GLAM: TRAINING A LATENT WORLD MODEL OVER GLOBAL SPATIOTEMPORAL MEMORY FOR ACTIVE EXPLORATION AND NAVIGATION

I-Tak Ieong[1,2], **Ruizhi Feng**[2], **Zhaoyang Lu**[2,3], **Yifei Cao**[2,4], **Jiayao Zhao**[2,5], **Leon Li**[†,2], **Senhua Zhu**[2], **Wenbo Ding**[*,1]
[1] Tsinghua University
[2] Lab for Brain-Inspired Embodied Intelligence, EBKernel Technologies Co., Ltd
[3] Northeastern University
[4] University of California, Los Angeles
[5] Shanghai Jiao Tong University
† Project Leader * Corresponding author
Correspondence to: yangyd26@mails.tsinghua.edu.cn

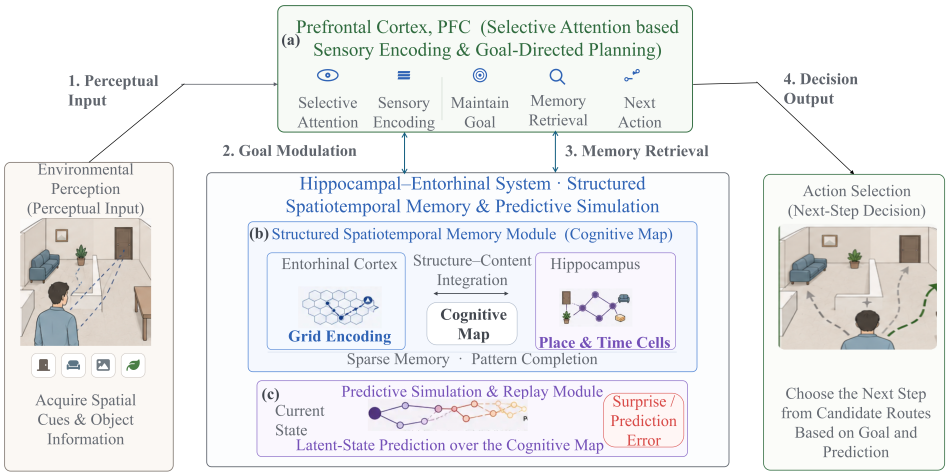

Figure 1: Neurocognitive motivation for active exploration and navigation. The schematic summarizes the functional analogy underlying the proposed formulation: selective attention and goal maintenance guide perceptual encoding, hippocampal–entorhinal memory supports structured spatial representation and predictive simulation, and planning operates over retrieved working memory to select subsequent actions.

ABSTRACT

Active exploration and semantic navigation require an embodied agent to build memory from partial observations, predict how the evolution of observed spatial memory may support future motion, and convert that prediction into actionable plans. We present GLAM, a goal-conditioned latent world model trained over global spatiotemporal memory, and GLAM NAV, the complete navigation system built around it. Given historical map tokens, a navigation goal, and the current robot pose, GLAM jointly predicts future map representations and robot-centric waypoint latents, allowing future spatial context and navigation intent to be inferred in a shared representation space. The model follows a JEPA-like latent prediction paradigm, operates directly on map-level latent tokens rather than RGB reconstruction, and uses a pretrained waypoint encoder–decoder to supervise and

decode navigation plans within GLAM NAV. Training data are collected by replaying ObjectNav expert trajectories in Habitat over HM3D v0.2 scene assets and slicing them into multi-timescale prediction samples. On a controlled HM3D-ObjectNav subset reproduction setting, GLAM NAV improves over a reproduced BSC-Nav baseline in both success rate and success weighted by path length.

1 INTRODUCTION

Object-goal navigation requires an embodied agent to find an instance of a specified object category in an unfamiliar environment from partial egocentric observations. The target may initially be unobserved or leave the field of view as the agent moves, making navigation depend on both accumulated spatial knowledge and decisions about where to explore next. Semantic exploration methods demonstrate the value of explicit maps, object-related priors, and map-grounded subgoal selection (Chaplot et al., 2020a; Ramakrishnan et al., 2022; Yokoyama et al., 2024). These advances motivate a predictive use of memory: beyond retrieving what has already been observed, an agent should anticipate the evolution of observed spatial memory associated with further goal-directed motion.

Latent world models offer a representation-level approach to this problem. Joint-embedding prediction and pretrained-feature dynamics show that future states can be modeled without making pixel reconstruction the primary learning objective (Assran et al., 2023; Zhou et al., 2025). More broadly, predictive spatial representations have long been linked to flexible navigation and planning in cognitive-map accounts (Stachenfeld et al., 2017; Pfeiffer & Foster, 2013; Behrens et al., 2018). These developments motivate a concrete question for object-goal navigation: *how can future spatial-memory prediction and goal-directed waypoint planning be learned jointly from a persistent, spatially indexed map?*

We present GLAM, a goal-conditioned latent world model that jointly predicts future map representations and robot-centric waypoint latents from global spatiotemporal memory. Here, global memory refers to an online representation of the explored environment rather than a complete or pre-built scene map. Given historical map tokens, a navigation goal, and the current robot pose, a shared backbone processes visual and waypoint queries in a single forward pass. The map branch predicts future visual-semantic map features, while the waypoint branch predicts a latent navigation plan. This formulation places spatial prediction and waypoint supervision in a common learning framework without requiring future RGB reconstruction.

