

BEYOND THE REMEMBERED WORLD: PREDICTIVE 4D BELIEF FOR PERSISTENT NAVIGATION IN EVOLVING WORLDS

Anonymous authors

Paper under double-blind review

ABSTRACT

Recent embodied navigation systems increasingly use persistent spatial memories and vision-language models to act across repeated visits to familiar environments. Persistent maps make prior observations available for subsequent decisions and reduce the need to reconstruct familiar scenes for every task. However, remembered observations need not describe the current world: objects may move outside the agent’s view, even while it travels toward a predicted destination. We study navigation under hidden world evolution, where an agent must infer the latent current state from irregular histories and revise its decisions as time and evidence accumulate. We introduce **EVOLVINGNAV**, an embodied agent that turns time-stamped 3D entity histories into a structured belief over whether the last observed state persists or the target has relocated. A frozen, zero-shot vision-language controller acts on this belief, while an event-driven filter forecasts states at candidate arrival times, incorporates newly informative, visibility-aware RGB-D evidence, and replans as the world evolves. We also introduce **EVOLVINGWORLD-BENCH**, a human-trace-grounded benchmark that evaluates state prediction and embodied decisions under controlled changes before and during navigation. Across simulation benchmarks and physical robot trials, **EVOLVINGNAV** improves navigation success and search efficiency relative to evaluated baselines. Paired controls show that temporal prediction is most useful when changes exhibit learnable regularity; ablations demonstrate the value of preserving uncertainty and updating beliefs from visual evidence.

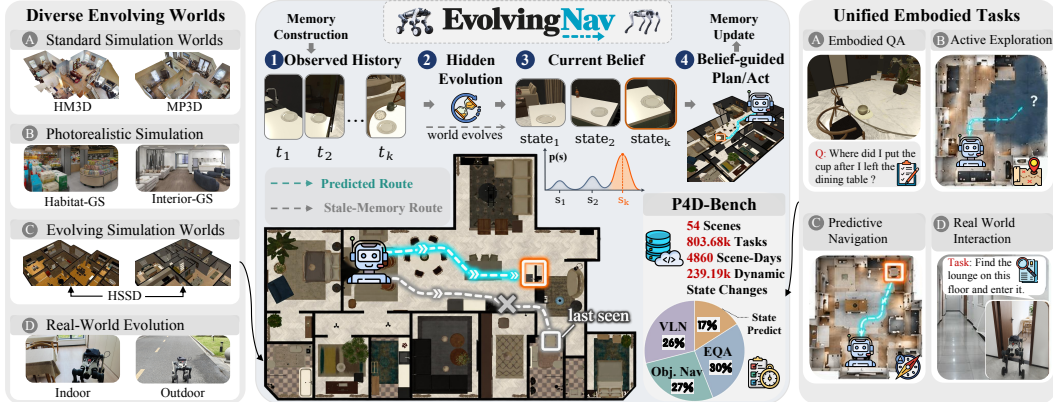


Figure 1: Overview of EVOLVINGNAV for predictive embodied reasoning and action in evolving worlds.

1 INTRODUCTION

Long-lived embodied agents navigate familiar environments across repeated tasks. Spatial maps and object memories preserve observations from earlier visits (Gu et al., 2024; Liu et al., 2024a). However, those observations may not describe the world when the agent acts. Human activity changes task-relevant states outside the robot’s view: a medication box observed on a bedside table may return to a cabinet before the next request. Further changes can occur while the robot travels, invalidating a

prediction made at the start of navigation. The agent must therefore plan from observations acquired minutes, hours, or days earlier and estimate what will be true when it inspects a target.

Partial observability makes the distinction between historical observations and current state consequential. Observations record past states, whereas the latent world state may evolve between them. This produces temporally stale memory: even accurate retrieval from a geometrically correct map can direct the agent to an outdated location. Searching from scratch discards useful history. A persistent agent instead needs a time-indexed predictive belief over the latent state. It uses past observations as evidence, represents uncertainty about unobserved changes, and supports revision during interaction. The belief must remain valid at the time of action, which may differ from the query time.

Memory systems retain object-centric observations and spatial relations, while predictive representations use temporal histories and recurring activities to estimate unobserved states (Patel et al., 2023; Saavedra-Ruiz et al., 2026). PredictiveGraphs, in particular, exposes Top- k navigation candidates and incorporates observations collected during active navigation (Saavedra-Ruiz et al., 2026). During execution, the world can continue to evolve. Candidate destinations require different travel times, so the agent must compare their probabilities at their respective arrival times. Actions can reveal evidence, while the time spent acting permits further hidden changes. This setting calls for a closed-loop mechanism that propagates, tests, and revises multiple hypotheses as actions and world evolution unfold together.

A reliable loop must retain timestamped histories, distinguish persistence at the last observed state from relocation, and preserve probability beyond known candidates. It must also condition negative evidence on visibility: a clear view of an empty location weakens the corresponding hypothesis, whereas an occluded view offers little evidence. A previously inspected location can become plausible again after further hidden evolution. Navigation therefore requires the agent to select both a destination and an informative inspection under finite travel and time budgets.

We formalize this problem as **Evolving-World Navigation**: locating and visually verifying task-relevant entities from intermittent histories while their latent states may change outside observation, both before a query and during execution. Portable-object navigation provides a measurable instantiation because relocations are consequential, externally caused, and often entirely unobserved, while the underlying distinction between historical observation and current belief applies more broadly to changing scene state. Given a target, a causal observation history, candidate states, and an action budget, the agent must infer where the target will be when inspected, gather physical evidence, and revise its plan without access to hidden transition events or current ground truth. To isolate complementary capabilities, we evaluate fixed-target episodes that test query-time inference, belief-guided search, and evidence-based recovery, and continuing-evolution episodes that test arrival-time prediction and replanning. Both expose only observations acquired along the executed trajectory.

We propose **EVOLVINGNAV** (Figure 1), a predictive 4D belief agent. A frozen, zero-shot vision-language controller invokes memory, prediction, inspection, exploration, and navigation tools. Within the agent, P4D-NAV encodes irregular timestamped histories and factorizes belief into persistence at the last observed state, relocation among alternative states, and residual mass outside the known set. This factorization preserves distinct persistence and relocation hypotheses without collapsing uncertainty to one destination. The belief draws on a persistent 4D memory that retains entity identity, spatial context, observation time, and evidence provenance.

An event-driven predict-observe-replan filter couples action and inference. Before selecting a destination, the agent propagates its current posterior to each candidate’s estimated arrival time and balances occupancy probability, expected new coverage, and travel cost. It then executes a short action chunk and uses opportunistic or deliberate RGB-D evidence to update the current-time belief. Clear negative evidence can redirect the route before the planned destination is fully inspected; inconclusive views preserve probability for later testing. Time-valid evidence rounds limit repeated counting of correlated views, while dynamic propagation can reopen a previously inspected state. Prediction therefore supplies a prior for action, and embodied interaction continually checks and corrects it.

To evaluate stale histories, hidden transitions, and the physical cost of testing uncertain states jointly, we introduce **EVOLVINGWORLD-BENCH**, a human-trace-grounded simulation benchmark with 54 evolving scenes and 803.68k tasks spanning state prediction, object navigation, vision-and-language navigation, and embodied question answering. It links the quality of a time-indexed belief to first inspection, recovery, path efficiency, and online replanning. Paired static, routine, and random worlds

hold scene and query factors fixed while varying temporal structure, testing whether gains arise from learnable historical regularity rather than location frequency, geometry, or additional exploration.

Experiments support this interpretation. Across the full navigation suite, **EVOLVINGNAV** reaches 61.32% First-Inspection SR and 86.18% Search SR, exceeding DynaMem by 15.99 and 14.24 points (Table 1). On episodes with execution-time target motion, it improves Dynamic SR by 18.4 points over PredictiveGraphs (Table 2). Component ablations show that the gain depends on preserving alternative hypotheses and revising them from online evidence, rather than only selecting a stronger first destination. Paired controls show the clearest advantage under learnable routines rather than matched random motion, while cross-benchmark, held-out-scene, frozen-VLM, and physical-robot evaluations test transfer beyond one simulator and controller. These results suggest that persistent embodied memory should support calibrated inference about the present and evidence-seeking action in addition to faithful historical storage.

Our contributions are fourfold:

- **Evolving-World Navigation:** We formulate navigation under hidden world evolution, where states may change outside the agent’s view and during execution.
- **Predictive 4D Belief Agent:** We introduce **EVOLVINGNAV**, coupling structured state prediction with an event-driven predict–observe–replan filter.
- **EvolvingWorld-Bench:** We establish a large-scale benchmark with 54 evolving scenes and 803.68k tasks for controlled evaluation of predictive navigation.
- **Comprehensive Validation:** Extensive simulation and real-robot experiments demonstrate consistent gains across environments, benchmarks, and frozen VLMs.

2 RELATED WORK

Navigation, memory, and changing worlds. Embodied navigation combines perception, spatial reasoning, and action; open-vocabulary maps and scene graphs ground language in structured space (Anderson et al., 2018; Savva et al., 2019; Chaplot et al., 2020; Xin et al., 2026; Huang et al., 2023; Gu et al., 2024; Werby et al., 2024; Li et al., 2026a; Huang et al., 2026). Foundation-model agents combine maps, tools, and navigation skills (Yokoyama et al., 2024; Long et al., 2024; Ziliotto et al., 2025; Li et al., 2026b; Zheng et al., 2022; Ginting et al., 2024; Zhang et al., 2025b), while lifelong benchmarks evaluate navigation and question answering across tasks (Khanna et al., 2024b; Yadav et al., 2025) and episodic question answering (Majumdar et al., 2024). Dynamic maps and 4D memories support updates and spatiotemporal retrieval (Sohn et al., 2025; Gorlo et al., 2026; Bescos et al., 2018; Narayana et al., 2020; Liu et al., 2024a).

