instructions are generated using predefined templates grounded in the verified relational tuples. For example, a tuple

$$(\text{sofa}, \textit{beside}, \text{chair})$$

may be expressed as "find the sofa beside the chair." Template based generation ensures that the language description remains consistent with the underlying geometric relation and avoids introducing unsupported relations through unconstrained language generation. All evaluated navigation methods receive the same relational instructions and episode splits.

A.2 Relational World Model Inference

The relational world model in RWMNav is training free and is instantiated using a frozen vision language model rather than a separately trained dynamics network. No task specific trajectory data, prediction loss, or parameter fine tuning is required. Instead, the model uses structured environmental context and the pretrained semantic knowledge of the VLM to infer plausible relational states behind unexplored frontiers.

Frontier Context Construction. For each candidate frontier f, RWMNav collects its spatial location, navigation distance, nearby observed entities, associated region information, and the relevant portion of the current environmental relation graph $\mathcal{G}_t$. These elements are organized into a structured frontier context:

$$\mathcal{C}_f = \text{Serialize}\left(\mathcal{G}_t, M_t, f\right), \tag{17}$$

where M_t denotes the current spatial semantic map. The serialization converts the current structured state into a compact representation that can be provided to the frozen VLM together with a fixed relational prediction prompt $\mathcal{P}$.

The resulting future relation prediction is written as

$$\hat{\mathcal{G}}^{f}_{t+\tau} = \mathcal{F}_{\text{RWM}}\left(\mathcal{C}_f; \mathcal{P}\right), \tag{18}$$

where $\mathcal{F}_{\text{RWM}}$ denotes the frozen VLM based relational inference procedure. Importantly, $\mathcal{F}_{\text{RWM}}$ contains no task specific trainable parameters.

Structured Relational Rollout. The relational prediction is decomposed into three reasoning stages. First, *Region Hypothesis* estimates the type of region that may become observable after crossing the candidate frontier. Second, *Entity Hypothesis* predicts entities that are plausible within the hypothesized region while considering entities already observed around the frontier. Finally, *Relation Completion* infers potential spatial and semantic relations among the predicted entities.

The resulting state is represented as

$$\hat{\mathcal{G}}^{f}_{t+\tau} = \left(\hat{V}^{\text{reg}}_f, \hat{V}^{\text{obj}}_f, \hat{E}^{\text{rel}}_f\right), \tag{19}$$

where $\hat{V}^{\text{reg}}_f$, $\hat{V}^{\text{obj}}_f$, and $\hat{E}^{\text{rel}}_f$ denote the predicted regions, entities, and relational edges, respectively. The VLM is constrained to return predictions in a predefined structured format, which is parsed into the corresponding relation graph before relational goal matching.

Observed and Predicted Relational States. RWMNav explicitly separates persistent relational memory from future relational hypotheses. The global relation graph $\mathcal{G}_t$ stores only objects, regions,

and relations supported by actual RGB-D observations. In contrast, $\hat{\mathcal{G}}_{t+\tau}^{f}$ is maintained only as a temporary prediction for evaluating frontier f. It is not directly written into the persistent relational memory.

After the agent explores the selected frontier and obtains new observations, predicted entities and relations are verified against the actual scene. Confirmed relations are incorporated into $\mathcal{G}_{t+1}$, whereas unsupported hypotheses are discarded. This mechanism prevents speculative VLM predictions from accumulating in the persistent environmental state.

A.3 Relational Goal Matching

Given the language derived goal graph $\mathcal{G}^q$ and a frontier conditioned future graph $\hat{\mathcal{G}}_{t+\tau}^{f}$, RWMNav evaluates both entity compatibility and relational consistency. Let

$$S_{\text{ent}}(f) \tag{20}$$

denote the semantic compatibility between target/reference entities in the goal graph and candidate entities in the predicted graph, and let

$$S_{\text{relation}}(f) \tag{21}$$

denote the consistency between the corresponding relational edges. The relational matching score is computed as

$$S_{\text{rel}}(f) = \alpha S_{\text{ent}}(f) + \beta S_{\text{relation}}(f). \tag{22}$$

The final frontier score further considers exploration utility and path cost:

$$S(f) = \lambda_r S_{\text{rel}}(f) + \lambda_e S_{\text{exp}}(f) - \lambda_c C(f), \tag{23}$$

where $S_{\text{exp}}(f)$ measures the expected exploration benefit and $C(f)$ denotes the navigation cost from the current agent position to frontier f. The frontier with the highest score is selected as the next high level navigation objective.

