Simen Owen Academic Proceedings Series

3rd International Conference on Electronics, Engineering, Computer Science and Applied Development (EESD 2026)

Article

Hierarchical World Model-Driven Visual Navigation Method for Embodied Robots in Dynamic Environments

Yujian Fu[1,*]

[1] Qixin Honors School, Zhejiang Sci-Tech University, Hangzhou, China
* Correspondence: Yujian Fu, Qixin Honors School, Zhejiang Sci-Tech University, Hangzhou, China

Abstract: Visual navigation is a crucial capability for embodied robots to achieve autonomous movement in indoor environments. Due to factors such as moving obstacles, temporary obstructions, and changes in object positions, robots are prone to being disturbed during navigation, which affects driving safety and path efficiency. Traditional visual navigation methods are usually based on end-to-end strategies, semantic maps, or single-level world models. These approaches often face difficulties in balancing long-term navigation goals and short-term risk responses in dynamic environments. To address this issue, this paper proposes a visual navigation method based on a hierarchical world model, which models global semantic planning and local dynamic prediction separately. The global semantic world model is responsible for maintaining the semantic topological memory of the environment and providing support for sub-objective selection. The local dynamic world model predicts short-term reachability and potential contact risks between the robot and surrounding obstacles. Subsequently, the risk-aware hierarchical strategy integrates the value of sub-objectives, predicted reachability, and dynamic risks to generate navigation actions. Experiments were conducted on two simulation platforms, iGibson and Habitat-Matterport 3D (HM3D), covering static scenes and simulated dynamic scenes. The experimental results show that the proposed method outperforms existing methods in both static and dynamic scenarios, demonstrating better comprehensive performance in multiple aspects and verifying the effectiveness of the hierarchical world model in dynamic visual navigation tasks. This indicates that modeling semantic memory and dynamic risk prediction separately can help improve the navigation performance of robots in dynamic indoor environments.

Keywords: embodied robot; visual navigation; world model; dynamic environment; risk prediction

Received: 03 June 2026
Revised: 13 July 2026
Accepted: 28 July 2026
Published: 01 August 2026

1. Introduction

Robots rely on first-person visual perception to sense the surrounding environment, determine their relative position with respect to a target, and plan safe and efficient movement paths. This capability, known as visual navigation, is widely deployed in service robots, warehouse automation systems, assistive devices, and mobile operation platforms. Real indoor environments are inherently dynamic [1]. Factors such as human movement, temporary object placement, furniture repositioning, and localized occlusions can alter spatial configurations, rendering previously acquired environmental representations outdated. Consequently, navigation strategies designed for static settings often exhibit reduced reliability in dynamic conditions, potentially leading to operational interruptions or safety incidents.

Recent advances in visual navigation have yielded notable improvements. End-to-end approaches directly map visual inputs to navigation actions, yet demonstrate limited robustness under environmental variation. Semantic mapping combined with topological

planning supports persistent environmental representation, but typically lacks explicit mechanisms for modeling temporal changes [2]. Single-level world models attempt to jointly encode long-term structural knowledge and short-term dynamics; however, persistent environmental evolution may cause interference between these two information streams, complicating the simultaneous optimization of navigation objectives and real-time safety assurance.

To address these challenges, this paper introduces a visual navigation method grounded in a hierarchical world model, aimed at strengthening the adaptability of embodied robots in dynamic indoor settings [3]. The approach decouples global semantic planning from local dynamic prediction during modeling and enables coordinated decision-making through layered control. Specifically, a global-local hierarchical world model is constructed to separately represent long-term semantic navigation structure and short-term spatial risk patterns; its efficacy is validated via comparative performance analysis and ablation studies. A dynamic risk-aware navigation strategy is formulated, which modulates low-level motion execution based on predicted local risk assessments while preserving high-level sub-goal integrity. Finally, the method is evaluated in an open test environment encompassing both static and dynamic scenarios, and benchmarked against representative baseline methods to demonstrate advantages in safety assurance, path optimality, and environmental adaptability.

The technical workflow proceeds as follows: At each time step, the system receives RGB-D observations, robot pose, historical action sequence, and target specification as joint inputs. A visual encoder extracts spatial features from the sensory data; the global semantic world model updates topological memory associated with the navigation target using these features. Concurrently, the local dynamic world model forecasts accessibility and potential safety constraints within the robot's forward region [3]. Based on this dual representation, the high-level strategy selects navigation sub-targets leveraging global memory, while the low-level strategy synthesizes current observations and risk predictions to generate precise motion commands. Training employs a composite objective comprising navigation loss, prediction loss, risk classification loss, and hierarchical consistency loss.