Predictive memory and human routines. Predictive navigation infers locations from partial histories or continuous placements (Kurenkov et al., 2023; Saavedra-Ruiz et al., 2026; Argenziano et al., 2026). PredictiveGraphs already combines future-state prediction, Top- k navigation, online observations, and agent tools; our complementary focus is an arrival-time filtering loop for continued hidden evolution, with time-valid evidence and candidate reopening. HOMER/HOMER+, STREAK, and personalized navigation model household routines or context drift (Patel & Chernova, 2023; Patel et al., 2023; Bartoli et al., 2025; Wang et al., 2026). HD-EPIC and ParaHome provide human–object interaction traces (Perrett et al., 2025; Kim et al., 2025). EVOLVINGWORLD-BENCH uses human traces and separately tracked simulated evidence to constrain executable histories; paired worlds test temporal regularity beyond location frequency.

3 PREDICTIVE 4D BELIEF NAVIGATION

EVOLVINGNAV turns timestamped 3D histories into a current-state belief and revises it with RGB-D evidence; P4D-NAV separates persistence from relocation inside the loop in Figure 2.

3.1 PROBLEM FORMULATION

An environment contains navigable space $\mathcal{X}$, entities $\mathcal{O}$, and candidate states $\mathcal{C}_o$ with inspection viewpoints. Before query time t_q , the agent has only the causal history

$$\mathcal{H}_{\leq t_q} = \{(I_k, D_k, \xi_k, t_k) \mid t_k \leq t_q\}, \quad (1)$$

where I_k, D_k, ξ_k are RGB, depth, and pose. The hidden state lies in $\tilde{\mathcal{C}}_o = \mathcal{C}_o \cup \{\text{unknown}\}$; later out-of-view transitions are also hidden.

Given a target, initial pose x_0 , memory, and action budget, the agent observes z_j after acting and must stop at a valid target viewpoint. It optimizes

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\pi} [\mathbb{I}(\text{success}) - \lambda_d L - \lambda_n N_{\text{inspect}}], \quad (2)$$

where L and N_{inspect} are travel and inspection costs.

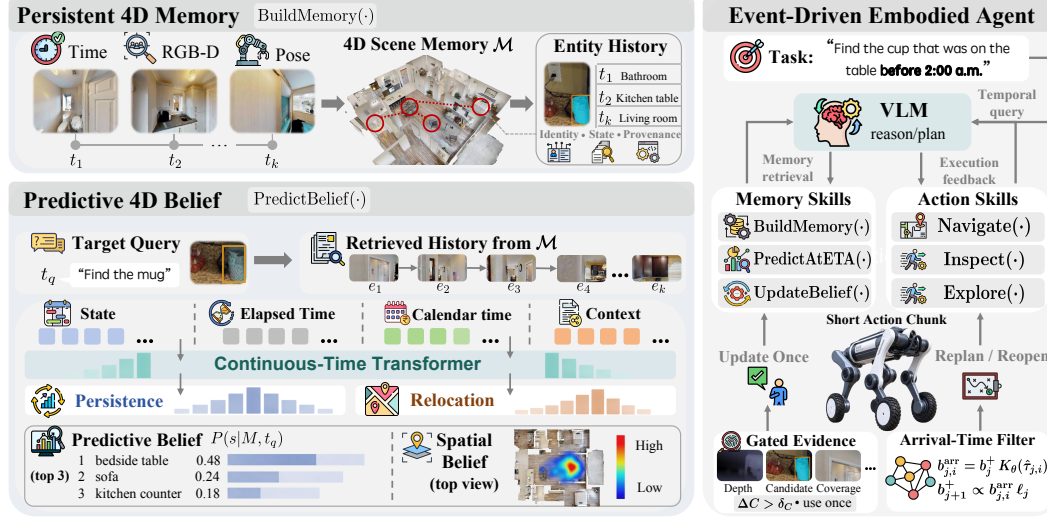


Figure 2: EVOLVINGNAV couples persistent 4D memory, predictive belief, and an embodied agent that updates and replans from stepwise evidence.

3.2 CAUSAL 4D MEMORY AND HISTORY ENCODING

The 4D memory registers RGB-D in a shared frame, associates entities using semantics, geometry, and time, and materializes versioned deltas causally:

$$\mathcal{M}_{\leq t} = \text{Materialize}(\Delta \mathcal{M}_0, \dots, \Delta \mathcal{M}_t) = (\mathcal{G}_t, \mathcal{V}_t, \mathcal{B}_t), \quad (3)$$

Here $\mathcal{G}_t, \mathcal{V}_t, \mathcal{B}_t$ store relations, time-valid versions, and provenance. Changes append versions without erasing history; perception and versioning details are in Appendix A.2 (Liu et al., 2024b; Ravi et al., 2025; Radford et al., 2021).

Target-conditioned history and continuous time. For target o , a typed query returns the most recent K causal events

$$H_o = \{e_k = (o_k, s_k, t_k, z_k, c_k, v_k)\}_{k=1}^K, \quad (4)$$

where o_k, s_k, t_k, z_k encode entity, candidate state, time, and $\{\text{seen}, \text{notseen}\}$ evidence; c_k, v_k encode confidence and visibility. Related context events are causal, candidate encodings capture function and geometry, and `unknown` absorbs unmapped states.

Each event token combines semantic, evidence, and temporal features:

$$\begin{aligned} \mathbf{x}_k &= E_o(o_k) + E_s(s_k) + E_z(z_k) + E_c(c_k) + E_v(v_k) \\ &\quad + \phi_{\Delta}(t_q - t_k) + \phi_{\text{cal}}(t_k), \end{aligned} \quad (5)$$

where ϕ_{Δ} encodes log elapsed time and ϕ_{cal} encodes hour and weekday. A query token specifies target, time, and last state; a compact Transformer yields

$$(\mathbf{h}_1, \dots, \mathbf{h}_K, \mathbf{h}_q) = \text{CTTransformer}(\mathbf{x}_1, \dots, \mathbf{x}_K, \mathbf{x}_q). \quad (6)$$

This represents irregular gaps, evidence, and event order without hidden labels.

3.3 PERSISTENCE-RELOCATION BELIEF

We predict whether the last observation persists and, otherwise, where the target moved. Let s_{last} be the last positive state, guaranteed by episode validity (Appendix B.4). The persistence branch predicts

$$\rho_q = P(s_{t_q}^o = s_{\text{last}} \mid H_o, t_q) = \sigma(\mathbf{w}_{\rho}^{\top} \mathbf{h}_q). \quad (7)$$

For each alternative $i \in \tilde{\mathcal{C}}_o \setminus \{s_{\text{last}}\}$, a shared pointer head computes

$$a_i = \frac{(W_q \mathbf{h}_q)^\top (W_c \mathbf{g}_i)}{\sqrt{d}} + \psi(\mathbf{h}_q, \mathbf{g}_i), \quad \pi_q(i) = \frac{\exp a_i}{\sum_{u \in \tilde{\mathcal{C}}_o \setminus \{s_{\text{last}}\}} \exp a_u}, \quad (8)$$

where $\mathbf{g}_i$ represents candidate i , d is the projection dimension, and ψ measures learned compatibility. The prior is

$$b_q^-(i) = \begin{cases} \rho_q, & i = s_{\text{last}}, \\ (1 - \rho_q)\pi_q(i), & i \neq s_{\text{last}}. \end{cases} \quad (9)$$

We train with $\mathcal{L}_{\text{state}} = -\log b_q^-(s_{t_q}^o)$ without change labels; the normalized belief represents returns and alternatives.

3.4 PREDICTIVE EMBODIED AGENT

A frozen VLM invokes memory, prediction, navigation, inspection, and exploration tools. After a short action chunk, new coverage, an ETA change, evidence, or arrival, the encoder produces $\mathbf{h}_j$ and the row-normalized operator

$$[K_\theta^{(j)}(\Delta t)]_{r,u} = \text{softmax}_u f_\theta(\mathbf{h}_j, \mathbf{g}_r, \mathbf{g}_u, \phi_\Delta(\Delta t)). \quad (10)$$

It shares the encoders in Equations (6) and (8); a separate head uses row-wise cross-entropy on chronological state pairs (Appendix A.4). From $b_0^- = b_q^-$, a completed chunk predicts the next current-time prior:

$$b_{j+1}^- = b_j^+ K_\theta^{(j)}(t_{j+1} - t_j). \quad (11)$$

Thus $b_j^\pm$ is current-time belief, distinct from arrival-time forecasts.

Arrival-aware action selection. For candidate viewpoints $\mathcal{W}_i$, let $d_{j,i} = d_{\text{geo}}(x_j, \mathcal{W}_i)$ and $\hat{\tau}_{j,i}$ be geodesic distance and estimated arrival-plus-inspection time. The agent selects

$$i_j^* = \arg \max_{i \in \mathcal{C}_o} \frac{p_{j,i}^{\text{arr}} \hat{q}_{j,i}^{\text{new}}}{d_{j,i} + \lambda_t \hat{\tau}_{j,i} + \lambda_{\text{inspect}}}, \quad (12)$$

where $p_{j,i}^{\text{arr}}$ is defined below and $\hat{q}_{j,i}^{\text{new}}$ is detection probability from unseen coverage. Higher unknown-state utility triggers candidate-expanding exploration.

