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

Anonymous authors
Paper under double-blind review

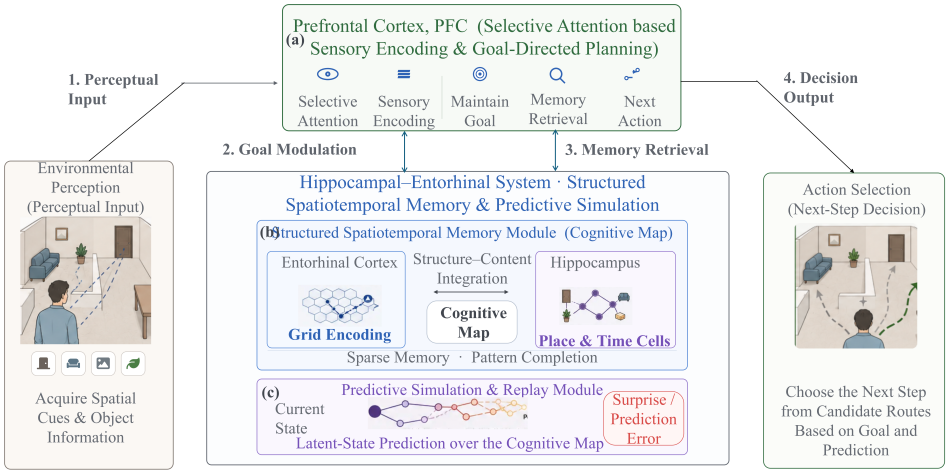


Figure 1: Neurocognitive motivation for active exploration and navigation. The schematic summarizes the functional analogy underlying the proposed formulation: selective attention and goal maintenance guide perceptual encoding, hippocampal–entorhinal memory supports structured spatial representation 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 in Habitat over HM3D v0.2 by sampling oracle-guided ObjectNav trajectories and their map transitions. On the full HM3D-Semantics v0.2 ObjectNav benchmark, GLAM NAV achieves a state-of-the-art success rate of 82.30% with 34.00% success weighted by path length.

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.

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 predictive model that jointly estimates future map representations and robot-centric waypoint latents from global spatiotemporal memory. Here, global memory is the online state accumulated from partial observations, not a complete or pre-built scene map. Given historical map tokens, a navigation goal, and the current robot pose, GLAM processes future-map and waypoint queries in a single forward pass. The two prediction branches provide prospective spatial context and navigation intent without reconstructing future RGB observations. GLAM is therefore the learned predictive component rather than the complete navigation planner.

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 principles rather than claiming a direct implementation of biological circuits. The complete GLAM NAV system combines a High-Level Planning Module, spatiotemporal memory, GLAM, and an Action Expert. High-level planning performs sensory encoding, mapping, retrieval, and goal checking; GLAM is invoked when the target cannot be grounded; and the Action Expert converts a retrieved target or predicted frontier into reachable motion. Only real observations update memory and authorize STOP.

Training uses oracle-guided ObjectNav trajectories collected in Habitat over HM3D scene assets (Savva et al., 2019; Yadav et al., 2023). We construct temporally aligned samples containing historical map tokens, future map targets, and supervised waypoints. 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 collected trajectories, rather than unrestricted counterfactual outcomes for arbitrary action sequences.

We evaluate the complete GLAM NAV system on the full HM3D-Semantics v0.2 ObjectNav benchmark. GLAM NAV obtains 82.30% SR and 34.00% SPL, compared with 78.50% SR and 47.70% SPL for our BSC-Nav reproduction on the same complete episode set. As shown in Table 1, GLAM NAV also achieves the highest SR among the reported full-benchmark results from SkillNav (Sang et al., 2026), Uni-NaVid (Zhang et al., 2025), Qwen-RobotNav (Zhang et al., 2026), and related methods. The gain in successful target acquisition is accompanied by a path-efficiency gap. Diagnostic ablations examine GLAM-guided frontier selection and High-Level Planning Module orchestration, and qualitative real-world demonstrations test online exploration and delayed memory reuse.