This study investigates whether explicit separation of semantic memory construction and dynamic prediction enhances visual navigation performance in evolving indoor environments. Such architectural decoupling contributes to improved operational safety, greater resilience against environmental variability, and generation of interpretable intermediate outputs—including sub-goals and quantitative risk indicators—that support transparent analysis of navigation reasoning [4]. For embodied robots operating in shared human-robot spaces—such as service robots—the resulting improvements in safety assurance and decision transparency are particularly valuable.

2. Related Works

2.1. Embodied Visual Navigation

Embodied visual navigation integrates visual perception, environmental understanding, and action decision-making, enabling robots to perform autonomous navigation based on real-time observations [5]. Prior research has demonstrated that robots can acquire navigation strategies using first-person RGB or RGB-D images and accomplish goal-directed movement in indoor environments. Compared with traditional map-dependent navigation methods, learning-based approaches reduce the need for manual mapping and maintain robust navigation performance under partial observability.

However, most existing methods are primarily validated in static environments. Their navigation performance often degrades significantly in the presence of dynamic elements such as pedestrians, moving furniture, or temporary obstacles [6]. Additionally, many approaches exhibit limited generalization across unseen environments due to strong scenario-specific training dependencies. To enhance robustness, this paper explicitly models dynamic risks as a distinct prediction target.

2.2. End-to-End Learning-Based Navigation

End-to-end reinforcement learning and imitation learning are widely adopted in visual navigation [7]. These approaches directly map raw visual input to navigation actions, enabling joint optimization of perception and control. By bypassing modular pipelines—such as separate perception, mapping, and planning stages—they mitigate error propagation across components and yield a more integrated and efficient navigation framework.

Nevertheless, such methods often suffer from limited interpretability and constrained generalization capability. The rationale behind a specific action—whether driven by target approach, obstacle avoidance, or spurious visual cues—remains difficult to discern [8]. In dynamic environments, robots must concurrently process long-term stable spatial information and short-term transient hazards; yet monolithic models typically struggle to reconcile these dual temporal scales. To enhance transparency and robustness, this work incorporates intermediate representations—including global sub-goals and local dynamic risk scores—to structure and clarify the decision-making process.

2.3. Semantic Mapping and Topological Planning

Semantic mapping and topological planning provide continuous navigation references for robots by preserving environmental information [5]. These methods record explored areas, object categories, room layouts, and accessible regions, enabling robots to reduce redundant exploration and optimize navigation objectives. Compared with traditional approaches, topological representation offers greater compactness and is better suited for long-term indoor navigation tasks.

Most semantic maps primarily capture the static structure of the environment and insufficiently account for local dynamic changes. For example, a corridor that was initially passable may become temporarily inaccessible due to pedestrian movement or shifting obstacles. To address this limitation, the proposed approach maintains global semantic memory for path planning while decoupling local dynamic prediction, thereby supporting independent long-term planning and short-term risk assessment.

2.4. World Models for Navigation

The world model captures regularities in environmental state evolution and forecasts future states, rewards, or potential risks to inform navigation decisions [9]. Even under partial observability, the robot can infer likely outcomes of candidate actions via the world model, thereby enhancing navigation efficiency.

Most existing approaches employ a single-level world model to jointly process visual inputs, spatial memory, action history, and task-related information [10]. While computationally straightforward, this design may lead to interference between long-term semantic representations and short-term dynamic variations. Consequently, local risk updates may lag behind environmental changes. To address this, the proposed method introduces a hierarchical world model that separately encodes global semantic topological structure and local dynamic predictions.

2.5. Dynamic Environment Navigation

Research on dynamic navigation primarily addresses obstacle avoidance, trajectory prediction, and online path adjustment [6]. These approaches enable robots to rapidly adapt their behavior in response to environmental changes and demonstrate robust responsiveness to moving pedestrians or transient obstacles. However, they predominantly prioritize immediate safety concerns and struggle to independently sustain long-term navigation objectives throughout the entire task.

Even with effective dynamic obstacle handling, robots may gradually drift away from the overarching navigation objective. In the absence of stable high-level planning, frequent directional adjustments, suboptimal path selection, and localized behavioral oscillation can occur. Conversely, high-level plans that inadequately incorporate real-time environmental dynamics—though globally sound—may lead to execution-phase safety

challenges. Thus, dynamic visual navigation must jointly optimize long-term goal consistency and short-term environmental responsiveness [11].