A.4 RWMNav Navigation Algorithm

Algorithm 1 summarizes the complete inference procedure of RWMNav. The entire framework operates in a zero shot manner without task specific model training.

B Additional Experimental Results and Visualizations

B.1 Additional Analysis of Relational Benchmark Construction

The relational benchmark is constructed from the original HM3D and MP3D ObjectNav scene splits while preserving the underlying environments and navigation settings. Unlike conventional ObjectNav episodes that specify only a target category, each relational episode additionally associates the target instance with one or more reference entities through geometrically verified relations. These relations are derived directly from the 3D semantic annotations rather than inferred from free form language generation.

A key requirement during episode construction is target uniqueness. For each candidate relational description, we enumerate all instances of the corresponding target category in the scene and retain the episode only when the specified relational constraints identify a unique target instance. If several instances satisfy the same relation, an additional valid relation is introduced when available. Otherwise, the candidate episode is discarded. Similarly, scenes in which no target instance satisfies the required relational configuration are excluded from the corresponding episode generation process. This procedure ensures that relational success can be evaluated against a well defined ground truth target rather than an ambiguous category level goal.

Natural language instructions are generated from predefined templates grounded in the verified target reference tuples. Therefore, the linguistic description and the underlying geometric relation remain consistent by construction. Importantly, all compared methods are evaluated using exactly the same relational episode files and target annotations, ensuring that performance differences arise from navigation and relational reasoning rather than from different goal generation procedures.

Algorithm 1 RWMNav Inference Procedure

Input: RGB-D observation o_t, pose p_t, relational instruction q, spatial semantic map M_t, relation graph $\mathcal{G}_t$
Output: Navigation action a_t, updated map M_{t+1}, updated relation graph $\mathcal{G}_{t+1}$
{— Relational Goal Parsing —}
Parse instruction q into relational goal graph $\mathcal{G}^q$
{— Perception and Environment Update —}
Detect open vocabulary entities from o_t
Project detected entities into M_t using depth and pose
Update observed relational state $\mathcal{G}_t$
Extract candidate frontier set $\mathcal{F}_t$
{— Frontier Conditioned Relational Prediction —}
for each $f \in \mathcal{F}_t$ **do**
    Construct structured context

$$\mathcal{C}_f \leftarrow \mathrm{Serialize}(\mathcal{G}_t, M_t, f)$$

    Predict region hypothesis using frozen VLM
    Predict candidate entities conditioned on the region hypothesis
    Complete potential relations among predicted entities
    Construct future relational graph

$$\hat{\mathcal{G}}^f_{t+\tau} \leftarrow (\hat{V}^{\mathrm{reg}}_f, \hat{V}^{\mathrm{obj}}_f, \hat{E}^{\mathrm{rel}}_f)$$

    Compute relational compatibility

$$S_{\mathrm{rel}}(f) \leftarrow \mathrm{Match}(\mathcal{G}^q, \hat{\mathcal{G}}^f_{t+\tau})$$

    Compute frontier utility

$$S(f) \leftarrow \lambda_r S_{\mathrm{rel}}(f) + \lambda_e S_{\mathrm{exp}}(f) - \lambda_c C(f)$$

end for
{— Frontier Selection and Execution —}
$f^*_t \leftarrow \arg\max_{f \in \mathcal{F}_t} S(f)$
Plan a collision free path toward f^*_t
Execute the next low level navigation action a_t
{— Online Relational Verification —}
Acquire new RGB-D observation o_{t+1}
Update spatial semantic map M_{t+1}
Verify predicted entities and relations using actual observations
Add verified relations to $\mathcal{G}_{t+1}$ and discard unsupported hypotheses
return $a_t, M_{t+1}, \mathcal{G}_{t+1}$