Visibility-qualified measurement update. A posed RGB-D view can update every covered candidate. The agent estimates $\hat{r}_{j,i} = P(\text{detect} \mid s_{t_j} = i, \mathbf{f}_{j,i})$ from projected candidate geometry, online depth, and calibrated view features, without target masks, poses, or simulator visibility flags. Given no detection,

$$\ell_j(i) = P(y_j = \emptyset \mid s_{t_j} = i) = \begin{cases} 1 - \hat{r}_{j,i}, & i \in \mathcal{C}_o, \\ 1, & i = \text{unknown}, \end{cases} \quad (13)$$

and the measurement posterior is

$$b_j^+(i) = \frac{b_j^-(i)\ell_j(i)}{\sum_{u \in \tilde{\mathcal{C}}_o} b_j^-(u)\ell_j(u)}. \quad (14)$$

An update requires over $\delta_C = 0.05$ new surface coverage and uses only that increment, avoiding repeated use of overlapping frames. Dynamic reopening starts a new time-valid round, so the same viewpoint can test a changed state; a verified match stops.

Posterior propagation and replanning. For action selection, the current posterior is forecast separately to each candidate’s ETA rather than replacing the current-time filter:

$$b_{j,i}^{\text{arr}} = b_j^+ K_\theta^{(j)}(\hat{\tau}_{j,i}), \quad p_{j,i}^{\text{arr}} = b_{j,i}^{\text{arr}}(i). \quad (15)$$

Fixed-target evaluation uses identity dynamics; continuing evolution applies Equations (11) and (15). Each epoch propagates, updates once, forecasts arrivals, and replans; pseudocode and evidence de-duplication are in Appendix A.4.

4 EVOLVINGWORLD-BENCH: EVALUATING NAVIGATION IN EVOLVING WORLDS

EVOLVINGWORLD-BENCH evaluates decisions from time-ordered observation histories while the world changes outside observation. Its 54 scenes and 803.68k task instances combine persistent histories, executable placements, and controlled dynamics.

4.1 TASKS AND CONTROLLED DYNAMICS

At the task-modality level shown in Figure 1, the pool comprises current-state prediction (17%), object navigation (27%), vision-and-language navigation (VLN; 26%), and embodied question answering (EQA; 30%) (Majumdar et al., 2024; Sakamoto et al., 2024). The complementary protocol-export accounting in Figure 3 is N1 (13.51%), N2 (13.37%), N3 (13.27%), N4 (26.52%), N5 (3.34%), and EQA (29.99%). In this second view, VLN-conditioned episodes are assigned to their corresponding navigation protocol rather than shown as a separate wedge. Prediction evaluates a calibrated region/receptacle belief independently of control; navigation uses the same target-time queries (Figure 3 and Table 7). N1 inspects the Top-1 destination; N2 uses the full belief for cost-aware search; N3 adds visibility-qualified updates and recovery. N1–N3 fix the target after the query. N4 advances the world clock during execution, permitting further hidden transitions. N5 applies these protocols to held-out scenes and households.

VLN grounds language goals in the same histories; EQA probes current and historical states without future evidence. Our experiments emphasize prediction and navigation, with EQA as a memory diagnostic. Controlled comparisons match public observations, candidates, geometry, starts, controllers, and budgets.

4.2 WORLD CONSTRUCTION AND PAST-ONLY EVALUATION

Human traces constrain generation; source trajectories are not copied (Figure 4). CASAS supplies longitudinal occupancy and timing statistics; ARAS and OPPORTUNITY add activity-context coverage; and HD-EPIC and ParaHome provide object-action and motion evidence (Cook et al., 2013; Perrett et al., 2025; Kim et al., 2025). HOMER+ is tracked separately as simulated long-horizon routine evidence (Patel et al., 2023). Event records preserve source identifiers and distinguish measured, environment-derived, simulated, and benchmark-authored quantities (Appendix B.1).

Across 54 HSSD scenes (Khanna et al., 2024a), household habits and stochastic exceptions generate chronological histories. Stable, routine, personal, and irregular regimes test persistence, activity regularity, owner-specific habits, and uncertainty (Table 6); labels are hidden. Habitat replay ensures semantically valid, collision-free placements and reachable inspection viewpoints.

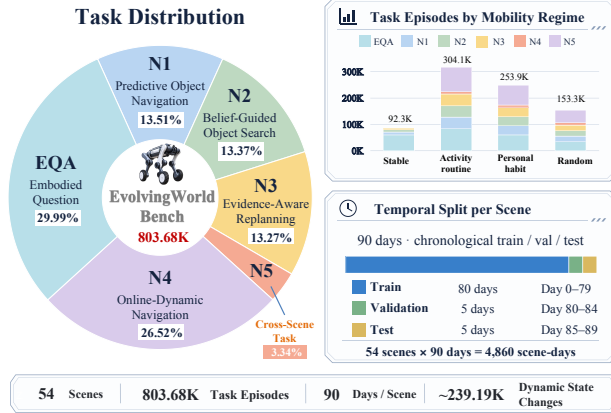


Figure 3: Overview of EVOLVINGWORLD-BENCH evaluation protocols.

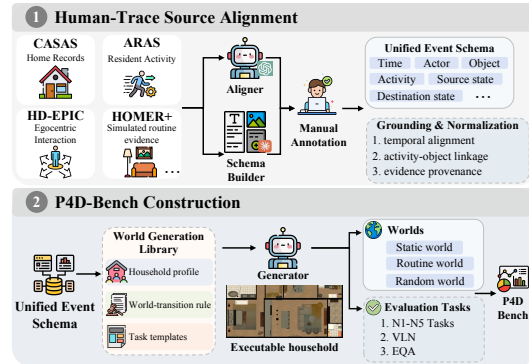


Figure 4: Construction of EVOLVINGWORLD-BENCH.

Patrols collect visibility-qualified RGB-D using only maps, time, travel cost, and observed history. Queries expose only observations available by the query time, the last supported state, elapsed/calendar time, candidates, and spatial context; current states, transitions, activities, mobility labels, and oracle viewpoints remain private. Splits group complete temporal sequences, paired worlds, and near-duplicate query/transition groups. Audits cover temporal/metadata leakage, split overlap, paired-world consistency, physical validity, and patrol access to hidden state. Stratification covers mobility, transition, staleness, spatial difficulty, and generalization.

5 EXPERIMENTS

We test whether predictive 4D memory improves current-state inference, embodied search, and recovery from contradicted beliefs. Controlled EVOLVINGWORLD-BENCH and robot comparisons share public histories, legal candidates, inspection viewpoints, low-level control, and action budgets. Cross-environment evaluations instead retain each released agent’s native policy and are used as system-level generalization evidence, not as a component-level causal comparison. Further protocol, implementation, and analysis details are summarized in Appendices A–D.

5.1 SETUP

Datasets. Our primary evaluation uses the complete EVOLVINGWORLD-BENCH N1–N5 suite for First-Inspection navigation, belief-guided search, evidence-aware replanning, online dynamics, and cross-scene transfer. FindingDory (Yadav et al., 2025) and GOAT-Bench (Khanna et al., 2024b) test long-horizon memory and repeated-goal navigation. We additionally evaluate one native task in each of MP3D, HM3D, Habitat-GS, and InteriorGS, and run 64 matched LYNX M20 trials per method.

Real-world setup. We evaluate on the DEEP Robotics LYNX M20 as the primary quantitative platform over 64 matched trials, with X30 and Lite3 evaluated on matched 32-block transfer subsets. Platform-specific perception and low-level navigation are wrapped by a common interface, while all methods receive identical pre-query histories, candidate locations, start conditions, and robot-specific control stacks. Each 360-second episode requires online confirmation within three inspections.

Controllers and baselines. Our main method uses a frozen GPT-5.6-Luna controller in a zero-shot setting: the VLM receives no navigation-task fine-tuning and invokes the fixed memory, prediction, inspection, exploration, and navigation tools. The prompt, tool interface, and inference budget remain fixed within each comparison. Baselines span navigation, structured memory, and predictive world models; paired runs with GPT-4o, GPT-5.5, GPT-5.6-Luna, and Qwen2.5-VL-3B/32B test controller dependence.

Metrics. Prediction uses Top-1, MRR, NLL, ECE, and true-state rank. Navigation uses SR, SPL, First-Inspection SR, Recovery SR, inspection count, distance, and time. Main comparisons report five-seed variation; paired static, routine, and random worlds isolate learnable temporal structure.

5.2 MAIN RESULTS

Table 1: Cross-benchmark comparison (%). $\pm$ denotes standard deviation over five seeds.