The design is functionally motivated by cognitive-map accounts that connect structured spatial memory with prospective navigation (Stachenfeld et al., 2017; Pfeiffer & Foster, 2013). As illustrated in Figure 1, we use persistent spatial representation and memory-based prediction as computational design principles rather than claiming a direct implementation of biological neural circuits. At the system level, we integrate GLAM into GLAM NAV, which combines online mapping, goal-conditioned memory retrieval, observation-based target verification, and geometric motion control. Predicted representations provide prospective context, whereas real observations remain the source of memory updates and target verification.

Training uses expert-controlled ObjectNav trajectories collected in Habitat over HM3D scene assets (Savva et al., 2019; Yadav et al., 2023a). We construct temporally aligned samples containing historical map tokens, future map targets, and expert waypoints at multiple prediction intervals. A waypoint encoder–decoder is pretrained through reconstruction and then frozen, providing a latent supervision space and a corresponding decoding interface. The world model is trained with masked regression objectives for future-map features and waypoint latents. Its predictions therefore concern future observations and plans represented in the expert data, rather than unrestricted counterfactual outcomes for arbitrary action sequences.

We evaluate the resulting navigation system against a reproduced BSC-Nav baseline (Ruan et al., 2026) on the same HM3D-ObjectNav evaluation subset. As reported in Table 1, GLAM NAV increases success rate from 78.50% to 86.89% and SPL from 47.70% to 48.35%. We also provide qualitative real-world demonstrations of online exploration, goal navigation, and reuse of previously acquired spatial memory. These experiments assess the navigation performance of the integrated system and the feasibility of reusing its spatial memory with real sensing and motion.

Our contributions are threefold: (i) a map-grounded latent model that jointly predicts future spatial-memory features and goal-conditioned waypoint latents; (ii) a training formulation combining multi-timescale trajectory supervision with a pretrained waypoint-latent interface; and (iii) an integrated navigation system evaluated through a matched-episode simulation comparison and qualitative real-world demonstrations.

2 Related Work

Goal-directed semantic exploration methods such as SemExp (Chaplot et al., 2020a), PONI (Ramakrishnan et al., 2022), and VLFM (Yokoyama et al., 2024) show that explicit maps and exploration policies are effective for object-goal navigation. Earlier navigation systems and planners, including cognitive mapping and planning, semi-parametric topological memory, target-driven visual navigation, distributed visual navigation policies, vision-and-language navigation, and Active Neural SLAM, established the value of explicit memory, long-horizon subgoal selection, semantically grounded planning, and map-grounded control (Gupta et al., 2017; Savinov et al., 2018; Zhu et al., 2017; Wijmans et al., 2020; Anderson et al., 2018; Chaplot et al., 2020b; Kuipers, 2000). These methods motivate the use of map structure and semantic cues, but they do not explicitly learn future map representations and waypoint plans in a shared latent predictive model.

Structured spatial memory is also central to recent navigation systems. ConceptGraphs (Gu et al., 2024) organizes open-vocabulary scene content into a persistent 3D scene graph, while BSC-Nav (Ruan et al., 2026) builds a brain-inspired spatial intelligence pipeline around landmark and survey memories. These directions are supported by large-scale embodied-AI simulators and 3D scene resources such as Habitat, HM3D-Semantics, and Matterport3D, which make persistent map construction and evaluation practical at scale (Savva et al., 2019; Yadav et al., 2023a;b; Chang et al., 2017). They motivate persistent map structure and retrieval, but they do not directly cast future map evolution and waypoint planning as a joint latent prediction problem over global spatiotemporal memory.

Latent world models provide the predictive side of our design. I-JEPA (Assran et al., 2023) predicts informative target representations rather than pixels, while DINO, MAE, and DINOv2 illustrate the strength of large-scale self-supervised visual features as a substrate for downstream representation learning (Caron et al., 2021; He et al., 2022; Oquab et al., 2024). DINO-WM (Zhou et al., 2025), PlaNet (Hafner et al., 2019), and MuZero (Schrittwieser et al., 2020) show how learned latent dynamics can support planning without reconstructing raw observations at every stage, and Banino et al. (Banino et al., 2018) further show that structured spatial codes can emerge in agents trained for navigation. Our focus is to bring this predictive perspective into semantic navigation by jointly modeling future map tokens and waypoint latents from shared global-memory context.