Our contributions are threefold: (i) GLAM, a map-grounded predictive model that jointly estimates 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) GLAM NAV, an integrated system evaluated on the full HM3D-Semantics v0.2 benchmark, through planner-level ablations, and in 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., 2023; 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. Recent foundation-model navigation systems span language-guided frontier selection, open-set grounding, scene-graph prompting, predictive world models, hierarchical review, behavioral skill intervention, cognitive-state modeling, and unified video-to-action policies (Yu et al., 2023; Yokoyama et al., 2024; Kuang et al., 2024; Yin et al., 2024; Nie et al., 2025; Qi et al., 2026; Sang et al., 2026; Cao et al., 2025; Zhang et al., 2025). Qwen-RobotNav further unifies navigation tasks through task-adaptive observation context and waypoint prediction (Zhang et al., 2026). Our focus is instead joint prediction of future map tokens and waypoint latents from shared global-memory context.

3 ACTIVE EXPLORATION AND NAVIGATION SYSTEM BASED ON GLAM NAV

Figure 2 organizes GLAM NAV into the High-Level Planning Module, spatiotemporal memory, the GLAM predictive model, and the Action Expert. GLAM denotes only the learned latent predictor; GLAM NAV denotes the complete closed loop that manages memory, routes decisions, and executes motion.

The **High-Level Planning Module** contains the four stages in the left panel. Sensory encoding maps RGB-D observations and poses to latent visual features. Mapping projects them into voxels, fuses map features, and estimates map surprise. Memory retrieval queries landmark and graph-voxel context using the target, and goal checking routes a match to execution, an uncertain candidate to confirmation, or an unmatched target to GLAM-guided exploration.

The **Spatiotemporal Memory** stores semantic, spatial, and temporal associations in landmark memory and maintains connectivity and latent map features in graph-voxel memory. The High-Level Planning Module retrieves from it and writes updates from real observations. GLAM reads a bounded map-token history, but predicted future tokens are not written back as observations.

The **GLAM Predictive Model** is invoked when the target cannot be grounded. Given target g, the current pose, and map tokens from $t - H + 1$ through t, it jointly predicts future map tokens and a robot-centric latent plan over $t + 1{:}t + F$. The decoded plan is associated with frontier f, while the

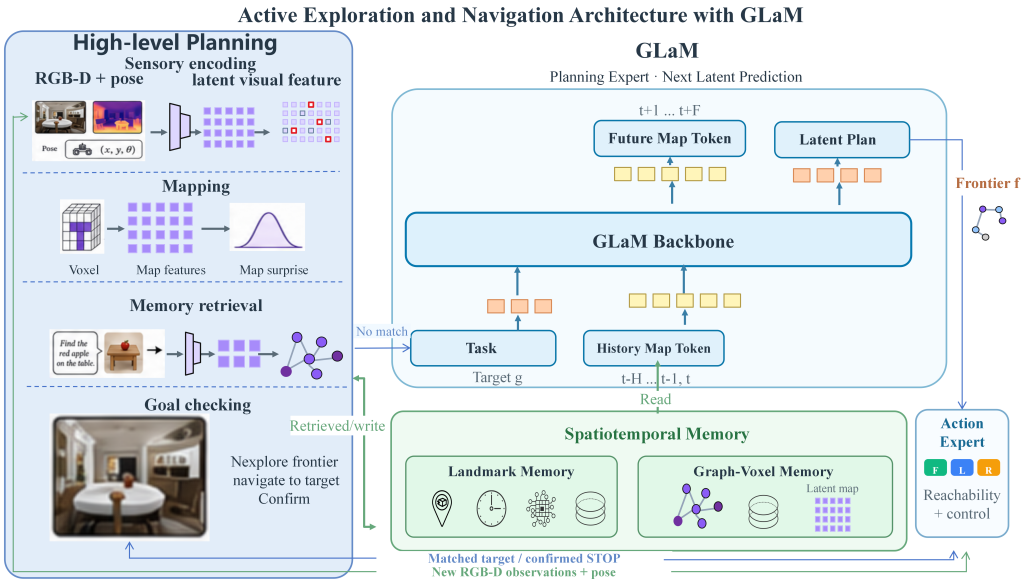


Figure 2: Overview of GLAM NAV. The High-Level Planning Module performs perception, mapping, retrieval, and goal checking. The GLAM predictive model produces future map tokens and a latent plan, and the Action Expert grounds a retrieved target or predicted frontier into reachable motion. Only real observations and poses update spatiotemporal memory.