Method	FindingDory		GOAT-Bench			EVOLVINGWORLD-BENCH		
	HL-SR $\uparrow$	HL-SPL $\uparrow$	SR $\uparrow$	SPL $\uparrow$	Repeat-SR $\uparrow$	First-Inspection SR $\uparrow$	Search SR $\uparrow$	SPL $\uparrow$
<i>Open-vocabulary and lifelong navigation</i>								
VLFM (Yokoyama et al., 2024)	13.42	10.86	15.79	5.45	11.82	20.65	28.36	24.67
ZSON (Majumdar et al., 2022)	27.89	21.81	29.64	2.45	16.37	26.84	32.13	28.55
FindingDory Agent (Yadav et al., 2025)	52.44 ^{PR}	–	–	–	–	–	–	–
LagMemo GLUE (Zhou et al., 2025)	32.79	23.25	20.61	1.74	11.27	30.18	42.27	33.82
<i>Structured spatio-temporal memory</i>								
HOV-SG (Werby et al., 2024)	21.47	15.20	28.37	2.23	12.38	36.46	52.32	40.61
DynaMem (Liu et al., 2024a)	30.30	22.78	14.54	1.71	15.79	45.33	71.94	56.83
<i>Predictive world and state modeling</i>								
SLaTe-PRO (Patel et al., 2023)	18.18	14.53	9.41	0.99	6.38	13.00	36.33	26.37
SGM+NEP (Kurenkov et al., 2023)	34.20	19.48	10.67	1.64	2.56	27.26	41.82	37.00
FlowMaps (Argenziano et al., 2026)	13.03	10.24	25.07	1.32	7.25	14.33	32.67	17.35
PredictiveGraphs (Saavedra-Ruiz et al., 2026)	24.24	13.50	10.27	1.80	7.48	36.18	43.67	40.36
EVOLVINGNAV (ours)	53.22 ± 3.87	38.83 ± 2.43	35.43 ± 2.16	12.98 ± 0.92	19.31 ± 0.68	61.32 ± 3.235	86.18 ± 2.07	70.15 ± 1.74

^{PR} Paper-reported under the cited native protocol; unmarked baseline scores are our reproductions under the method-preserving protocol in Appendix A.7.



Figure 5: Predictive, last-seen, and nearest-first search on EVOLVINGWORLD-BENCH.

Table 2: N4 online-dynamic navigation with execution-time target motion.

Method	Dynamic SR ↑	Online Recovery SR ↑	Excess Distance (m) ↓	Revisit Success ↑
SLaTe-PRO (Patel et al., 2023)	34.7	21.5	18.9	8.7
SGM+NEP (Kurenkov et al., 2023)	41.3	27.8	15.6	12.4
FlowMaps (Argenziano et al., 2026)	29.6	18.9	22.4	6.5
PredictiveGraphs (Saavedra-Ruiz et al., 2026)	46.8	36.7	13.2	18.9
EVOLVINGNAV (ours)	65.2	58.7	8.4	32.6

Comparison across embodied benchmarks. EVOLVINGNAV consistently improves the evaluated metrics across FindingDory, GOAT-Bench, and EVOLVINGWORLD-BENCH. Across EVOLVINGWORLD-BENCH N1–N5, it improves First-Inspection SR from 45.33% (DynaMem) to 61.32% and Search SR from 71.94% to 86.18%, showing that arrival-time prediction and en-route evidence improve the first fully covered inspection while retained hypotheses support recovery. The gains on FindingDory and GOAT-Bench further show that predictive memory remains effective beyond the proposed evolving-world setting.

Online navigation under execution-time dynamics. Figure 5 shows how a routine-conditioned belief avoids an obsolete last-seen location. To isolate the dynamic closed loop rather than dilute it with episodes in which motion is merely permitted, Table 2 reports the N4 subset in which the target actually changes after the query and before termination.

Relative to PredictiveGraphs, EVOLVINGNAV improves Dynamic SR by 18.4 points and Online Recovery SR by 22.0 points, while reducing excess distance by 4.8 m and increasing successful revisits by 13.7 points. These gains directly test the execution-time contribution: current-time propagation and time-valid evidence rounds let the agent recover after a target moves, while candidate reopening turns revisits into useful actions rather than permanent exclusions.

Generalization and physical deployment. Across five simulators, ours leads seven of ten metrics (Table 8). On 64 matched M20 trials, Table 3 improves both success and efficiency; details are in Appendix D.

Table 3: Primary LYNX M20 search results.

Method	First-Inspection SR ↑	Search SR ↑	Recovery SR ↑	Dist. (m) ↓	Time (s) ↓	Inspect. ↓
Last Seen + Search	17.2	26.6	12.8	56.2	284	2.19
Time Frequency	21.9	31.3	14.0	52.9	277	2.05
Retrieval + Reasoning	26.6	37.5	17.1	50.3	281	2.03
EVOLVINGNAV	34.4	48.4	24.3	43.8	248	1.84

Across these evaluations, the gains appear at three levels: cross-benchmark results improve initial destination selection, N4 isolates recovery under execution-time motion, and robot trials translate the same belief interface into shorter paths and fewer inspections. Their agreement argues against the gain coming from benchmark-specific exploration or a stronger controller alone.

5.3 ABLATION STUDIES

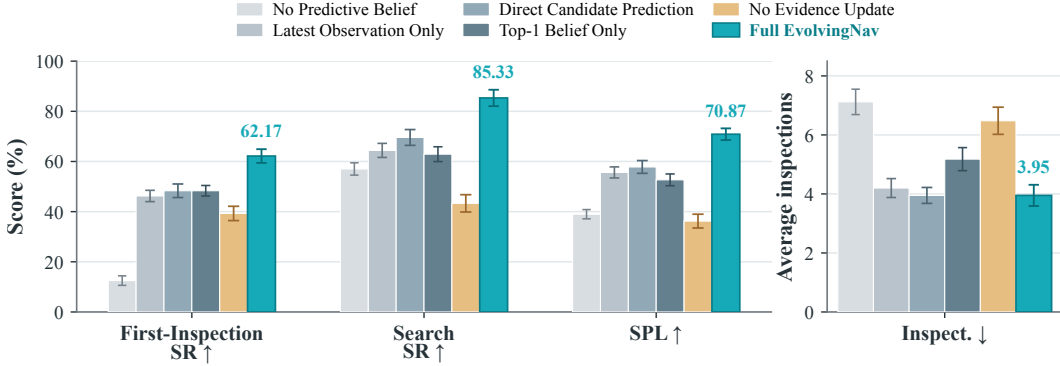


Figure 6: Core component ablation; error bars show five-seed standard deviations.

Core components. The full agent reaches 62.17 First-Inspection SR, 85.33 Search SR, and 70.87 SPL with 3.95 inspections (Figure 6). Without predictive belief, First-Inspection SR falls to 12.51; Latest Observation Only loses 20.91 Search SR points, showing that temporally persistent history guides both the initial destination and recovery. Direct candidate prediction loses 13.84/15.75 First-Inspection/Search points, supporting the persistence–relocation factorization. Top-1-only search loses 22.41 Search SR points, while removing evidence updates loses 42.02 points and raises inspections to 6.48. Thus the gain comes from retaining alternatives and revising them online, not from a single stronger guess or a larger search budget. The pattern separates the roles of the components: predictive belief chooses where inspection begins, the full distribution preserves fallback hypotheses, and evidence updating converts new views into route changes. Because No Evidence Update uses the same predictor, its gap isolates the action-time filter rather than improved offline state classification.

Prediction and controller robustness. Against the Direct Transformer, ours gains 4.08/9.47 Top-1 points and reduces NLL by 0.2367/0.3379 on temporal/LOSO splits. It also exceeds the semantic prior, supporting temporal structure beyond location frequency. The larger LOSO gain indicates transfer beyond scene-specific destination statistics, while the lower NLL provides more useful alternatives for recovery. Additional ablations are in Appendix C.

Table 4: Current-state prediction on temporal and LOSO splits.

Predictor	Temporal			LOSO		
	Top-1 ↑	MRR ↑	NLL ↓	Top-1 ↑	MRR ↑	NLL ↓
Last Seen	0.00	0.1699	3.4011	50.00	0.5850	1.8825
Time Frequency	33.85	0.5462	1.8400	49.53	0.6166	1.6455
Direct Transformer	34.16	0.5586	1.7292	51.71	0.6884	1.3157
Shared semantic prior	36.02	0.5957	1.5766	59.47	0.7462	1.0448
Ours	38.24	0.5971	1.4925	61.18	0.7523	0.9778
	±1.42	±0.0291	±0.0867	±1.53	±0.0176	±0.0228

6 CONCLUSION

We study persistent navigation under unobserved changes before and during execution. EVOLVING-NAV couples a persistence–relocation belief from 4D histories with an event-driven predict–observe–replan filter. Across EVOLVINGWORLD-BENCH, external benchmarks, and robots, it improves first-inspection selection and recovery, especially under learnable temporal structure. Future work will address interacting objects and continuously changing goals. More broadly, the results suggest that persistent embodied memory should be evaluated not only by what it stores, but by whether it supports calibrated inference and evidence-seeking action when the remembered world is no longer current.