3 Active Exploration and Navigation System Based on GLaM Nav

Figure 2 illustrates the overall organization of GLaM Nav. At the system level, GLaM Nav integrates four interacting parts: perception and retrieval, structured spatiotemporal memory, the GLaM predictive module, and a planning-and-control layer that grounds latent predictions into executable motion. Throughout the paper, GLaM refers to the learned latent world model, whereas GLaM Nav refers to the complete navigation system built around that model.

The **Perception, Retrieval, and Goal Verification** module is responsible for perceptual encoding, working-memory querying, and target verification. It receives RGB-D observations, the agent pose, and the motion state, and converts them into semantically useful latent features. These features are projected into the global map structure used by the rest of the system. The same module also retrieves goal-relevant map context and decides whether the target has already been verified in current observation or memory. If verified, the system navigates directly toward the retrieved location. Otherwise, it triggers predictive exploration and planning.

The **Spatiotemporal Memory Module** is responsible for memory construction, storage, updating, and retrieval. Selected latent visual features, depth-projected coordinates, instance information, and robot poses are organized into a sparse spatial memory that combines landmark memory with graph-voxel memory. During memory updating, the system compares new observations with existing

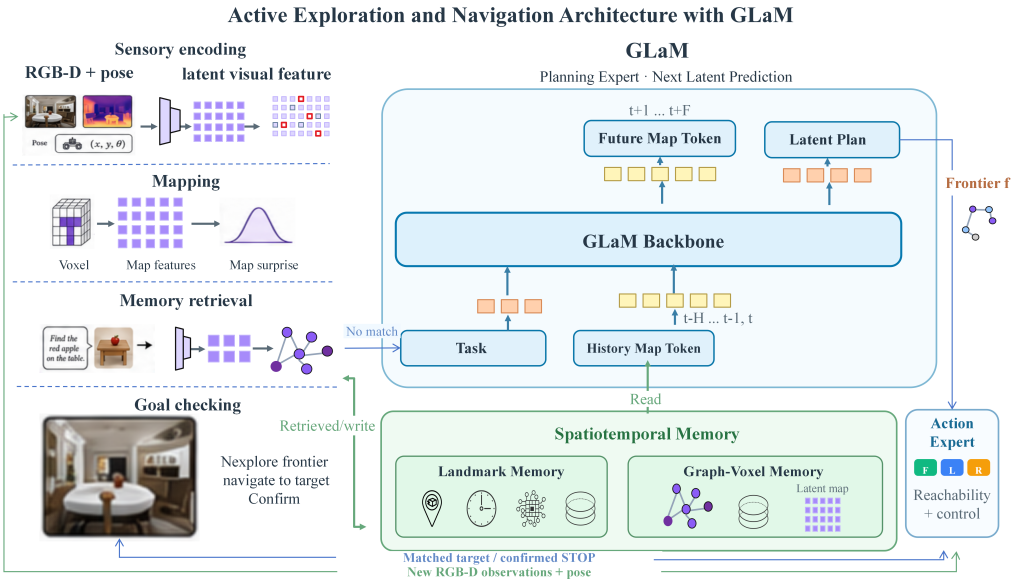

Figure 2: Overview of the GLAM NAV architecture for active exploration and object-goal navigation. The system closes the loop among sensory encoding, mapping, memory retrieval, structured spatiotemporal memory, GLAM-based latent prediction, and action execution. Real RGB-D observations update memory, while the learned predictive module supplies future map tokens and latent plans that guide navigation decisions.

memory and decides whether the corresponding units should be added, fused, or refreshed. During retrieval, semantic information together with its associated latent features is used to identify the most relevant candidate navigation locations from the current memory.

The **GLAM backbone** is the predictive core of GLAM NAV. It takes historical map tokens, the navigation goal, and the current robot pose as input and predicts future map representations together with latent waypoint plans. The model does not reconstruct future RGB frames. Instead, following the latent-space learning logic of the original Cog-WM design, it predicts high-level latent map representations and uses those predictions to support planning. In this formulation, GLAM jointly learns two outputs: future spatial-memory prediction and goal-conditioned waypoint-latent prediction. The first output estimates how global memory is expected to evolve over future horizons. The second output represents navigation plans in a robot-centric latent space that can later be decoded into waypoint increments.

The **Planning Expert** receives retrieved map context and the future map and waypoint predictions produced by GLAM. It proposes or grounds a feasible subgoal in the current observed map and sends executable commands to the low-level controller. This preserves a clear boundary between latent prediction and physical control: GLAM predicts future map structure and navigation intent, while the rest of GLAM NAV remains responsible for geometric execution, reachability checking, and collision-aware motion.

The full navigation loop is as follows. GLAM NAV first surveys its surroundings and updates the online map. It then retrieves relevant regions from landmark memory and graph-voxel memory based on the goal. If the goal has already been verified, the system navigates directly to that location. Otherwise, it uses GLAM to predict future map context and waypoint latents from the current memory state, decodes the waypoint plan, and executes the selected motion branch. After execution, the system updates memory with real observations only, verifies the goal again, and repeats the exploration–retrieval cycle until success or termination.