map branch supplies prospective context. Retrieval, verification, reachability, and control remain outside GLAM.

The **Action Expert** grounds either a retrieved target or a GLAM-proposed frontier through reachability checking and collision-aware control. New observations and poses return to the High-Level Planning Module, update memory, and start the next retrieval cycle. STOP is issued only after goal checking confirms the target from real sensory evidence.

The resulting loop alternates between evidence-driven navigation and predictive exploration. At each cycle, the High-Level Planning Module first updates the observed map and searches memory for the target. A successful match is grounded directly by the Action Expert; otherwise, GLAM reads the current task and map-token history to propose future spatial context and a frontier-directed latent plan. After the Action Expert executes the grounded motion, the next real observation replaces prediction as the basis for retrieval and verification. This read–predict–act–observe structure keeps long-horizon prediction separate from persistent memory and physical control.

4 GLAM: TRAINING A LATENT WORLD MODEL OVER GLOBAL SPATIOTEMPORAL MEMORY

As shown in Figure 2, GLAM is the predictive model inside GLAM NAV, not the overall planner or controller. It maps the task, current pose, and historical map tokens to future map-token predictions and latent plans. The High-Level Planning Module remains responsible for perception, memory operations, verification, and routing, while the Action Expert handles reachability and execution.

4.1 PREDICTIVE-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 Data and Supervision

We construct the supervision for GLaM from oracle-guided ObjectNav trajectories collected in Habitat with HM3D v0.2 scene assets (Yadav et al., 2023). RGB-D observations sampled along a shortest path to a reachable goal viewpoint are fused into the same geometric and visual-semantic memory representation used by the model. Each labeled segment records the memory before motion, the memory obtained after reaching the next frontier or goal viewpoint, and the corresponding displacement. Historical memory states provide the conditioning tokens in Equation 1; arrival states provide future-map targets, and the displacements provide waypoint targets. Consecutive segments are grouped over multiple temporal horizons to form the joint map-and-waypoint training sequences optimized by Equation 4. Oracle paths are used only to construct supervision; navigation at evaluation time remains closed-loop.

The resulting dataset contains 11,000 trajectories across 145 scenes, comprising 10,000 training trajectories and 1,000 held-out internal trajectories, 484,131 map-observation events, 495,131 RGB-D frames, and 27,428 supervised segments. Because both trajectory splits use the same scenes, the internal split separates held-out trajectories but does not measure unseen-scene navigation. These collection splits are distinct from the navigation evaluation episodes in Section 5. Appendix A specifies the observation schedule, voxel construction, frontier labels, and split statistics.

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-VL 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 supervised 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 the complete GLaM Nav system in simulation on the HM3D ObjectNav benchmark (Yadav et al., 2023), using model weights learned from the map-transition supervision in Section 4.2. HM3D ObjectNav is built on HM3D-Semantics v0.2 and contains 216 semantically annotated indoor reconstructions, split into 145 training scenes, 36 validation scenes, and 35 test scenes. Each episode specifies a scene, an initial robot pose, a target category, and valid target instances. Unlike the oracle-guided training collector, the evaluated agent must perform exploration, retrieval, and planning online from RGB-D observations without a pre-built map or privileged target coordinates. The internal trajectory split reported in Section 4.2 is not used as the navigation evaluation set described here.