REFERENCES

- Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.
- Francesco Argenziano, Miguel Saavedra-Ruiz, Sacha Morin, Charlie Gauthier, Daniele Nardi, and Liam Paull. Flowmaps: Modeling long-term multimodal object dynamics with flow matching. *arXiv preprint arXiv:2606.20209*, 2026.
- Ermanno Bartoli, Fethiye Irmak Doğan, and Iolanda Leite. Streak: Streaming network for continual learning of object relocations under household context drifts. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1550–1557. IEEE, 2025.
- Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. *IEEE robotics and automation letters*, 3(4):4076–4083, 2018.
- 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*, 2020.
- Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 23568–23576, 2025.
- An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.
- Diane J. Cook, Aaron S. Crandall, Brian L. Thomas, and Narayanan C. Krishnan. CASAS: A smart home in a box. *Computer*, 46(7):62–69, 2013.
- Hongyu Ding, Sizhuo Zhang, Ziming Xu, Jinwen Guo, Hongxiu Liu, Xingzhi Cheng, Zixuan Chen, Haifei Qi, Duo Wang, Hao Xu, et al. Uni-lavira: Language-vision-robot actions translation for unified embodied navigation. *arXiv preprint arXiv:2605.27582*, 2026.
- Muhammad Fadhil Ginting, Sung-Kyun Kim, David D Fan, Matteo Palieri, Mykel J Kochenderfer, and Ali-akbar Agha-Mohammadi. Seek: Semantic reasoning for object goal navigation in real world inspection tasks. *arXiv preprint arXiv:2405.09822*, 2024.
- Nicolas Gorlo, Lukas Schmid, and Luca Carlone. Describe anything anywhere at any moment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35002–35013, 2026.
- Qiao Gu, Ali Kuwajerwala, Sacha Morin, Krishna Murthy Jatavallabhula, Bipasha Sen, Aditya Agarwal, Corban Rivera, William Paul, Kirsty Ellis, Rama Chellappa, et al. Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5021–5028. IEEE, 2024.
- Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10608–10615, 2023.
- Xun Huang, Shijia Zhao, Yunxiang Wang, Xin Lu, Wanfa Zhang, Rongsheng Qu, Weixin Li, Yunhong Wang, and Chenglu Wen. Msgnav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37154–37163, 2026.
- Mukul Khanna, Yongsan Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X. Chang, and Manolis Savva. Habitat synthetic scenes dataset (HSSD-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

- Mukul Khanna, Ram Ramrakhya, Gunjan Chhablani, Sriram Yenamandra, Théophile Gervet, Matthew Chang, Zsolt Kira, Devendra Singh Chaplot, Dhruv Batra, and Roozbeh Mottaghi. GOAT-Bench: A benchmark for multi-modal lifelong navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.
- Jeonghwan Kim, Jisoo Kim, Jeonghyeon Na, and Hanbyul Joo. Parahome: Parameterizing everyday home activities towards 3d generative modeling of human-object interactions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1816–1828. IEEE, 2025.
- Andrey Kurenkov, Michael Lingelbach, Tanmay Agarwal, Emily Jin, Chengshu Li, Ruohan Zhang, Li Fei-Fei, Jiajun Wu, Silvio Savarese, and Roberto Martín-Martín. Modeling dynamic environments with scene graph memory. *arXiv preprint arXiv:2305.17537*, 2023.
- Kailing Li, Tianwen Qian, Lijin Yang, Yuqian Fu, Jingyu Gong, Xiaoling Wang, and Liang He. Bridging the 2d-3d gap: A hierarchical semantic-geometric map for vision language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15243–15252, June 2026a.
- Yijian Li, Changze Li, Hantian Shi, Jiaying Luo, Jiyuan Cai, Ming Yang, and Tong Qin. Agenticnav: Zero-shot vision-and-language navigation as a tool-calling harness. *arXiv preprint arXiv:2606.10577*, 2026b.
- Peiqi Liu, Zhanqiu Guo, Mohit Warke, Soumith Chintala, Chris Paxton, Nur Muhammad Mahi Shafullah, and Lerrel Pinto. Dynamem: Online dynamic spatio-semantic memory for open world mobile manipulation. *arXiv preprint arXiv:2411.04999*, 2024a.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024b.
- Yuxing Long, Wenzhe Cai, Hongcheng Wang, Guanqi Zhan, and Hao Dong. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882*, 2024.
- Arjun Majumdar, Gunjan Aggarwal, Bhavika Devnani, Judy Hoffman, and Dhruv Batra. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in neural information processing systems*, 35:32340–32352, 2022.
- Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16488–16498. IEEE, 2024.
- Manjunath Narayana, Andreas Kolling, Lucio Nardelli, and Phil Fong. Lifelong update of semantic maps in dynamic environments. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6164–6171. IEEE, 2020.
- Maithili Patel and Sonia Chernova. Proactive robot assistance via spatio-temporal object modeling. In *Proceedings of the 6th Conference on Robot Learning*, pp. 881–891, 2023.
- Maithili Patel, Aswin Prakash, and Sonia Chernova. Predicting routine object usage for proactive robot assistance. In *Conference on Robot Learning*, 2023.
- Toby Perrett, Ahmad Darkhalil, Saptarshi Sinha, Omar Emara, Sam Pollard, Kranti Parida, Kaiting Liu, Prajwal Gatti, Siddhant Bansal, Kevin Flanagan, Jacob Chalk, Zhifan Zhu, Rhodri Guerrier, Fahd Abdelazim, Bin Zhu, Davide Moltisanti, Michael Wray, Hazel Doughty, and Dima Damen. HD-EPIC: A highly-detailed egocentric video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmlR, 2021.

- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.
- Miguel Saavedra-Ruiz, Charlie Gauthier, Kumaraditya Gupta, Shima Shahfar, Kirsty Ellis, Steven Parkison, and Liam Paull. Predictive spatio-temporal scene graphs for semi-static scenes. *arXiv preprint arXiv:2605.00121*, 2026.
- Koya Sakamoto, Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Map-based modular approach for zero-shot embodied question answering. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10013–10019. IEEE, 2024.
- Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9338–9346. IEEE, 2019.
- Tin Stribor Sohn, Maximilian Dillitzer, Jason J. Corso, and Eric Sax. R^4 : Retrieval-augmented reasoning for vision-language models in 4d spatio-temporal space. *arXiv preprint arXiv:2512.15940*, 2025.
- Hongcheng Wang, Jinyu Zhu, and Hao Dong. User-centric object navigation: A benchmark with integrated user habits for personalized embodied object search. *arXiv preprint arXiv:2602.06459*, 2026.
- Abdelrhman Werby, Chenguang Huang, Martin Büchner, Abhinav Valada, and Wolfram Burgard. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. *Robotics: Science and Systems*, 2024.
- Zihao Xin, Wentong Li, Yixuan Jiang, Ziyuan Huang, Bin Wang, Piji Li, Jianke Zhu, Jie Qin, and Shengjun Huang. Agentvln: Towards agentic vision-and-language navigation. *arXiv preprint arXiv:2603.17670*, 2026.
- Karmesh Yadav, Yusuf Ali, Gunshi Gupta, Yarin Gal, and Zsolt Kira. Findingdory: A benchmark to evaluate memory in embodied agents. *arXiv preprint arXiv:2506.15635*, 2025.
- Naoki Yokoyama, Sehoon Ha, Dhruv Batra, Jiuguang Wang, and Bernadette Bucher. Vlfm: Vision-language frontier maps for zero-shot semantic navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 42–48. IEEE, 2024.
- Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852*, 2024.
- 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025a.
- Mingjie Zhang, Yuheng Du, Chengkai Wu, Jinni Zhou, Zhenchao Qi, Jun Ma, and Boyu Zhou. Apexnav: An adaptive exploration strategy for zero-shot object navigation with target-centric semantic fusion. *IEEE Robotics and Automation Letters*, 2025b.
- Kaiyu Zheng, Rohan Chitnis, Yoonchang Sung, George Konidaris, and Stefanie Tellex. Towards optimal correlational object search. In *IEEE International Conference on Robotics and Automation*, 2022.
- Haotian Zhou, Xiaole Wang, He Li, Zhuo Qi, Jinrun Yin, Haiyu Kong, Jianghuan Xu, and Huijing Zhao. Lagmemo: Language 3d gaussian splatting memory for multi-modal open-vocabulary multi-goal visual navigation. *arXiv preprint arXiv:2510.24118*, 2025.

Ziyu Zhu, Xilin Wang, Yixuan Li, Zhuofan Zhang, Xiaojian Ma, Yixin Chen, Baoxiong Jia, Wei Liang, Qian Yu, Zhidong Deng, et al. Move to understand a 3d scene: Bridging visual grounding and exploration for efficient and versatile embodied navigation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8120–8132. IEEE, 2025.

Filippo Ziliotto, Tommaso Campari, Luciano Serafini, and Lamberto Ballan. Tango: training-free embodied ai agents for open-world tasks. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24603–24613. IEEE, 2025.

A ADDITIONAL METHOD DETAILS

This section gives the formal episode interface, memory representation, baseline specification, and implementation details that support the method described in the main paper.

A.1 FORMAL EPISODE INTERFACE

Each navigation episode references one time-indexed prediction query. Public fields include schema version, episode and query identifiers, split, task type, scene and household identifiers, world variant, target description, query time, agent start pose, candidate-state identifiers, budgets, and success criteria. Evaluation-private fields include the current state, target pose, valid goal viewpoints, hidden transition class, oracle path, mobility label, and query-time activity. Train and validation releases may expose analysis labels in metadata; the formal test release replaces them with null values and computes grouped metrics in a private evaluator.

The high-level track exposes

NAVIGATE_TO(state), INSPECT(state), EXPLORE, STOP, NOT_FOUND,

with a fixed navigation backend. The low-level track exposes Habitat-style forward, turn, look, and stop actions. The main paper uses the high-level track to isolate predictive state selection; low-level results measure sensitivity to control and perception.