4 GLaM: Training a Latent World Model over Global Spatiotemporal Memory

As shown in Figure 2, GLaM should be understood as the predictive core inside GLaM Nav. Within the full navigation loop, GLaM receives task condition and historical map tokens, interacts with the spatiotemporal memory modules, and produces two coupled outputs: future map-token prediction and latent planning representations. These outputs provide predictive evidence before action grounding in GLaM Nav, while memory construction, retrieval, and control remain system-level functions.

4.1 Model Training Paradigm

GLaM learns a goal-conditioned relationship between global spatiotemporal memory, future map states, and subsequent navigation waypoints. Its functional organization is inspired by spatial indexing, place–event binding, and predictive replay in the hippocampal–entorhinal system, together with goal-conditioned memory retrieval and planning control associated with the prefrontal cortex (Whittington et al., 2020; Pfeiffer & Foster, 2013; Eichenbaum, 2017). These correspondences motivate the functional organization of the model rather than a direct implementation of biological neural circuits.

Given historical map tokens, a goal description, and the current robot pose, GLaM jointly predicts future map representations and robot-centric waypoint latents over a fixed-length horizon:

$$\left(\widehat{\boldsymbol{Z}}_{t+1:t+F}, \widehat{\boldsymbol{E}}^{w}_{t+1:t+F}\right) = F_{\Theta}(\boldsymbol{Z}_{t-H+1:t}, g, p_t, \boldsymbol{Q}^{v}, \boldsymbol{Q}^{w})\,. \tag{1}$$

Here, $\boldsymbol{Z}_{t-H+1:t}$ denotes the historical map-token sequence, g is the goal description, and p_t is the current robot pose. The parameters H and F specify the history length and prediction horizon, respectively. The learnable queries $\boldsymbol{Q}^{v}$ and $\boldsymbol{Q}^{w}$ correspond to future-map prediction and waypoint-latent prediction. The two outputs, $\widehat{\boldsymbol{Z}}_{t+1:t+F}$ and $\widehat{\boldsymbol{E}}^{w}_{t+1:t+F}$, represent the predicted future map tokens and waypoint latents, respectively.

GLaM follows a JEPA-like latent-space prediction paradigm (Assran et al., 2023): it directly regresses continuous future representations rather than reconstructing future RGB observations. Visual and waypoint queries are supplied together in a single forward pass, allowing both outputs to be predicted from the same memory, goal, and pose context. The predicted waypoint latents are subsequently decoded into normalized planar pose increments, as described in Section 4.3. We evaluate the practical utility of this joint predictive representation through the downstream tasks of active exploration and semantic navigation.

4.2 Training Datasets

Training data are collected by replaying ObjectNav expert trajectories in Habitat using HM3D v0.2 scene assets. HM3D provides large-scale reconstructed indoor environments for simulated semantic-navigation interactions, including RGB-D observations, robot poses, and navigation meshes (Yadav et al., 2023b).

During collection, each episode preserves the original scene, start pose, goal category, and expert rollout defined by the benchmark. Along each expert trajectory, we record RGB-D observations, robot poses, executed actions, and the map features derived from these observations, while also storing the frontier sequence encountered during exploration together with the updated map features associated with frontier progression. We then slice the replayed trajectories into training pairs with historical map tokens, future map targets, and future waypoint latents. This keeps the description of data construction explicit while maintaining the same sample definition used by the model.

The collected subset contains 577 ObjectNav episodes across 145 scenes, comprising 55,392 recorded interaction steps. At an observation rate of 1 frame per second, these steps correspond to approximately 15.4 hours of observations. Slicing the trajectories with multiple prediction intervals and history windows produces 12,101 multi-timescale prediction samples. These statistics describe the subset collected for this study rather than the full HM3D dataset.

4.3 MAP-TOKEN AND WAYPOINT LATENT ENCODING

Map-token encoding. Map tokens are constructed from four RGB-D views. Visual features are extracted from the observations, projected into a voxel grid, and fused within each voxel. Voxel coordinates are expressed in the current robot frame and encoded using three-dimensional spatial rotary positional embeddings (RoPE). The resulting token sequences are truncated or zero-padded to a fixed length, with validity masks identifying non-padded positions. A linear input projection maps the visual features to the 2560-dimensional representation used by the backbone.