Collectively, prior work has advanced visual navigation through improvements in visual strategy learning, semantic memory, world modeling, and dynamic obstacle management. These components are typically developed in isolation, and a unified framework for integrating long-term semantic planning with short-term risk mitigation remains underdeveloped. Particularly in dynamic settings, ensuring alignment between high-level navigation intent and low-level action execution continues to pose a significant challenge [11]. To bridge this gap, this paper introduces a hierarchical world model: global semantic memory supports sub-goal selection; local dynamic models forecast reachability and proximity risks; and hierarchical control strategies tightly couple planning with execution. Validation includes comparative benchmarking, ablation studies, and targeted dynamic environment evaluations.

3. Methodology

3.1. Overall Architecture

Figure 1 presents the overall architecture of the hierarchical world model-driven visual navigation method. This framework consists of RGB-D input, visual-spatial feature extraction, global semantic world model, local dynamic world model, and hierarchical navigation control module [12]. At each time step, the robot receives inputs such as RGB images, depth information, pose, the action taken in the previous time step, target information, proximity alerts, and the state of dynamic obstacles, and processes them uniformly during the model update and policy training processes.

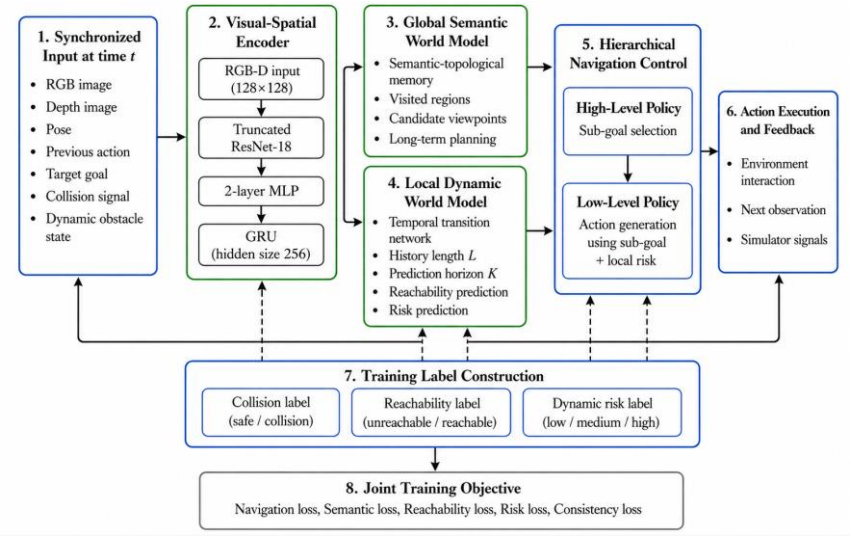

Figure 1. Overall Architecture of the Hierarchical World Model-Driven Visual Navigation Method.

As shown in Figure 1, RGB-D observations and the robot's state first pass through the visual-spatial encoder to extract environmental features. Subsequently, the encoded features are input into two parallel world model branches. The global branch maintains semantic topological memory and supports long-term navigation planning; the local branch predicts short-term environmental changes, including navigability and proximity risks. On this basis, the high-level strategy selects navigation sub-goals based on the global memory, and the low-level strategy integrates current observations, sub-goal information, and risk prediction results to generate specific actions [13].

This method employs hierarchical modeling to separately process long-term semantic information and short-term dynamic changes. The global model focuses on the environmental structure associated with the target, providing a stable navigation direction; the local model evaluates action safety based on real-time environmental variations [11]. Compared to approaches that encode all navigation information into a

single representation, this design reduces interference between different levels of information, enabling proximity-related assessments to more effectively inform action selection.

3.2. Data Synchronization and Problem Formulation

To ensure consistency across heterogeneous data sources, sensor measurements and simulator states are aligned to a common temporal discretization. At time step t, the model input comprises:

$$\mathcal{X}_t = \{I_t, D_t, p_t, a_{t-1}, g, c_t, e_t\} \quad (1)$$

Here, I_t and D_t denote the RGB image and depth image captured at the current time step, respectively; p_t denotes the robot's current pose; a_{t-1} denotes the action executed at the preceding time step; g denotes the target location; c_t denotes the proximity alert signal; and e_t denotes the state of dynamic obstacles. RGB-D observations, robot state, action history, and proximity alerts are all sampled synchronously within the same simulation time step. Given the relatively low update rate of dynamic obstacle states, nearest-neighbor temporal matching is applied to align them with the current time step, thereby preserving temporal coherence of the input representation.

Based on this input representation, the navigation policy is formulated as:

$$a_t \sim \pi_\theta(a_t \mid \mathcal{X}_{1:t}, M_t^g, M_t^l) \quad (2)$$

In this formulation, a_t denotes the action selected at the current time step; π_θ denotes the policy mapping; $\mathcal{X}_{1:t}$ denotes the observation history from initialization to the present; and M_t^g and M_t^l denote the global semantic world model and the local dynamic world model, respectively [14]. The robot's action space consists of four discrete primitives: forward motion, counterclockwise rotation, clockwise rotation, and halt.