Following the standard ObjectNav protocol, an episode is counted as successful when the agent executes STOP within the allowed budget, is within $1.0\,\mathrm{m}$ of a valid target instance, and the target is observable from the stopping location. The main evaluation uses the complete HM3D-Semantics v0.2 ObjectNav evaluation set. We reproduce BSC-Nav (Ruan et al., 2026) as the reference and run the complete GLaM Nav system on exactly the same full episode list, including all starts, target categories, and scenes in the benchmark evaluation. This navigation evaluation is independent of the internal trajectory split used during data collection. Results from other methods are taken from their full-benchmark reports under HM3D-Semantics v0.2.

We report Success Rate (SR) and Success weighted by Path Length (SPL). SR measures the fraction of episodes in which the agent reaches a valid target and stops successfully within the episode budget. SPL additionally penalizes successful trajectories that are longer than the corresponding shortest path.

Table 1: Full HM3D-Semantics v0.2 ObjectNav results. Published values follow the original papers and the consolidated comparison of Sang et al. (2026); BSC-Nav and GLAM NAV are evaluated on the same complete episode set.

Method	SR (%) ↑	SPL (%) ↑
L3MVN Yu et al., 2023	36.30	15.70
SG-Nav Yin et al., 2024	49.60	25.50
VLFM Yokoyama et al., 2024	62.60	31.00
Qwen-RobotNav-8B Zhang et al., 2026	71.20	33.00
WMNav Nie et al., 2025	72.20	33.30
Uni-NaVid Zhang et al., 2025	73.70	37.10
MerNav Qi et al., 2026	74.80	37.60
Qwen-RobotNav-4B Zhang et al., 2026	75.60	30.60
SkillNav Sang et al., 2026	75.90	43.20
BSC-Nav Ruan et al., 2026	78.50	**47.70**
GLAM NAV (ours)	**82.30**	34.00

5.2 HM3D-OBJECTNAV RESULTS

Table 1 compares full-benchmark results on HM3D-Semantics v0.2. GLAM NAV achieves the highest SR at 82.30%, exceeding reproduced BSC-Nav by 3.80 points, SkillNav by 6.40 points, and Qwen-RobotNav-4B by 6.70 points. Its SPL of 34.00% is below BSC-Nav by 13.70 points and SkillNav by 9.20 points, indicating that the principal gain is broader target coverage rather than shorter successful trajectories. The exactly matched BSC-Nav reproduction controls the episode list and evaluation protocol, while the remaining published results provide benchmark-level comparison. This success–efficiency trade-off motivates the waypoint-guidance and High-Level Planning Module diagnostics in Section 5.4.

5.3 QUALITATIVE REAL-WORLD DEMONSTRATIONS

We deploy GLAM NAV in an open-plan office on an Astribot S1 equipped with an NVIDIA Jetson AGX Orin, an RGB-D camera, LiDAR, and proprioceptive sensors. Body-independent memory and prediction communicate with the platform controller through a metric navigation interface. We qualitatively test online exploration and subsequent reuse of the acquired spatial memory.

5.3.1 ACTIVE EXPLORATION AND SEMANTIC NAVIGATION

The robot is asked to find a plant without a pre-built map or target coordinates. As it moves, visual observations expand a connected spatial memory containing newly observed landmarks. The left column of Figure 3 shows the map growing from the initial node until the robot reaches the plant region, demonstrating coupled online mapping and goal-directed motion.

5.3.2 SPATIOTEMPORAL MEMORY RETRIEVAL AND SPATIAL QUESTION ANSWERING

The second experiment tests delayed memory reuse. The system retrieves the stored relation that a black trash can is beside the plant and, when asked to revisit the plant, converts its historical location into a route over the accumulated memory. The right column of Figure 3 illustrates this return trajectory, showing that stored object locations and local spatial relations remain available after they leave the current view.

Together, the demonstrations test online memory construction and delayed retrieval with real sensing and motion. Broader quantitative evaluation across environments and embodiments remains necessary.