Waypoint latent encoding. Each waypoint is represented as a robot-centric planar pose increment $(\Delta x, \Delta y, \Delta\theta)$. We normalize its translation and encode its heading using a sine–cosine representation:

$$\begin{aligned} \boldsymbol{u} &= \left[\mathrm{clip}\left(\frac{\Delta x}{s_x}, -1, 1\right), \mathrm{clip}\left(\frac{\Delta y}{s_y}, -1, 1\right), \sin\Delta\theta, \cos\Delta\theta\right]^\top, \\ \boldsymbol{e}^w &= \phi(\boldsymbol{u}) \in \mathbb{R}^{2560}. \end{aligned} \tag{2}$$

The translation scales s_x and s_y are estimated exclusively from the training split. The encoder ϕ is a lightweight two-layer multilayer perceptron with dimensions $4 \to d_a \to 2560$, followed by LayerNorm, where d_a denotes the hidden-layer width.

Before world-model training, ϕ is jointly pretrained with a lightweight decoder ρ by minimizing the mean squared error between normalized waypoints $\boldsymbol{u}$ and their reconstructions $\rho(\phi(\boldsymbol{u}))$. This pretraining is independent of the world model. Both modules are subsequently frozen: the encoder provides waypoint-latent supervision during world-model training, and the decoder converts predicted latents into normalized waypoint representations at inference:

$$\widehat{\boldsymbol{u}}_j = \rho\big(\widehat{\boldsymbol{e}}_j^w\big)\,. \tag{3}$$

Inference uses the same decoder ρ that was pretrained jointly with ϕ.

4.4 TRAINING LOSS DESIGN

During world-model training, Qwen3.5-0.8B backbone, input projections, learnable queries, and two regression heads are optimized. The objective combines masked mean squared errors for future-map prediction and waypoint-latent prediction:

$$\mathcal{L} = \lambda_v\ \mathrm{MSE}_{m^v}\Big(\widehat{\boldsymbol{Z}}, \boldsymbol{Z}^*\Big) + \lambda_w\ \mathrm{MSE}_{m^w}\Big(\widehat{\boldsymbol{E}}^w, \phi(\boldsymbol{U}^*)\Big)\,. \tag{4}$$

Here, $\boldsymbol{Z}^* = \{\boldsymbol{z}_i^*\}$ contains future-map targets produced by the frozen visual encoder and deterministic map-construction pipeline. The sequence $\boldsymbol{U}^* = \{\boldsymbol{u}_j^*\}$ contains normalized expert waypoints, with ϕ applied to each waypoint to obtain its target latent representation. The binary masks m_i^v and m_j^w identify valid visual and waypoint target positions, respectively, while λ_v and λ_w control the relative contributions of the two objectives.

For d-dimensional target vectors $\boldsymbol{x}_\ell^*$ and a binary validity mask m, masked MSE is defined as

$$\mathrm{MSE}_m\Big(\widehat{\boldsymbol{X}}, \boldsymbol{X}^*\Big) = \frac{\sum_\ell m_\ell\, \|\widehat{\boldsymbol{x}}_\ell - \boldsymbol{x}_\ell^*\|_2^2}{d \sum_\ell m_\ell}, \qquad \sum_\ell m_\ell > 0. \tag{5}$$

Thus, each objective is normalized by both its number of valid target positions and the dimensionality of its target representation. The waypoint-latent loss uses $d = 2560$, whereas the map-prediction loss uses the dimensionality of the future-map target features. Both objectives supervise continuous representations directly; no language-model cross-entropy term is used.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTING

We evaluate GLAM in simulation on the HM3D ObjectNav benchmark (Yadav et al., 2023b), a standard embodied object-navigation benchmark in the Habitat ecosystem built on HM3D-Semantics

Table 1: HM3D-ObjectNav results on the 50% subset reproduction setting. Relative improvements are computed with respect to the reproduced BSC-Nav baseline on the same evaluation episodes.

Method	SR (%)	SPL (%)
BSC-Nav Ruan et al., 2026	78.50	47.70
GLAM NAV (ours)	86.89	48.35
Relative improvement	10.69% ↑	1.36% ↑
Absolute gain	+8.39 pts	+0.65 pts

v0.2 (Yadav et al., 2023a). The benchmark contains 216 semantically annotated real-world indoor 3D reconstructions, split into 145 training scenes, 36 validation scenes, and 35 test scenes. Scenes cover apartments, offices, and other multi-room indoor layouts, and target categories include six common indoor objects such as chair, sofa, potted plant, bed, toilet, and TV. Each episode specifies a scene, an initial robot pose, a target category, and corresponding target instances. During evaluation, the agent starts from a random pose and orientation and must complete exploration, target retrieval, and path planning from RGB-D observations without access to a pre-built map or privileged target coordinates.

Following the standard ObjectNav protocol, an episode is counted as successful when the agent executes the STOP action within the allowed budget, is within $1.0\,\mathrm{m}$ of a valid target instance, and the target is observable from the stopping location. To enable a controlled comparison against a strong brain-inspired baseline and to reduce reproduction overhead, we evaluate on a 50% subset of the HM3D benchmark under the same split and protocol. We reproduce BSC-Nav (Ruan et al., 2026) as the reference method and then run GLAM on the same evaluation episodes, so that both methods are compared under matched scene distributions, target categories, and task difficulty. Under this subset reproduction setting, we do not directly compare the reported numbers with methods evaluated on the full benchmark, different HM3D versions, different input modalities, or different evaluation protocols.