3.3. Visual-Spatial Encoder

RGB and depth images are concatenated into a four-channel RGB-D input, which is uniformly resized to 128×128. A truncated ResNet-18 backbone extracts visual features, followed by encoding through two multilayer perceptron (MLP) layers and gated recurrent units (GRUs) with a hidden dimension of 256. This architecture jointly preserves spatial structure and captures short-term temporal dynamics [15].

$$z_t = \mathrm{GRU}_{\theta_r}\left(\phi_{\theta_v}([I_t, D_t]), z_{t-1}\right) \quad (3)$$

Here, z_t denotes the encoded latent features, z_t represents the concatenated RGB-D input, $\varphi\phi_{\theta_v}$ denotes the convolutional visual encoder, and ϕ_{θ_v} denotes the recurrent encoder [11]. The resulting features z_t serve as inputs for subsequent semantic prediction, dynamic prediction, and policy decision-making modules.

3.4. Training Label Construction

The supervision signals during the training process are all automatically generated by the simulator. Among them, the interaction label is obtained based on the contact state between the robot and the environment. The accessibility label is determined using the shortest path planner in the simulator, and is assigned by verifying whether candidate sub-targets can be reached within the specified time limit. The dynamic risk label is generated by analyzing potential proximity events or near-contact situations with moving obstacles over a future time horizon:

$$y_t^r = \mathbb{I}\left(\max_{1\le k\le K} c_{t+k} = 1 \text{ or } \min_{1\le k\le K} d_{t+k}^{obs} < \tau_d\right) \quad (4)$$

Among them, y_t^r r represents the binary classification risk label, $\mathbb{I}(\cdot)$ is the indicator function, K represents the prediction time range, c_{t+k} represents the proximity event signal at the future time point, d_{t+k}^{obs} represents the distance between the robot and the nearest dynamic obstacle, and τ_d represents the safety distance threshold [10]. In this way, the local model can identify potential risks in advance and adapt the navigation strategy accordingly.

3.5. Global Semantic World Model

The global model is represented as $M_t^g = (V_t, E_t, H_t)$, where V_t denotes the set of nodes, E_t encodes the connection relationships among nodes, and H_t constitutes the

node feature matrix [4, 6]. Each node corresponds to an explored region, a key observation location, or a candidate area with semantic relevance. Node features encompass visual embeddings, semantic confidence scores, visit frequency, estimated distance to target locations, and historical risk statistics. Edges encode passability relationships between spatial regions.

$$H_t = GAT_{\theta_g}(H_{t-1}, E_{t-1}, [z_t, p_t, s_t]) \quad (5)$$

Here, H_t denotes the updated node features, computed by GAT_{θ_g}, a two-layer graph attention network module. H_{t-1} and E_{t-1} represent the graph state at the preceding time step, z_t denotes the current visual feature input, p_t encodes the robot's pose, and s_t reflects the semantic prediction output. As exploration progresses, new nodes are dynamically incorporated into the graph to represent unvisited areas or high-confidence semantic targets; edge weights are adaptively adjusted based on inter-region distances and traversability assessments.

3.6. Local Dynamic World Model

The local model leverages recent visual inputs and action history to forecast subsequent environmental evolution and the likely consequences of the robot's actions. It implements a time series prediction network based on gated recurrent units (GRUs), with two parallel output branches: one estimating future navigational feasibility and the other assessing potential interaction hazards [11].

$$\hat{h}_{t+1:t+K}, \hat{q}_t, \hat{r}_t = F_{\theta_d}(z_{t-L:t}, a_{t-L:t-1}) \quad (6)$$

Here, F_{θ_d} denotes the local dynamic model, $\hat{h}_{t+1:t+K}$ denotes the predicted potential state sequence, $\hat{q}_t$ denotes the navigational feasibility prediction, $\hat{r}_t$ denotes the interaction hazard prediction, L denotes the length of historical observations, and K denotes the prediction horizon. The two branches are supervised independently using feasibility and hazard labels generated by the planner (see Eq. 4).

3.7. Hierarchical Policy and Safety Aggregation

The high-level strategy selects navigation sub-goals using global graph information, while the low-level strategy generates precise actions from the current state. To prevent selection of candidate nodes with high semantic value but elevated safety risks, this work computes a composite score for each node by jointly evaluating semantic value, accessibility, and dynamic risk [1, 8].

$$S(v_i) = Q_h(v_i \mid H_t, g) - \alpha_t \hat{r}_t(v_i) + \beta_t \hat{q}_t(v_i) \quad (7)$$

$S(v_i)$ denotes the security perception score of candidate node v_i, $Q_h(v_i \mid H_t, g)$ denotes the high-level strategy's estimated node value, $\hat{r}_t(v_i)$ denotes the predicted risk, $\hat{q}_t(v_i)$ denotes the predicted accessibility, and α_t and β_t denote dynamically adjusted weight parameters. The candidate node achieving the highest composite score is selected as the current sub-goal. Incorporating the risk term ensures alignment with global navigation objectives while enabling adaptive avoidance of hazardous regions [10].