The observation history available at query time t_q is

$$\mathcal{H}_{\leq t_q} = \{(I_k, D_k, \xi_k, t_k) \mid t_k \leq t_q\}, \quad (16)$$

where I_k, D_k, ξ_k denote RGB, depth, and camera pose. The task objective in Equation (2) balances successful target verification against travel and inspection costs. N1–N3 fix the target after the query to isolate query-time inference and search, whereas N4 allows additional hidden transitions while the agent moves.

A.2 DETAILED 4D MEMORY CONSTRUCTION

Figure 7 illustrates how observations are converted into persistent entity versions. Each observation contributes spatial evidence and temporal validity, allowing the memory to preserve both previous and current states. Memory is materialized causally from versioned deltas as defined in Equation (3), retaining semantic–spatial relations, time-valid entity versions, and observation provenance. Persistent identity is separated from mutable state: an observed change appends a time-valid version without erasing earlier states, and each memory query uses only evidence available at its decision time.

For image coordinate $\tilde{\mathbf{u}} = [u, v, 1]^\top$ with depth $D_t(u, v)$, the corresponding world point is

$$\mathbf{x}^w = \text{proj}_3 \left(\mathbf{T}_{w \leftarrow c, t} \begin{bmatrix} D_t(u, v) \mathbf{K}^{-1} \tilde{\mathbf{u}} \\ 1 \end{bmatrix} \right), \quad (17)$$

where $\mathbf{K}$ is the camera intrinsic matrix and $\mathbf{T}_{w \leftarrow c, t}$ is the camera-to-world transform. Cross-view association uses semantic compatibility, 3D overlap, temporal consistency, and source confidence. The m -th version of entity o is

$$h_o^m = ([t_{o,s}^m, t_{o,e}^m], s_o^m, c_o^m, f_o^m, \mathcal{P}_o^m), \quad (18)$$

containing a validity interval, candidate state, confidence, visual feature, and evidence handles. An observed state change appends a version without erasing the earlier trajectory, preserving both time-ordered history and observation provenance.

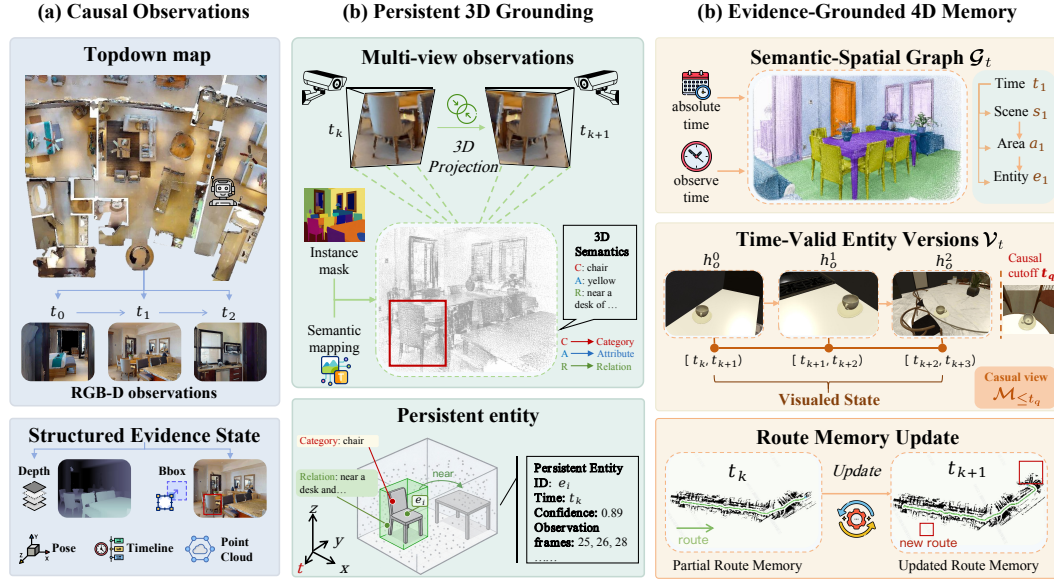


Figure 7: Temporally versioned 4D memory construction.

A.3 BELIEF ENCODING AND TRAINING

The retrieved history and event representation are defined in Equations (4) and (5); the Transformer in Equation (6) produces the query representation used by both prediction heads. Event tokens retain entity, state, evidence type, confidence, and visibility alongside elapsed-time and calendar features. The query token specifies the target, query time, and last positive state. The persistence probability in Equation (7) measures agreement with that last positive observation, including departures followed by returns. The shared pointer head in Equation (8) scores every alternative, including unknown, from the query and candidate representations. Its candidate encoder combines context, functional role, geometry, and instance identity. The persistence and pointer heads jointly define the normalized belief in Equation (9), trained with current-state negative log-likelihood without a separate change label or oracle mobility type.

A.4 EVIDENCE UPDATES AND ONLINE PREDICTION

Transition-kernel parameterization and training. The transition operator in Equation (10) is not inferred from a single marginal belief. It is an additional row-conditional prediction head that shares the continuous-time history encoder and candidate embeddings with the query-time predictor. Its training set contains chronological tuples $(H_{\leq t_a}, s_{t_a}, \Delta t, s_{t_a+\Delta t})$ extracted only from training world histories. For each tuple, the known source state selects one row and the future state supervises

$$\mathcal{L}_{\text{trans}} = -\log[K_{\theta}(\Delta t \mid H_{\leq t_a})]_{s_{t_a}, s_{t_a+\Delta t}}, \quad \mathcal{L} = \mathcal{L}_{\text{state}} + \lambda_K \mathcal{L}_{\text{trans}}. \quad (19)$$

Horizons are sampled from the empirical action-chunk and candidate-arrival range. The softmax over destination states makes every row sum to one. No transition label, future state, or mobility tag is available at test time. In N1–N3, where the target is fixed after the query, K is the identity; learned propagation is activated only for N4.

Decision epochs and ETA. The filter begins with $b_0^- = b_q^-$ and $b_0^+ = b_0^-$ when no query-time measurement is available. An epoch is triggered after a short action chunk, when a view adds sufficient candidate coverage, when the route or ETA changes materially, when target evidence is obtained, or when the robot reaches an inspection viewpoint. After a chunk of observed duration $\Delta t_j = t_{j+1} - t_j$, Equation (11) first produces the prior at the new current time. For candidate i , the navigation backend then estimates

$$\hat{\tau}_{j,i} = \frac{d_{\text{geo}}(x_j, \mathcal{W}_i)}{\hat{v}_j} + \tau_{\text{inspect}}(i), \quad \hat{v}_j = \alpha \hat{v}_{j-1} + (1 - \alpha) \frac{\Delta d_j}{\Delta t_j}. \quad (20)$$

All ETAs are recomputed after obstruction or replanning. Candidate-specific arrival beliefs are forecasts from the current posterior, not filtering states at the current time, and never replace b_j^+ unless their horizon is actually executed.

Complete event-driven filter. The implementation follows the same ordering at every decision epoch:

1. Initialize $t_0 = t_q$, $b_0^- = b_q^-$, incorporate any valid query-time measurement once to obtain b_0^+ , and initialize an evidence ledger for every candidate.
2. From b_j^+ , forecast each $b_{j,i}^{\text{arr}}$ to its own ETA, score known candidates and EXPLORE, and select the highest-utility action.
3. Execute only a short action chunk and record its actual elapsed time, displacement, pose, and RGB-D evidence.
4. Propagate to the new current time: $b_{j+1}^- = b_j^+ K_\theta^{(j)}(t_{j+1} - t_j)$.
5. Create measurements only for candidates whose evidence ledger admits new information; apply Equation (14) once per evidence identifier to obtain b_{j+1}^+ . Stop on a verified positive detection.
6. Update speed and ETAs, forecast all candidate-specific arrival beliefs from b_{j+1}^+ , reopen eligible states, and replan if the preferred action changes; otherwise execute the next short chunk.

This explicitly separates the current-time filter $b_j^- \rightarrow b_j^+$ from the counterfactual arrival forecasts $\{b_{j,i}^{\text{arr}}\}_i$.

Time-valid evidence rounds without leakage. Candidate i maintains an evidence-round index $m(i)$ and $C_j^{(m)}(i)$, the union of surface samples observed within that round. A no-detection event is created only if the previously unused coverage satisfies $\Delta C_j^{(m)}(i) > \delta_C$, with $\delta_C = 0.05$. Candidate surface samples from 4D memory are projected into the live camera, and online depth marks samples as visible when they lie in the frustum and agree with the measured depth. The resulting online-estimated coverage and other view features form $\mathbf{f}_{j,i}$ for the calibrated detector model g_ϕ in Equation (13). Crucially, $\hat{r}_{j,i}$ is evaluated on the newly admitted surface samples and rays, not on the entire overlapping image. One view can update several candidates with different strengths. Opportunistic views and deliberate inspections use the same rule, although the latter usually provide more coverage. A candidate is counted as a sufficiently covered inspection when $C_j^{(m)}(i) \geq \tau_{\text{cov}} = 0.70$; this designation does not permanently remove it from the candidate set.

Every measurement has a unique evidence identifier and is incorporated once. Its RGB-D frame and pose remain in memory as provenance, but its likelihood is not multiplied again or reintroduced as an independent history token into the current filter. Ground-truth target masks, poses, unoccluded fractions, and simulator visibility flags are retained only by the private evaluator; an oracle-visibility diagnostic is reported separately from fair comparisons. Incremental masking and evidence identifiers do not assert that video frames are statistically independent; they prevent direct reuse of the same surface evidence. The remaining dependence is the filter’s conditional-measurement approximation and is absorbed by the validation-fitted calibrator. We never multiply repeated whole-view detection probabilities from overlapping frames.