We report Success Rate (SR) and Success weighted by Path Length (SPL). SR measures the fraction of episodes in which the agent successfully reaches the target and stops correctly within the allowed budget, and therefore primarily reflects whether the target can be found reliably. SPL additionally accounts for path efficiency, so it decreases when the agent succeeds but follows an unnecessarily long route. In addition to these primary metrics, final reports should include path length to first verified target observation, collision rate, timeout rate, false stop rate, and runtime including both the frozen feature service and the trainable predictor. For the training set itself, reporting should cover the number of scenes, replayed episodes, recorded interaction steps, and resulting multi-timescale prediction samples. Where additional learner-visited states are introduced in future data aggregation experiments, they should be reported separately rather than merged into the original expert-only manifest, following the logic of no-regret dataset aggregation in imitation learning (Ross et al., 2011).

5.2 HM3D-OBJECTNAV RESULTS

Table 1 reports higher SR and SPL for GLAM NAV than for the reproduced BSC-Nav baseline on the matched evaluation subset. SR increases from 78.50% to 86.89%, a gain of 8.39 percentage points, while SPL increases from 47.70% to 48.35%, a gain of 0.65 percentage points. These point estimates indicate a larger improvement in task completion than in the aggregate path-efficiency metric. They do not, by themselves, establish shorter trajectories on commonly successful episodes or isolate the contribution of future-map prediction. Paired trajectory analysis and component ablations are required to assess those explanations.

5.3 QUALITATIVE REAL-WORLD DEMONSTRATIONS

We further evaluate whether the exploration and navigation behavior of GLAM NAV transfers from simulation to a physical indoor environment. The system was deployed in an open-plan research office on an Astribot S1 wheeled dual-arm humanoid equipped with an NVIDIA Jetson AGX Orin,