Control weights are updated dynamically in response to recent prediction errors:

$$\alpha_{t+1} = \alpha_t + \eta_\alpha \mid \hat{r}_t - y_t^r \mid, \beta_{t+1} = \beta_t - \eta_\beta \mid \hat{r}_t - y_t^r \mid \quad (8)$$

where η_α and η_β represent the weight adaptation rates. An increase in risk prediction error triggers greater weighting of safety factors and reduced dependence on accessibility-driven behavior [10].

3.8. Training Objective and Algorithm

The model is trained end-to-end using a composite objective comprising navigation loss, semantic prediction loss, accessibility prediction loss, risk classification loss, and hierarchical consistency loss.

$$\mathcal{L} = \mathcal{L}_{nav} + \lambda_1 \mathcal{L}_{sem} + \lambda_2 \mathcal{L}_{reach} + \lambda_3 \mathcal{L}_{risk} + \lambda_4 \mathcal{L}_{cons} \quad (9)$$

$\mathcal{L}_{nav}$ denotes the navigation strategy loss, $\mathcal{L}_{sem}$ denotes the semantic prediction loss, $\mathcal{L}_{reach}$ denotes the reachability prediction loss, $\mathcal{L}_{risk}$ denotes the risk classification loss, $\mathcal{L}_{cons}$ denotes the high- and low-level strategy consistency loss, and λ_1 to λ_4 denote the respective weighting coefficients [8].

[[TABLE_001]]

The proposed method integrates three key components: synchronous multimodal representation learning, a world model with decoupled global and local reasoning pathways, and an adaptive safety-aware fusion mechanism for hierarchical control [4].

4. Results and Analysis

4.1. Experimental Setup

This paper conducted experiments in two publicly available embodied navigation environments. The main experiments were carried out based on the iGibson platform, which supports visual rendering, physical simulation, indoor scene interaction, and motion planning. To evaluate the model's cross-scenario generalization ability, further tests were conducted in the Habitat-Matterport 3D (HM3D) environment. Since the original HM3D scenes consist primarily of static environments, dynamic obstacles were introduced into the Habitat simulator to assess model robustness under time-varying conditions.

In the experiment, both RGB images and depth maps were resized to a resolution of 128×128. Depth values were clipped to the range of 0.1–5.0 meters and normalized. The robot's pose was represented in relative coordinates, and tasks exceeding 500 steps were terminated. Dynamic obstacles were simulated by sampling random initial positions, velocities, and motion trajectories. Three scenario configurations were employed: static, mild dynamic, and high-density dynamic. Data were partitioned by scenario into training, validation, and test sets at ratios of 70%, 15%, and 15%, respectively.

The model was implemented using PyTorch and trained with the Adam optimizer. The batch size was set to 64, the learning rate to 3×10^{-4}, the history length L to 4, and the prediction horizon K to 5. Under identical experimental conditions, performance was compared against four baseline approaches: an end-to-end visual policy trained via reinforcement learning, a semantic mapping method integrated with classical path planning, a neural SLAM–based navigation framework, and a single-level world model. Evaluation metrics included success rate (SR), weighted path length success rate (SPL), and contact frequency (CF). Reported results reflect the mean and standard deviation across five independent runs.

4.2. Performance Comparison

Table 1 summarizes the navigation results of the proposed method in the dynamic test scenarios of iGibson and HM3D. All indicators report the average values and standard deviations of five independent runs. In the iGibson environment, the method achieved the highest success rate and the lowest collision rate, while the success-weighted path length was slightly higher than that of the single-level world model. In the HM3D environment, the method also achieved the best performance in success rate and collision rate, and its success-weighted path length was close to that of the neural SLAM navigation approach. Overall, the hierarchical world model prioritizes navigation safety and task completion reliability in dynamic environments over strict path efficiency.