Reopening candidates and executing unknown. There is no permanent exclusion set. Each candidate stores cumulative coverage, the time of its latest clear negative, and cumulative detection probability. After propagation, a previously inspected candidate becomes eligible when its belief exceeds $\tau_b = 0.10$, predicted return probability

$$P_{\text{return}}(i) = \sum_{r \neq i} b_j^+(r) K_\theta^{(j)}(r, i; \Delta t) \quad (21)$$

exceeds $\tau_{\text{return}} = 0.05$, or a new viewpoint offers $\Delta C(i) > \delta_C = 0.05$. Thus a weak or occluded view supports multi-view reinspection, and a location verified empty can regain mass after sufficient world evolution.

Reopening by a new viewpoint alone retains the current round, so only genuinely new geometry is admitted. Reopening caused by propagated belief or predicted return after elapsed time starts round $m(i) + 1$ and resets only that round’s coverage gate; the earlier coverage map and negative observations remain in the provenance ledger. Consequently, the same viewpoint can provide a new measurement after the world may have changed, while adjacent frames within one state-validity interval cannot repeatedly suppress the same hypothesis.

The unknown state invokes an executable EXPLORE action with

$$U_j(\text{EXPLORE}) = \frac{b_j(\text{unknown}) \hat{\eta}_j^{\text{discover}}}{c_j^{\text{explore}} + \lambda_{\text{scan}}}. \quad (22)$$

The controller selects a semantic frontier, uncovered region, or unopened container, performs an open-vocabulary scan, adds discovered states to $\mathcal{C}_o$, redistributes unknown mass, and replans. NOT_FOUND is allowed only after the exploration budget is exhausted (or no valid frontier remains) and both unknown mass and remaining searchable mass fall below fixed validation thresholds. The same filter supports N1–N4 without a static/dynamic policy router.

A.5 FULL BASELINE SPECIFICATION

Last Seen assigns all mass to the latest positive state. Frequency Prior estimates $P(s \mid o)$ from training data. Markov Transition estimates $P(s_{k+1} \mid s_k, o)$; Time-Conditioned Prior additionally conditions on public hour and weekday bins. Instance Hotspot uses only the target instance’s training trajectory. GRU Direct and Transformer Direct receive the same event tokens and candidate encoder as P4D-NAV but directly normalize candidate logits without the persistence factorization. P4D-Belief Prior uses Equation (9) once at episode start. Full P4D-NAV applies Equations (12), (14) and (15) at event-driven decision epochs.

The Oracle Activity diagnostic may use the hidden activity and source location but is excluded from fair rankings. Oracle Current State receives the private target state and measures remaining perception and navigation error. External predictive methods are adapted only through their public inputs; any use of privileged simulator state must be identified and excluded from the main comparison.

A.6 IMPLEMENTATION DETAILS

Prediction model and optimization. The history encoder uses three Transformer layers, hidden width 128, four attention heads, maximum history length $K = 64$, and dropout 0.1. Candidate encoders are shared across scenes. The transition scorer f_θ is a two-layer width-128 MLP with GELU activation; it shares the history and candidate encoders but not the persistence, relocation, or transition output heads. We set $\lambda_K = 1$ in Equation (19). We optimize with AdamW at learning rate 3×10^{-4} , weight decay 10^{-2} , and batch size 64 for at most 100 epochs. Gradients are clipped to an ℓ_2 norm of 1.0. Early stopping uses patience 10 on validation NLL, and the checkpoint with the lowest validation NLL is retained. We report seeds $\{0, 1, 2, 3, 4\}$. Experiments use Habitat-Sim 0.3.3 and Habitat-Lab 0.3.3 on $3 \times$ NVIDIA GeForce RTX 5080 GPUs.

Perception and controller. The frozen perception stack uses Grounding DINO (Liu et al., 2024b) with box threshold 0.35 and text threshold 0.25, followed by SAM 2 (Ravi et al., 2025) only for mask refinement. The main high-level agent uses GPT-5.6-Luna as a frozen zero-shot VLM controller, without navigation-task fine-tuning. Candidate viewpoints and low-level planning are fixed in the high-level track; a Habitat low-level action track evaluates the complete embodied stack.

Detection calibration and information boundaries. The function g_ϕ in Equation (13) is a lightweight logistic calibrator whose inputs are online-estimated candidate coverage, range, viewing angle, projected size, image quality, category, and validation-set detector recall. It is fitted only on validation observations and remains frozen throughout test evaluation. The agent never accesses ground-truth target masks, ground-truth target poses, unoccluded target fractions, or simulator visibility flags.

All architecture choices, optimization hyperparameters, calibration models, perception thresholds, policy thresholds, and checkpoint-selection rules were determined exclusively from the training and validation splits. The test split was not used for model selection or hyperparameter tuning and was accessed only for final evaluation.

A.7 EXPERIMENTAL DETAILS

Benchmark protocols. The complete agent is evaluated on EVOLVINGWORLD-BENCH, FindingDory, and GOAT-Bench under their native task definitions. FindingDory reports high-level goal selection and conditional low-level execution; GOAT-Bench reports official SR and SPL, with Repeat-SR used only as a marked memory-reuse diagnostic. The cross-domain matrix uses R2R-CE on MP3D, HM3D-OVON on HM3D, native PointNav on Habitat-GS, SAGE-Bench on InteriorGS, and EVOLVINGWORLD-BENCH on HSSD. Scores are never pooled across these heterogeneous tasks, and published values are transferred only when split, sensors, actions, and success rules match.

Result provenance and comparison scope. The entries in Tables 1 and 8 come from two explicitly separated sources. A cell marked ^{PR} is transcribed from the cited paper only when the benchmark version, evaluation split, metric, and success rule match; the marker applies to that cell rather than

to an entire method row. Every unmarked baseline entry is reproduced by us. Missing values are left unreported rather than inferred from another task or checkpoint. Accordingly, Table 1 evaluates memory and prediction under each benchmark’s task definition, whereas Table 8 measures cross-environment system performance under native navigation tasks. Neither table is interpreted as a single architecture-controlled ablation; causal attribution to our predictive belief is instead provided by the paired-world control and component ablations.

Test-time information boundary. All reproduced methods receive only information public in the target benchmark. On EVOLVINGWORLD-BENCH, observations are truncated at the current decision time. Count and Markov baselines use only the target category, public time, and latest supported state; structured memories such as HOV-SG and DynaMem replay the same causal RGB-D and poses; predictive models receive public state or edge histories and the legal candidate graph. Released navigation agents consume only their native RGB-D/video window, goal specification, and proprioception. The paper-reported FindingDory cell specifically uses the cited release’s Qwen2.5-VL-3B checkpoint and native 96-frame history; it is not recomputed with our planner. No reproduced method receives the private current state, future observations, query-time activity, target pose, oracle viewpoint, or simulator visibility flags. Oracle rows, where present, are diagnostics and are excluded from fair rankings.

Training and model selection. Non-parametric baselines are estimated from the public training split, with smoothing or time bins fixed on validation. Internal neural baselines use the same event tokens, candidate encoder, training scenes, and validation rule as P4D-NAV. External predictive architectures that require target-domain fitting are trained only on EVOLVINGWORLD-BENCH days 0–79 and selected on days 80–84; days 85–89 remain test-only. When an official checkpoint exists for a native benchmark, we retain that checkpoint and its prescribed observation window. Task-adapted systems are labeled separately from zero-shot systems in Table 8; interface conversion is not counted as navigation-policy training.

Prediction splits. Each scene history follows a 90-day chronological split: days 0–79 are used for training, days 80–84 for validation, and days 85–89 for testing. Temporal evaluation pools the held-out test periods and reports 322 changed-only queries; leave-one-scene-out evaluation rotates the held-out scene while preserving the same chronological boundaries and reports 644 change-balanced queries. All predictors share query identities, time-ordered histories, and legal candidate sets. The main table reports five-seed variation; no future observation or private transition label is available at inference time.

Real-world protocol. LYNX M20 is the primary quantitative platform, with X30 and Lite3 used for transfer. The primary M20 comparison uses 64 matched trials, whereas X30 and Lite3 use matched 32-block transfer subsets. Each method receives identical pre-query history, candidate locations, start condition, perception interface, and robot-specific low-level stack. Success requires autonomous online confirmation within three inspections during a 360-second episode. Missed detections, navigation failures, timeouts, and human interventions count as failures. Indoor/outdoor balance, platform specifications, and condition-wise breakdowns are provided in Appendix D.2.

Reporting controls. We use two method-preserving adaptation regimes. Memory-only and prediction-only systems retain their native representation or predictor, map their outputs onto the legal candidate states, and use the shared planner, inspection viewpoints, and action budget; this isolates temporal reasoning from low-level control. End-to-end navigation agents retain their released perception, mapping, and policy, with adapters limited to sensor conventions, task-goal formatting, action vocabulary, and STOP semantics. We do not add our VLM planner or predictive belief to those agents. Their cross-domain scores therefore measure full-system compatibility, while the controlled EVOLVINGWORLD-BENCH and real-robot comparisons support module-level claims. Results are stratified by world variant, staleness, transition class, mobility regime, and generalization split. Paired routine/static/random episodes retain scene, query, and marginal placement factors while changing only the hidden transition mechanism.

B BENCHMARK CONSTRUCTION AND EVALUATION PROTOCOL

This section documents episode validity, leakage controls, mobility assignment, and the metrics used for all reported comparisons.