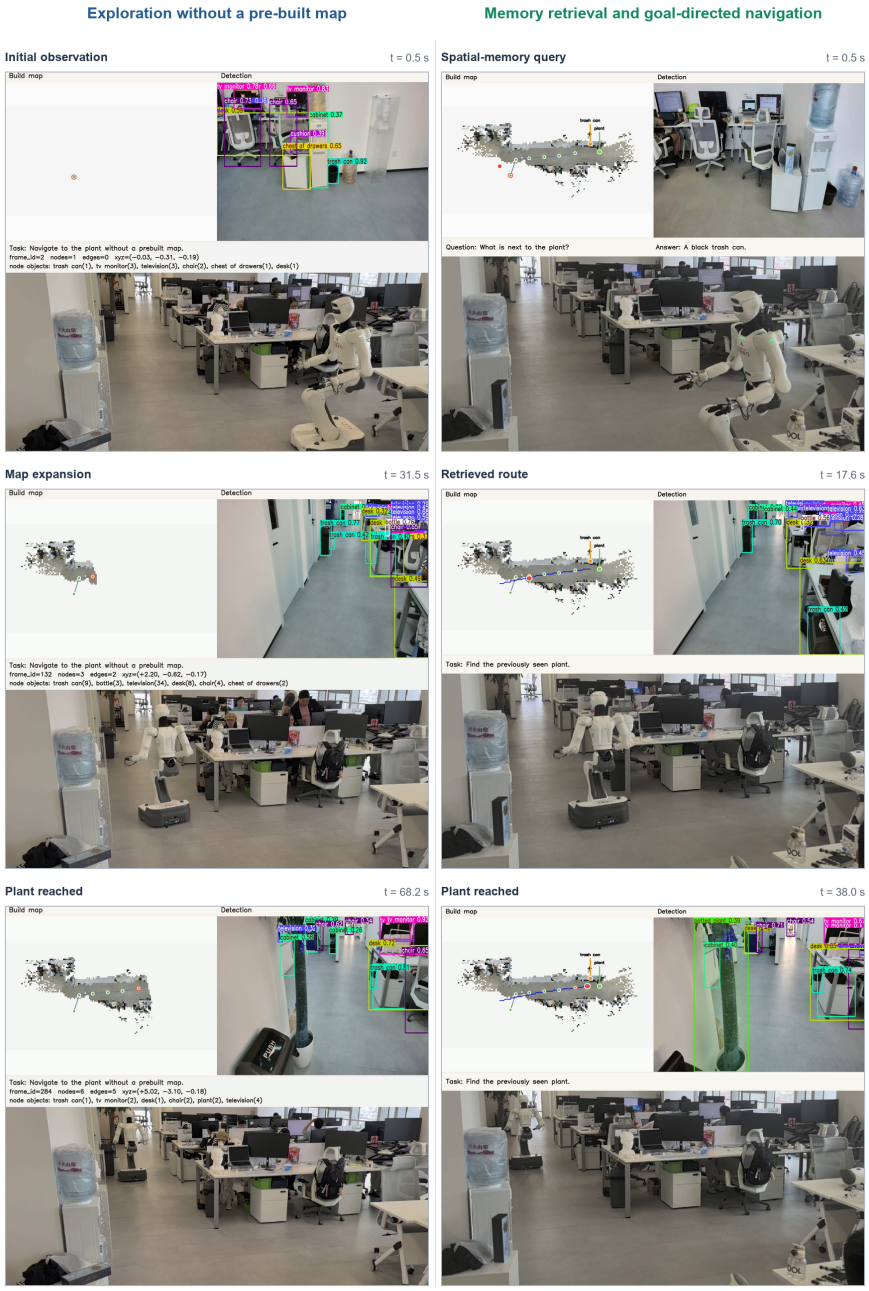

Figure 3: Real-world deployment of GLAM NAV on an Astribot S1. Left: the robot explores without a pre-built map, expands its spatiotemporal memory, and reaches the plant region. Right: the system answers a spatial-memory query, retrieves the historical plant location, and navigates back to it. Each frame combines the online map, first-person semantic perception, and an external view of the robot.

an RGB-D camera, LiDAR, and proprioceptive sensors. Our implementation separates the body-independent global memory and latent prediction modules from the body-specific motion controller. The predictive model therefore communicates with the robot through a metric navigation interface, while the platform controller executes the resulting motion commands. The experiments reported here validate this interface on Astribot S1.

5.3.1 Active Exploration and Semantic Navigation

The first experiment instructs the robot to navigate to a plant in an initially unknown environment. The robot receives no pre-built 2D or 3D map and no target coordinates. It incrementally associates visual observations with spatial nodes while moving through the office, producing a global spatiotemporal memory whose topology expands with the explored free space. As shown in the left column of Figure 3, the map grows from a single initial node to a connected route containing newly observed semantic landmarks. The robot then reaches the plant region using the same online memory for localization, retrieval, and planning. This sequence demonstrates that GLaM Nav can couple map construction with goal-directed motion in a real environment rather than relying on an environment-specific map prepared before deployment.

5.3.2 Spatiotemporal Memory Retrieval and Spatial Question Answering

The second experiment tests whether information acquired during exploration remains useful after the target leaves the current field of view. The robot is first asked what is located next to the plant. By retrieving the stored semantic and spatial relations, the system answers that a black trash can is beside the plant. The robot then receives the instruction to find the previously observed plant. As shown in the right column of Figure 3, GLaM Nav retrieves the corresponding historical location, converts it into a route over the accumulated memory, and navigates back to the target region. The result shows that the map stores more than an instantaneous detector output: it supports persistent object-location associations and local spatial relations that can be reused for both navigation and question answering.

Together, the two experiments provide complementary evidence for real-world operation. The first evaluates online memory construction under active motion, while the second evaluates delayed retrieval and reuse of previously acquired spatial knowledge. Although a broader quantitative study across buildings and robot embodiments remains necessary, these demonstrations show that the same memory and prediction interface used in simulation can operate with real sensing and robot motion without a pre-built map.

5.4 Ablation Studies

Detailed ablation studies and additional experimental results will be released in a subsequent public version of the manuscript and project materials.

6 Limitations

Prediction scope and data coverage. GLaM learns from expert-controlled trajectories, so its supervision covers the states and behaviors represented in those rollouts rather than all states that the deployed policy may encounter. Recovery from navigation errors, unfamiliar exploration patterns, and alternative routes may therefore require experience not represented in the current training set. Moreover, the model conditions on memory, a goal, and pose rather than an externally specified future action sequence. Its joint predictions should consequently be interpreted as goal-conditioned map and waypoint estimates, not as a general simulator of arbitrary action-conditioned futures.

Uncertainty and control feasibility. The current model produces deterministic latent predictions over a fixed horizon. Under partial observability, the same observed memory may admit multiple unseen layouts and feasible routes, but the regression objective does not explicitly represent these alternatives or calibrate predictive uncertainty. Furthermore, low map-feature or waypoint-latent error does not guarantee accurate target localization, collision-free motion, or consistency between the two predicted outputs. The system therefore relies on observation-based verification and geometric control rather than treating a latent prediction as a verified environmental fact.

Memory fidelity and capacity. Spatial memory depends on visual features, depth projection, and pose alignment. Errors in these components can affect both retrieval and the context supplied to the predictor. Although the stored map is persistent, the model accesses a bounded token representation; truncation may omit spatial detail or previously observed evidence as exploration expands. This creates a trade-off between memory coverage, token resolution, and inference cost. The present

evaluation does not establish robustness to substantial localization drift, persistent scene changes, or substantially larger environments.