Table 1. Performance Comparison on Dynamic Navigation Scenes, Mean ± SD, N = 5

Method	iGibson SR ↑	iGibson SPL ↑	iGibson CR ↓	HM3D SR ↑	HM3D SPL ↑	HM3D CR ↓
PPO-based End-to-End Policy	0.631 ± 0.031	0.451 ± 0.026	0.309 ± 0.036	0.589 ± 0.038	0.412 ± 0.031	0.334 ± 0.041
Semantic Map + A* Planner	0.674 ± 0.027	0.502 ± 0.029	0.274 ± 0.032	0.636 ± 0.035	0.468 ± 0.030	0.298 ± 0.036
Neural SLAM-based Navigation	0.701 ± 0.024	0.536 ± 0.021	0.247 ± 0.028	0.663 ± 0.032	0.502 ± 0.029	0.279 ± 0.033
Single-Level World Model	0.724 ± 0.026	0.552 ± 0.024	0.226 ± 0.027	0.681 ± 0.030	0.511 ± 0.027	0.263 ± 0.031

Proposed	0.752 ± 0.023	0.558 ± 0.025	0.197 ± 0.026	0.699 ± 0.029	0.507 ± 0.028	0.244 ± 0.030

Compared with the optimal baseline method, the proposed method increased success rate by 2.8 percentage points and decreased collision rate by 12.8% in iGibson. In HM3D, success rate increased by 1.8 percentage points and collision rate decreased by 7.2%. Meanwhile, its success-weighted path length was slightly lower than that of the single-layer world model, indicating a deliberate trade-off between path efficiency and robustness when navigating under dynamic risk conditions [13].

4.3. Ablation Study and Mechanism Verification

Table 2 presents the ablation experiment results under the iGibson dynamic scenario. After removing the local dynamic model, the incidence of physical contact with dynamic obstacles increased from 0.197 ± 0.026 to 0.249 ± 0.033. The most pronounced degradation occurred in safety performance. This indicates that short-term risk prediction enables robots to proactively adjust navigation behavior, thereby mitigating interactions with moving entities. In contrast, removing the global semantic model primarily reduced navigation efficiency, with both success rate (SR) and success-weighted path length (SPL) declining. The change in physical contact incidence was relatively modest, suggesting that global semantic memory mainly supports environmental comprehension and goal-oriented planning. Further removal of the safety aggregation module led to a measurable decline in overall performance, indicating that this component contributes to more consistent and reliable action selection by integrating navigational objectives with real-time risk assessment.

Table 2. Ablation Results on iGibson Dynamic Scenes, Mean ± SD, N = 5

Variant	SR ↑	SPL ↑	CR ↓
Full Model	0.752 ± 0.023	0.558 ± 0.025	0.197 ± 0.026
w/o Global Semantic Model	0.716 ± 0.028	0.521 ± 0.027	0.214 ± 0.030
w/o Local Dynamic Model	0.709 ± 0.030	0.529 ± 0.031	0.249 ± 0.033
w/o Safety Aggregation	0.731 ± 0.025	0.540 ± 0.026	0.223 ± 0.029
w/o Hierarchical Consistency Loss	0.738 ± 0.024	0.545 ± 0.025	0.211 ± 0.027

From the overall fusion results, different modules fulfill distinct functional roles during navigation. The variant without the local dynamic model maintains relatively high path efficiency, yet exhibits a marked rise in physical contact incidence—demonstrating that optimizing for shorter paths alone may overlook emergent hazards in dynamic settings. The complete model achieves a balanced trade-off among task completion, path efficiency, and operational safety through the coordinated integration of global semantic planning, local risk prediction, and safety-aware decision refinement.

4.4. Convergence Analysis

Figure 2 shows the success rate (SR) change curves of the proposed method and four baseline methods during training. Each curve represents the average across five independent random initialization trials, with the shaded region indicating the standard deviation. In the early training phase, the proposed method exhibits a slightly slower SR improvement compared to the single-layer world model, likely attributable to the joint optimization of the global semantic graph and the local dynamic model, which introduces additional complexity and delays initial convergence [3]. As training progresses, the proposed method consistently outperforms all baselines and demonstrates superior stability in later stages, exhibiting reduced performance fluctuation relative to the single-layer world model. At final evaluation, the proposed method achieves an SR of 0.758 ± 0.020, exceeding the best baseline result of 0.711 ± 0.026.