B.2 Additional Analysis of Relational World Model

RWMNav employs a training free relational world model instantiated with a frozen vision language model. The model does not learn a separate transition network from navigation trajectories. Instead, it uses the current environmental relation graph, spatial semantic map, and frontier specific context to infer plausible relational structures that may become observable after exploring a candidate frontier. This design preserves the zero shot setting while allowing the agent to perform prospective reasoning over unexplored regions.

For each candidate frontier, the current local context is organized into a structured representation containing nearby observed entities, regional information, spatial connectivity, and relevant relations already verified in the environment. Based on this information, the relational prediction process proceeds through three conceptual stages. First, the model predicts the potential region that may lie beyond the frontier. Second, it estimates the entities that are likely to occur in that region. Finally, it completes the possible relations among the predicted entities. The resulting prediction forms a

frontier conditioned future relation graph that can be directly compared with the language derived relational goal graph.

This prediction is fundamentally different from the persistent environmental relation graph. The latter contains only entities and relations supported by actual RGB-D observations, whereas frontier conditioned predictions are maintained as temporary hypotheses for decision making. They are never directly inserted into the persistent relational memory. Once the agent explores the selected region, new observations are used to verify the predicted entities and relations. Consistent relations are incorporated into the environmental graph, while unsupported hypotheses are discarded. This separation limits the accumulation of speculative VLM predictions during long horizon navigation.

The benefit of such prediction is particularly important under partial observability. A target instance and its reference entities may not be simultaneously visible from the current viewpoint, making purely reactive semantic matching insufficient. By estimating which relational structures may emerge along different exploration directions, RWMNav can prefer frontiers that are not only semantically relevant to the target category but are also more likely to satisfy the complete relational goal.

B.3 Visualization of Navigation Results

We further visualize a representative relational navigation episode to illustrate the complete decision process of RWMNav. Figure 4 presents nine intermediate navigation states during exploration. At each stage, the agent incorporates newly observed semantic and spatial evidence into its environmental representation, updates the verified relational state, and evaluates unexplored frontiers using their predicted future relational structures.

At the beginning of navigation, only limited relational evidence is available and frontier selection therefore relies more strongly on predicted future states. As exploration proceeds, additional objects and regions become observable, allowing the agent to progressively verify or reject previous hypotheses. Consequently, navigation decisions become increasingly constrained by the observed relational structure rather than by semantic category relevance alone. This iterative interaction between future state prediction and environmental feedback enables the agent to gradually reduce uncertainty and approach the target instance specified by the relational instruction.

Figure 5 shows the final state of the same navigation episode. After progressively exploring candidate regions and verifying the relational evidence through actual observations, the agent reaches the designated target instance satisfying the relational constraints in the instruction. The complete trajectory demonstrates how RWMNav combines prospective relational reasoning with online environmental feedback to guide navigation from an initially partial observation toward the correct compositional goal.

C Discussion on Limitations and Future Directions

Although RWMNav demonstrates the effectiveness of relational world modeling for compositional zero shot object navigation, several limitations remain and suggest directions for future research.

Scope of Relational Goal Definitions. The current benchmark focuses on relational goals that can be reliably derived from the available 3D semantic annotations. These relations are defined using explicit geometric predicates to ensure reproducible ground truth and unambiguous target identification. However, real world instructions may involve more diverse relations, including functional, contextual, and viewpoint dependent descriptions. Extending the benchmark toward richer relation types will require more expressive annotations and carefully defined reference frames. Future work could further incorporate learned relational grounding while preserving the consistency between language descriptions and physical scene geometry.