7 Conclusion

GLAM combines future-map prediction and goal-conditioned waypoint-latent prediction in a single latent world model over global spatiotemporal memory. By training on replayed HM3D ObjectNav trajectories and supervising both future map tokens and waypoint latents, the model connects memory, prediction, and planning in one representation space. The HM3D-ObjectNav subset evaluation suggests that this formulation improves target-search reliability while maintaining competitive path efficiency against a strong structured-memory baseline.

References

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Suenderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. doi: 10.1109/CVPR.2018.00387.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, June 2023. doi: 10.1109/CVPR52729.2023.01499.

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J. Chadwick, Thomas Degris, Joseph Modayil, Greg Wayne, Hubert Soyer, Fabio Viola, Brian Zhang, Ross Goroshin, Neil Rabinowitz, Razvan Pascanu, Charlie Beattie, Stig Petersen, Amir Sadik, Stephen Gaffney, Helen King, Koray Kavukcuoglu, Demis Hassabis, Raia Hadsell, and Dharshan Kumaran. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429–433, 2018. doi: 10.1038/s41586-018-0102-6.

Timothy E. J. Behrens, Thomas H. Muller, Jill C. R. Whittington, Stephannie Mark, Aleksandra B. Baram, Kimberly L. Stachenfeld, and Zeb Kurth-Nelson. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, 100(2):490–509, 2018. doi: 10.1016/j.neuron.2018.10.002.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, 2021. doi: 10.1109/ICCV48922.2021.00951.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017. doi: 10.1109/3DV.2017.00081. URL https://doi.org/10.1109/3DV.2017.00081.

Devendra Singh Chaplot, Dhiraj Gandhi, Abhinav Gupta, and Ruslan Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4247–4259, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/2c75cf2681788adaca63aa95ae028b22-Abstract.html.

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*, 2020b. URL https://openreview.net/forum?id=HklXn1BKDH.

Howard Eichenbaum. Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, 18:547–558, 2017. doi: 10.1038/nrn.2017.74.

Qiao Gu, Alihusein Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, Chuang Gan, Celso Miguel de Melo, Joshua B. Tenenbaum, Antonio Torralba, Florian Shkurti, and Liam Paull. ConceptGraphs: Open-vocabulary 3D scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028, 2024. doi: 10.1109/ICRA57147.2024.10610243.

Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7272–7281, 2017. doi: 10.1109/CVPR.2017.769.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565. PMLR, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022. doi: 10.1109/CVPR52688.2022.01553.

Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1–2):191–233, 2000. doi: 10.1016/S0004-3702(00)00017-5.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a68SUt6zFt.

Brad E. Pfeiffer and David J. Foster. Hippocampal place-cell sequences depict future paths to remembered goals. *Nature*, 497:74–79, 2013. doi: 10.1038/nature12112.

Santhosh Kumar Ramakrishnan, Devendra Singh Chaplot, Ziad Al-Halah, Jitendra Malik, and Kristen Grauman. PONI: Potential functions for objectgoal navigation with interaction-free learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18823–18832, 2022. doi: 10.1109/CVPR52688.2022.01832.

Stephane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 627–635. PMLR, 2011. URL https://proceedings.mlr.press/v15/ross11a.html.

Shouwei Ruan, Liyuan Wang, Caixin Kang, Qihui Zhu, Songming Liu, Xingxing Wei, and Hang Su. Brain-inspired spatial intelligence for embodied agents. *Nature Communications*, 17:8061, 2026. doi: 10.1038/s41467-026-74358-5.

Nikolay Savinov, Alexey Dosovitskiy, and Vladlen Koltun. Semi-parametric topological memory for navigation in large-scale environments. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HkeCWnAqKQ.

Manolis Savva, Jitendra Malik, Devi Parikh, Dhruv Batra, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, and Vladlen Koltun. Habitat: A platform for embodied AI research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019. doi: 10.1109/ICCV.2019.00943. URL https://doi.org/10.1109/ICCV.2019.00943.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. MuZero: Mastering go, chess, shogi and atari without rules. *Nature*, 588(7839): 604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Kimberly L. Stachenfeld, Matthew M. Botvinick, and Samuel J. Gershman. The hippocampus as a predictive map. *Nature Neuroscience*, 20(11):1643–1653, 2017. doi: 10.1038/nn.4650.

James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. The Tolman–Eichenbaum Machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263.e23, 2020. doi: 10.1016/j.cell.2020.10.024.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. DD-PPO: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1gX8C4YPr.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-matterport 3D semantics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4927–4936, 2023a. doi: 10.1109/CVPR52729.2023.00477.

Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, and Devendra Singh Chaplot. Habitat-Matterport 3D Semantics Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4927–4936, 2023b. doi: 10.1109/CVPR52729.2023.00477.

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. 15684–15691, 2024. doi: 10.1109/ICRA57147.2024.10610712.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79115–79135. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhou25t.html.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J. Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017. doi: 10.1109/ICRA.2017.7989381. URL https://doi.org/10.1109/ICRA.2017.7989381.