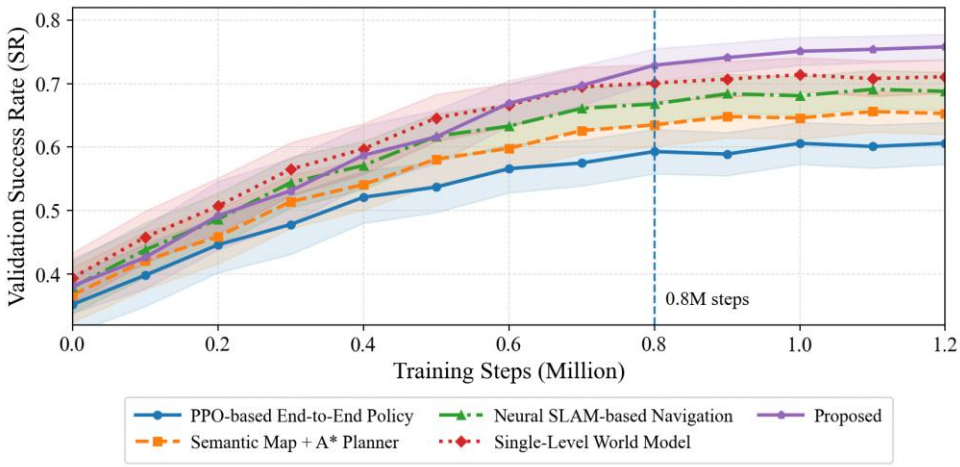

Figure 2. Training Convergence Curves of Validation SR, Mean ± SD, N = 5.

The evolution of the training curve reflects the hierarchical optimization dynamics. The concurrent learning of global semantic memory and local dynamic prediction necessitates an initial adaptation period, leading to comparatively modest early-stage performance gains. With progressive stabilization of the risk prediction module, lower-level navigation strategies gain access to increasingly reliable risk assessments, thereby enabling more consistent and robust decision-making throughout the navigation task.

4.5. Stability, Significance and Robustness Analysis

Figure 3 further presents the multiple run results of navigation safety metrics in the iGibson and HM3D environments. The error bars represent the 95% confidence interval. Compared with the single-layer world model, the method proposed in this paper significantly improves navigation safety in iGibson ($p = 0.018$); a similar improvement is also observed in HM3D, reaching statistical significance ($p = 0.041$).

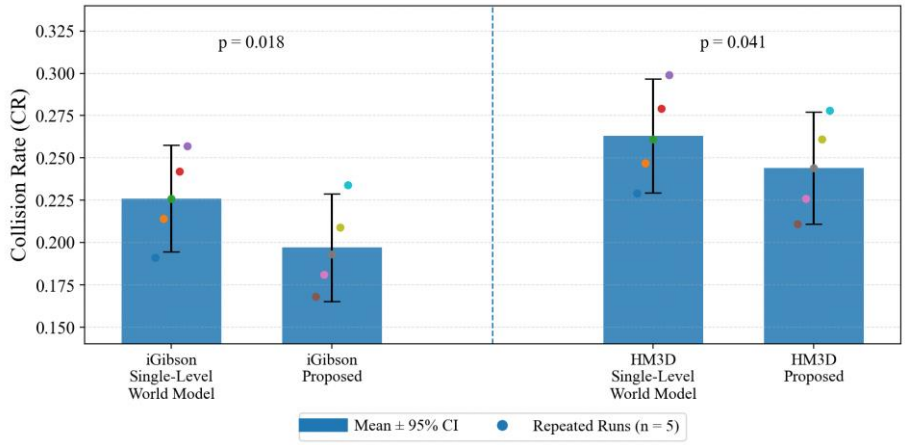

Figure 3. Collision Rate Comparison with 95% Confidence Intervals, N = 5.

In the iGibson environment, smaller result fluctuations indicate that this method exhibits more stable navigation performance under varying operational conditions. In contrast, the larger fluctuation range in HM3D suggests that the model still faces certain challenges when transferring across diverse scene configurations [3, 9]. This may be related to the test setup where dynamic elements are introduced within the static reconstruction environment. Besides performance metrics, the proposed method also provides intermediate information such as sub-object selection and dynamic risk scoring, enabling researchers to further analyze the decision-making basis behind the robot's actions—for instance, whether the action is primarily driven by goal-oriented planning or

responsive local adaptation. From a practical deployment perspective, further verification is still needed by considering factors such as sensor noise, system delay, and hardware security mechanisms.

5. Conclusion

The main challenge posed by dynamic environments for robot navigation lies in the fact that environmental information previously acquired by the robot may quickly become outdated. A navigation strategy effective in static environments often requires frequent path adjustments when encountering moving obstacles or spatial reconfigurations, potentially compromising safety. To address this, this paper proposes a visual navigation method based on a hierarchical world model, which explicitly separates stable environmental features from rapidly changing ones. Evaluation results in the iGibson and HM3D environments demonstrate that the method maintains robust navigation performance under dynamic conditions.

Further analysis reveals that distinct characteristics of environmental information exert different influences on navigation behavior. The local dynamic model primarily handles nearby changes—such as those introduced by moving obstacles—and contributes significantly to short-term risk anticipation. When this module is disabled, collision frequency increases, confirming that forecasting imminent environmental changes supports proactive behavioral adaptation. In contrast, the global semantic model plays a more decisive role in target localization and high-level route planning. However, enhanced risk mitigation entails trade-offs: in certain test scenarios, the robot selects longer, safer paths, resulting in a modest reduction in Success-weighted by Path Length (SPL) compared to a single-layer world model.