5.4 ABLATION STUDIES

The main SR–SPL trade-off motivates diagnostics at two levels of GLAM NAV: frontier proposal by the GLAM predictive model and retrieval, verification, and grounding in the High-Level Planning Module. These experiments use a separate 4B variant and 100 validation episodes from four HM3D scenes, so they identify qualitative mechanisms rather than numerically decomposing the main result. Decoded waypoint guidance improves SR by 6.00 points but lowers SPL by 0.66 points.

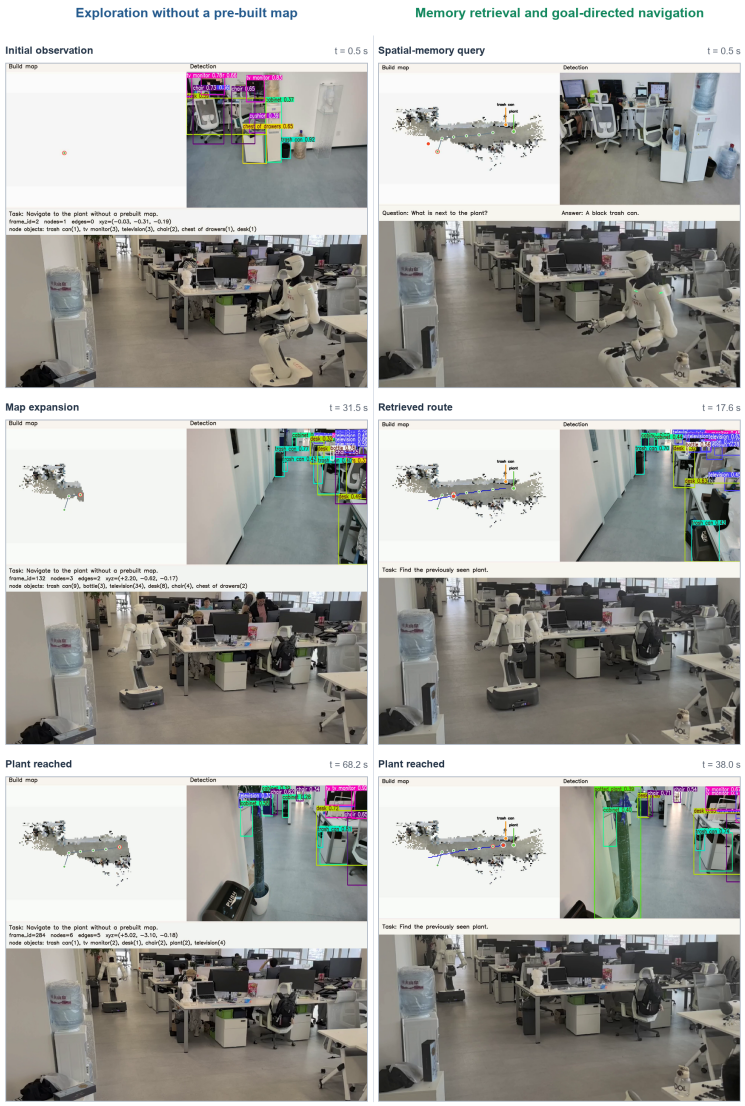


Figure 3: Real-world deployment of GLAM NAV on an Astribot S1. Left: the robot explores without a pre-built map, expands its spatiotemporal memory, and reaches the plant region. Right: the system answers a spatial-memory query, retrieves the historical plant location, and navigates back to it. Each frame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

Because future map tokens do not directly score frontiers, this comparison evaluates the latent-plan branch rather than isolating future-map supervision.

With the GLAM checkpoint fixed, the remaining ablations modify orchestration within the High-Level Planning Module. Querying memory after the arrival scan improves SPL by 2.04 points without changing SR; path-cost gating has negligible effect, and an additional RGB-based STOP check reduces SR by 11.00 points. A separate reference with revised cluster scoring and failed-target handling obtains 82.00% SR and 27.82% SPL. Grounding a semantic target to a reachable observation viewpoint improves SR/SPL by 2.00/0.95 points, whereas support-normalized cluster scoring degrades both. These results favor timely retrieval and reachable subgoals, but interventions are tested independently and gains across groups are not additive.