Reliability of Future Relational Prediction. RWMNav uses a frozen vision language model to predict possible regions, entities, and relations behind unexplored frontiers. Such predictions provide useful prospective guidance, but they may also contain incorrect or overconfident hypotheses due to the imperfect commonsense knowledge of the underlying model. To limit this effect, predicted relations are used only for frontier evaluation and are not directly inserted into the persistent environmental relation graph. Nevertheless, inaccurate predictions may still influence short term



Figure 4: Visualization of the navigation process of RWMNav. The sequence shows representative intermediate states during relational goal navigation, where the agent progressively updates its environmental representation and refines frontier selection using future relational predictions and newly acquired observations.

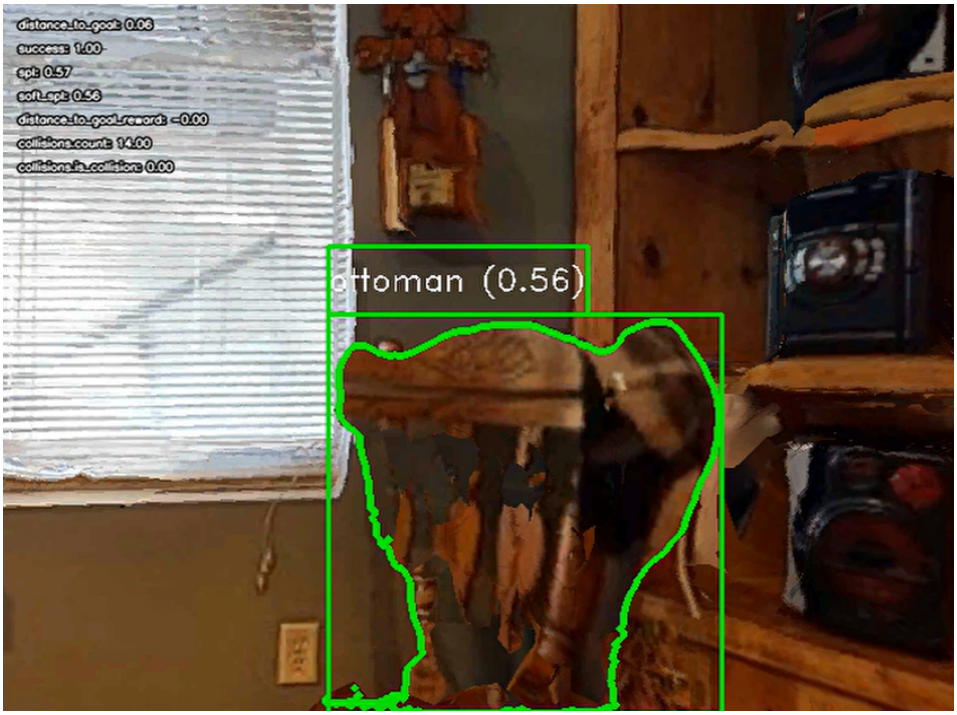


Figure 5: Successful navigation result of RWMNav. The agent reaches the designated target instance after progressively verifying the target entity, reference entities, and their relational configuration during exploration.

exploration decisions. Future work could introduce uncertainty estimation or multi hypothesis relational prediction to better calibrate the confidence of predicted future states.

Dependence on Open Vocabulary Perception. The quality of relational reasoning depends on reliable detection and grounding of both target and reference entities. Missed detections, segmentation errors, or incorrect semantic associations may therefore propagate into the environmental relation graph and affect subsequent frontier evaluation. This issue becomes more significant under heavy occlusion or when several visually similar instances are present. Improving open vocabulary perception and incorporating temporal observation consistency could further enhance the robustness of relational state construction.

Inference Efficiency and Generalization. RWMNav evaluates multiple candidate frontiers through frontier conditioned relational prediction, which introduces additional inference cost compared with purely reactive navigation methods. The current implementation therefore balances the number of candidate frontiers with prediction quality during each decision cycle. More efficient candidate filtering or lightweight relational predictors could reduce this overhead. In addition, although the current framework is evaluated on indoor navigation benchmarks, the underlying formulation is not restricted to specific object categories or environments. Extending relational world modeling to more diverse scenes and embodied tasks remains an important direction for future study.