Although the method achieves consistent performance across multiple simulated environments, its generalization to novel scenarios remains limited. Performance variations emerge when dynamic elements are introduced into large-scale reconstructed environments such as HM3D, indicating room for improvement in cross-environment adaptability. Spatial representations learned during training do not always transfer effectively to new dynamic settings—particularly when layout configurations or obstacle distributions differ substantially—potentially affecting risk assessment accuracy. Importantly, the model's outputs—including sub-goals and risk scores—provide interpretable signals about navigational intent, enabling researchers to distinguish between goal-directed motion and reactive avoidance behavior.

All experiments reported in this work were conducted in simulation. Dynamic obstacles in these simulations follow predefined motion patterns, whereas human movement in real-world settings exhibits greater complexity and unpredictability. Additionally, while the hierarchical architecture enables specialized processing of diverse environmental cues, it incurs higher computational overhead. This poses challenges for deployment on resource-constrained platforms. Future work will include physical robot validation, integration of realistic human motion datasets, and exploration of computationally efficient model architectures to strengthen practical applicability.

References

1. A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," arXiv preprint arXiv:2412.03572, 2024.
2. D. Zheng, S. Huang, L. Zhao, Y. Zhong, and L. Wang, "Towards learning a generalist model for embodied navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 13624–13634.
3. T. Wang, X. Zhao, W. Cai, and C. Sun, "ImagineNav++: Prompting Vision-Language Models as Embodied Navigator through Scene Imagination," arXiv preprint arXiv:2512.17435, 2025.
4. Y. Zhuang, J. Ren, X. Ye, J. Shen, T. Yue, M. Faayez, ..., and T. Shu, "Simworld-robotics: Synthesizing photorealistic and dynamic urban environments for multimodal robot navigation and collaboration," in *Adv. Neural Inf. Process. Syst.*, vol. 38, 2026, pp. 51854–51895.
5. D. Nie, X. Guo, Y. Duan, R. Zhang, and L. Chen, "WMNav: integrating vision-language models into world models for object goal navigation," arXiv preprint arXiv:2503.02247, 2025.

6. Y. Cong and H. Mo, "An overview of robot embodied intelligence based on multimodal models: Tasks, models, and system schemes," *Int. J. Intell. Syst.*, vol. 2025, no. 1, p. 5124400, 2025.
7. Z. Li, H. Zheng, F. Zhao, A. Chan, J. Zhou, S. Lin, ..., and Q. Wu, "One agent to guide them all: Empowering mllms for vision-and-language navigation via explicit world representation," arXiv preprint arXiv:2602.15400, 2026.
8. S. Zhou, Y. Du, Y. Yang, L. Han, P. Chen, D. Y. Yeung, and C. Gan, "Learning 3d persistent embodied world models," in *Adv. Neural Inf. Process. Syst.*, vol. 38, 2026, pp. 104807–104829.
9. S. Janny, H. Poirier, L. Antsfeld, G. Bono, G. Monaci, B. Chidlovskii, ..., and C. Wolf, "Reasoning in visual navigation of end-to-end trained agents: a dynamical systems approach," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2025, pp. 12111–12121.
10. H. He, Y. Ma, B. Squicciarini, W. Wu, and B. Zhou, "From seeing to experiencing: Scaling navigation foundation models with reinforcement learning," arXiv preprint arXiv:2507.22028, 2025.
11. H. Wang, A. H. Tan, A. Fung, and G. Nejat, "X-Nav: Learning End-to-End Cross-Embodiment Navigation for Mobile Robots," *IEEE Robot. Autom. Lett.*, vol. 11, no. 1, pp. 698–705, 2025.
12. H. MaBouDi, M. Roper, M. G. Guiraud, M. Juusola, L. Chittka, and J. A. Marshall, "A neuromorphic model of active vision shows how spatiotemporal encoding in lobula neurons can aid pattern recognition in bees," *bioRxiv*, 2023.
13. R. Partsey, E. Wijmans, N. Yokoyama, O. Dobosevych, D. Batra, and O. Maksymets, "Is mapping necessary for realistic pointgoal navigation?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 17232–17241.
14. Y. Pan, K. Hu, H. Cao, H. Kang, and X. Wang, "A novel perception and semantic mapping method for robot autonomy in orchards," *Comput. Electron. Agric.*, vol. 219, p. 108769, 2024.
15. P. Karkus, S. Cai, and D. Hsu, "Differentiable slam-net: Learning particle slam for visual navigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2815–2825.