Table 2: Diagnostic ablations on 100 HM3D validation episodes. Interventions are evaluated independently from the reference configuration in their group.

Variant	SR (%) ↑	SPL (%) ↑
Waypoint guidance		
Default frontier selection	77.00	**29.45**
GLAM waypoint-latent guidance	**83.00**	28.79
+ Path-cost gating	**83.00**	28.63
+ Post-arrival retrieval	**83.00**	**30.83**
+ RGB-based STOP check	72.00	23.39
Revised cluster scoring and failed-target handling	82.00	27.82
+ Reachable-viewpoint grounding	**84.00**	**28.77**
+ Support-normalized cluster score	49.00	15.62

6 LIMITATIONS

Evaluation and supervision scope. The main navigation result is measured on the full HM3D-Semantics v0.2 ObjectNav evaluation set. BSC-Nav and GLAM NAV share the same episode list and implementation-level protocol, whereas comparisons with other published methods may still reflect differences in input modalities, implementation details, or controller configurations. In addition, GLAM is supervised near oracle shortest paths rather than across the full state distribution of the deployed planner, which may underrepresent recovery behavior and alternative routes.

Prediction uncertainty and planner feasibility. The deterministic fixed-horizon objective does not represent multiple plausible unseen layouts or calibrated uncertainty. Because GLAM conditions on memory, goal, and pose rather than a specified future action sequence, its outputs are goal-conditioned predictions rather than unrestricted action-conditioned simulations. Low latent error also does not guarantee target visibility, reachability, or collision-free execution; the High-Level Planning Module and Action Expert must therefore retain observation-based verification and geometric control.

Memory fidelity and generalization. Visual, depth, and pose errors can corrupt retrieval and predictive context, while the bounded token representation trades spatial coverage against resolution and inference cost. The real-world results remain qualitative, and we do not yet establish robustness to substantial localization drift, persistent scene changes, unseen environments, or larger-scale deployments.

7 CONCLUSION

We presented GLAM, a predictive model for future map tokens and goal-conditioned waypoint latents, and GLAM NAV, its integration with the High-Level Planning Module, spatiotemporal memory, and Action Expert. Predictions guide exploration, while real observations govern memory updates and STOP. On the full HM3D-Semantics v0.2 ObjectNav benchmark, GLAM NAV achieves a state-of-the-art 82.30% SR with 34.00% SPL. Ablations associate waypoint guidance with broader coverage and timely retrieval and reachable-viewpoint grounding with improved high-level planning. Path efficiency, uncertainty, and generalization to planner-induced trajectories remain open challenges.

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.

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.

Yihan Cao, Jiazhao Zhang, Zhinan Yu, Shuzhen Liu, Zheng Qin, Qin Zou, Bo Du, and Kai Xu. CogNav: Cognitive process modeling for object goal navigation with LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9550–9560, 2025. doi: 10.1109/ICCV51701.2025.00891.

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.

Yuxuan Kuang, Hai Lin, and Meng Jiang. OpenFMNav: Towards open-set zero-shot object navigation via vision-language foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 338–351. Association for Computational Linguistics, June 2024. doi: 10.18653/v1/2024.findings-naacl.24.

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

Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. WMNav: Integrating vision-language models into world models for object goal navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025. doi: 10.1109/IROS60139.2025.11246684.

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.

Dekang Qi, Shuang Zeng, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, and Mu Xu. MerNav: A highly generalizable memory–execute–review framework for zero-shot object goal navigation. *arXiv preprint arXiv:2602.05467*, 2026. doi: 10.48550/arXiv.2602.05467.

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.

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.

Ruijie Sang, Yiqun Duan, Pinhan Fu, Ruilin Wang, Wei Sui, and Xianda Guo. SkillNav: Score-level skill intervention for zero-shot object goal navigation. *arXiv preprint arXiv:2607.15758*, 2026. doi: 10.48550/arXiv.2607.15758.

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, 2023. doi: 10.1109/CVPR52729.2023.00477.

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 *Advances in Neural Information Processing Systems*, volume 37, pp. 5285–5307, 2024.

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.

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.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2412.06224.

Jiazhao Zhang, Gengze Zhou, Hale Yin, Yiyang Huang, Zixing Lei, Qihang Peng, Haoqi Yuan, Jie Zhang, Xudong Guo, Xiaoyue Chen, An Yang, Fei Huang, Zhibo Yang, Junyang Lin, Dayiheng Liu, Jingren Zhou, Zhuoyuan Yu, Jingyang Fan, Zhixuan Liang, Pei Lin, Ye Wang, Haoyang Li, Anzhe Chen, Kun Yan, Xiao Xu, Jiahao Li, Lulu Hu, Minying Zhang, Shurui Li, Wenhu Xiao, Shuai Bai, Xuancheng Ren, Chenxu Lv, Chenfei Wu, and Xiong-Hui Chen. Qwen-RobotNav technical report: A scalable navigation model designed for an agentic navigation system. *arXiv preprint arXiv:2606.18112*, 2026. doi: 10.48550/arXiv.2606.18112.

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.

A Data Collection Details

Section 4.2 explains how collected transitions supervise the two prediction branches of GLAM. This appendix documents the underlying oracle-guided collection procedure in Habitat using HM3D v0.2 scene assets.

A.1 Task Construction and Oracle Trajectories

Tasks are allocated by scene and object category. Within each category, a target instance is sampled and matched to an original ObjectNav episode from which it is reachable. The start position, orientation, target instance, and route are then fixed and reused during recollection. Habitat's pathfinder computes a shortest path from the start to the nearest reachable valid viewpoint associated with the selected object, rather than to the object center. Observations are rendered by directly setting the agent pose at interpolated path positions spaced by at most approximately $0.5\,\mathrm{m}$. The resulting data are therefore oracle-guided observation trajectories rather than closed-loop action rollouts.

A.2 Observations and Visual Features

The RGB-D camera has a resolution of 320×320, a 90° horizontal field of view, and a height of $1\,\mathrm{m}$. Four views at 90° yaw intervals are acquired at the start and at selected frontier nodes; the first starting view preserves the original episode orientation. Intermediate positions receive a forward-facing observation. Before leaving each nonterminal node, the collector acquires an additional view aligned with the next node. At the endpoint, a forward-facing observation is supplemented by a view directed toward the annotated target center. This target-directed view is stored separately and is not fused into the map.

Each RGB image is processed by the Qwen3-VL-4B visual encoder and its final spatial merger to obtain

$$\boldsymbol{V} \in \mathbb{R}^{10\times10\times2560}, \tag{6}$$

which is stored in half precision. The encoder is used only to extract visual features and does not generate text during collection.

A.3 Spatial-Memory Construction

Valid depth pixels are back-projected using the camera intrinsics and pose to update a geometric voxel map with a horizontal resolution of $5\,\mathrm{cm}$, representing free, occupied, and unknown space. Visual tokens are associated with image pixels and their depth-projected surface points to construct a separate semantic voxel map,

$$\mathcal{M} = \{(\boldsymbol{k}_i, \boldsymbol{f}_i)\}_{i=1}^{N}, \qquad \boldsymbol{k}_i \in \mathbb{Z}^3, \quad \boldsymbol{f}_i \in \mathbb{R}^{2560}. \tag{7}$$

Here, $\boldsymbol{k}_i$ identifies a semantic voxel with a horizontal resolution of $10\,\mathrm{cm}$ and one of four height bands, whose boundaries are 0.5, 1.0, and $1.6\,\mathrm{m}$ relative to the starting position. The semantic voxels thus have nonuniform vertical extents rather than an isotropic cubic shape. A visual token may contribute to multiple voxels, and each voxel may receive observations from multiple frames. Within-frame representative-feature selection and cross-frame updates follow the collector's `max_simi` and `replace_by_vis` rules, respectively. The additional departure-aligned view is fused into semantic memory before a view-specific voxel snapshot is saved, without changing the geometry used for frontier selection.

A.4 Frontier Selection and Transition Labels

Frontiers separate observed free space from unexplored regions. The collector selects frontier nodes along the oracle path within its known-free prefix, before the path first enters an unknown or occupied location. Among candidates satisfying reachability, obstacle-clearance, and known-free line-of-sight constraints, it selects the point providing the greatest forward progress, subject to a minimum advancement of $0.3\,\mathrm{m}$. If no candidate is valid, the collector advances by at most $0.5\,\mathrm{m}$ along the path to acquire more observations. This perception step updates the map but does not itself create a frontier-transition label. A final segment terminates at the goal viewpoint.

For each labeled segment s, the collector retains the map at frontier selection, $\mathcal{M}_s^{\mathrm{sel}}$; the departure map after incorporating the additional aligned view, $\mathcal{M}_s^{\mathrm{dep}}$; and the arrival map after completing the observations, $\mathcal{M}_s^{\mathrm{arr}}$. The supervision is

$$\begin{aligned} \boldsymbol{d}_s &= \boldsymbol{x}_s^{\mathrm{arr}} - \boldsymbol{x}_s^{\mathrm{dep}}, \\ \Delta\mathcal{M}_s &= \mathrm{Diff}\left(\mathcal{M}_s^{\mathrm{dep}}, \mathcal{M}_s^{\mathrm{arr}}\right). \end{aligned} \tag{8}$$

Here, $\boldsymbol{x}$ denotes a position in metric world coordinates and $\boldsymbol{d}_s$ is the displacement to the selected frontier. The final goal segment uses the same departure-to-arrival convention, with the goal viewpoint as its destination. These world-coordinate displacements are distinct from the robot-centric planar pose increments used by the waypoint encoder. The map difference stores added voxel keys and features, updated features and point information at existing keys, and deleted keys; it is not a numerical subtraction of embeddings.

Each trajectory stores RGB-D observations, visual features, camera parameters, poses, map-update events, departure-view snapshots, segment labels, the final accumulated voxel map, file hashes, and replay-validation metadata. Although observations span the route, the sparse labeled segments need not cover it continuously. Reaching an oracle goal viewpoint does not by itself establish target visibility or provide a successful ObjectNav STOP label.

A.5 Dataset Statistics

The completed collection contains 11,000 trajectories across 145 HM3D scenes. The training split comprises 10,000 trajectories, 440,265 map-observation events, and 24,963 supervised segments, including 14,963 frontier segments and 10,000 goal segments. The internal test split contains 1,000 trajectories, 43,866 map-observation events, and 2,465 supervised segments, including 1,465 frontier segments and 1,000 goal segments. Both splits use the same 145 source scenes, so the internal test split does not measure generalization to unseen scenes. Their cumulative oracle-path lengths are $98.53\,\mathrm{km}$ and $9.82\,\mathrm{km}$, respectively.

In total, the dataset contains 484,131 map-observation events and 27,428 supervised segments. Map-observation events include in-place four-direction observations and additional departure-aligned views and should not be interpreted as executed primitive-action steps. Including the terminal target-directed views stored without map fusion, the collection contains 495,131 RGB-D frames. At a reference rate of $1\,\mathrm{Hz}$, the map-observation counts correspond to approximately 122.3 hours for training and 12.2 hours for internal testing, or 134.5 hours in total. This normalization does not measure navigation time or collection runtime. The segment counts refer to saved supervision before training-time filtering, temporal windowing, or data augmentation; restricting supervision to frontier segments yields 16,428 segments across both splits.

B LLM Use Declaration

Large Language Models (ChatGPT) were used exclusively to improve the clarity and fluency of English writing. They were not involved in research ideation, experimental design, data analysis, or interpretation. The authors take full responsibility for all content.