B.1 CONSTRUCTION AND AUDIT DETAILS

Design principles. EVOLVINGWORLD-BENCH follows four principles: temporal continuity, past-only observability, physical executability, and controlled dynamics. Examples are sampled from persistent scene histories rather than independent placements. Public memories contain only evidence available to the robot by the query time. Candidate states must admit collision-free placements and reachable, visibility-checked inspection viewpoints. Paired worlds hold the scene, target, query, and marginal placement factors fixed while varying the hidden transition mechanism. Together, these controls separate predictive reasoning from scene frequency, route geometry, and privileged simulator access.

Source alignment and accounting. Source records constrain the generator rather than serving as episode templates. We distinguish raw source rows or sequences (N_{raw}), records after source-specific normalization (N_{std}), distinct normalized records referenced by at least one retained generation rule (N_{used}), and total rule references with reuse allowed (N_{ref}). Table 5 reports counts from the audited consumption manifests; reuse therefore does not inflate the number of independent source observations.

The raw unit is a sensor row for CASAS, ARAS, and OPPORTUNITY, an annotation row for HD-EPIC, and a sequence for ParaHome and HOMER+. CASAS events are mapped to the room ontology and split into sessions at inactivity gaps above 900 seconds; the resulting occupancy, time-of-day, and weekday statistics parameterize household schedules.

ARAS and OPPORTUNITY sensor streams are segmented into activity intervals and normalized to the shared activity, room, and object vocabularies, supplying complementary activity compatibility and temporal-context statistics. HD-EPIC narration verbs and nouns are mapped through canonical action and object aliases to estimate object-action frequencies, without assigning HSSD destinations. ParaHome annotations and object transforms provide displacement and transition-timing evidence. HOMER+ remains explicitly marked as simulated and contributes only long-horizon activity order and terminal-state priors, not independent human observations or test-time model inputs.

Mapping rules and measured parameters. All adapters emit a common schema containing time, activity, actor context, object category, source state, destination state, and provenance. Measured quantities comprise CASAS occupancy timing, ARAS/OPPORTUNITY activity context, HD-EPIC action-object frequencies, ParaHome displacement and transition timing, and HOMER+ simulated order priors. Legal receptacles, reachable viewpoints, and collision-free placements are derived from HSSD geometry. Category-to-receptacle allowlists, affordance constraints, stochastic exception rates, paired static/routine/random interventions, the 90-day horizon, and query sampling are benchmark-authored and labeled as such. Thus no source trajectory is copied verbatim and no source count is expanded into a claimed number of independent human observations. The final manifest spans 54 scenes and 4,860 scene-days, producing approximately 239.19k dynamic state changes and 803.68k task instances after physical-validity, observability, and leakage audits.

Household generation and behavioral checks. For each HSSD scene (Khanna et al., 2024a), resident schedules, room preferences, orderliness, object ownership, and placement habits are fixed at the household level. Activities, resident region trajectories, and object transitions are generated jointly in chronological order over multi-day histories; stochastic events, delayed returns, and exceptions prevent deterministic histories. Household profiles, transition rules, and task templates instantiate executable N1–N5, VLN, and EQA episodes. Every relocation is attributed to an activity, resident transition, object lifecycle event, or tidying event. Personal objects alternate between placed and carried states with owner-specific return habits; irregular objects use affordance-valid destinations without temporal or actor-specific predictability. Audits check that personal objects do not simply track their owners and that actor or activity identity does not predict irregular destinations beyond affordance. The audited configuration yields 42.6 region transitions per resident-day, 6.56 relocations per personal object-day, 52.6% within-region relocations, and zero terminal collisions.

The realized transition density in Figure 8 provides a direct check of the paired-world intervention. Routine worlds retain repeatable, activity-linked time bands, whereas random worlds spread changes across the day while matching the marginal movement process. Thus the control removes predictable

Table 5: Audited source contributions.

Source	N_{raw}	N_{std}	N_{used}	N_{ref}
CASAS Aruba	1,602,820	20,252	428	480
ARAS	5,184,000	5,08	312	360
HD-EPIC	59,454	10,786	742	864
ParaHome	212 seq.	860	186	216
OPPORTUNITY	869,387	2,551	318	372
HOMER+ (sim.)	65 seq.	1,991	356	420

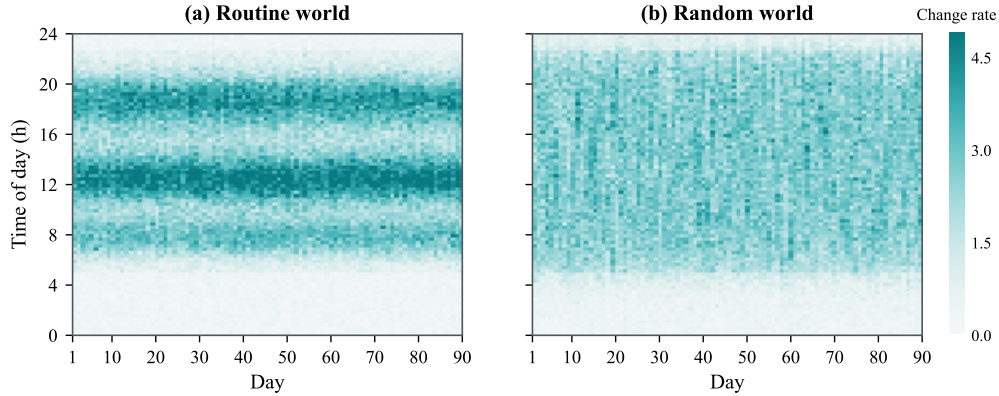


Figure 8: Temporal change density in paired routine and random worlds.

temporal phase without changing the 90-day horizon or scene setup. The “Change rate” color scale is the mean number of object relocations per scene-day in each one-hour time-of-day bin.

B.2 HUMAN REVIEW PROTOCOL

Every retained benchmark episode passes a two-stage human quality-control process. Five trained researchers perform the primary annotation using a structured checklist, and one supervising researcher conducts the final audit. For each object transition, reviewers verify identity continuity, a physically valid and reachable destination, collision-free placement, consistency with the scene geometry and visibility model, and a transition time that lies between the relevant observations without exposing future information. They also record whether the relocation reason is compatible with the attributed activity, actor context, object affordance, or household routine. A destination may be temporarily unobserved by the agent, as required by the benchmark, but it cannot be impossible, outside the navigable scene, or revealed through the public history. Reviewers additionally check that state changes, returns, and negative evidence are temporally ordered and semantically consistent.

Tasks are reviewed for target–episode correspondence, referent clarity, executability, and research relevance. The target category or instance must be instantiated in the private world state (or explicitly belong to a valid not-found track), candidate states and viewpoints must be reachable, and the action budget must permit a meaningful search. Language goals must identify the intended object without answer-correlated names or future-state hints. Finally, each task is checked to ensure that it probes the paper’s setting—stale observations, hidden evolution, predictive belief, or visibility-aware evidence updates—rather than reducing to a static or trivially solvable navigation case. Reviewers assign accept, revise, or reject decisions; disagreements are resolved by majority vote, while unresolved or high-risk cases are adjudicated by the supervisor. Rejected cases are regenerated or removed, and the audit log records the decision and reason for every retained item.

Candidate metadata and query sampling. Chronological Habitat replay records navigability, geodesic cost, camera frustum coverage, unoccluded target fraction, and success viewpoints for each candidate. These fields support the physical validity and visibility checks described in Appendix B.4; candidates without collision-free placements and reachable inspection viewpoints are excluded. Query times sample both the natural temporal distribution and mechanism-relevant windows, including routine stages and intervals after a personal object is dropped. Each navigation query is augmented with a start pose, candidate inspection viewpoints, a path budget, and frozen success conditions. The patrol policy cannot access current object states, resident locations, hidden activities, or generator latents. Query-time activity and hidden transitions are retained only by the evaluator.

Language tasks and release documentation. EQA questions cover existence, location, temporal order, activity association, count, comparison, and compositional relations over the same time-indexed memories. Answers are evaluated against a normalized ontology, and questions requiring future evidence are excluded. VLN replaces the object label with a language goal whose referents and constraints are grounded in the same versioned scene history. The broader task annotations support studies beyond the prediction and navigation experiments emphasized in this paper. Across tracks, only memory representations, current-state beliefs, and resulting decisions differ between controlled

methods. Dataset cards and audit summaries document source mappings, authored assumptions, physical validity, candidate coverage, and leakage tests.

Splits and leakage checks. Each 90-day history uses days 0–79 for training, days 80–84 for validation, and days 85–89 for testing. Complete temporal sequences are split together so that no future portion of a household day enters an earlier decision. Held-out homes are reserved for cross-scene evaluation; paired static/routine/random members remain in the same split. Near-duplicate queries and transition groups are assigned atomically. Serialized releases are checked for future timestamps, private event fields, answer-correlated filenames or indices, repeated event groups across splits, inconsistent paired variants, invalid physics, unreachable goals, and accidental patrol dependence on hidden state. Episode-level checks and mobility assignment are further specified in Appendices B.4 and B.5.

B.3 MOBILITY REGIMES AND TASK PROTOCOLS

Mobility labels are used only for stratified evaluation and are hidden from the agent.

Table 6: Mobility regimes in EVOLVINGWORLD-BENCH.

Regime	Mechanism	Frozen signal	Role
Stable/rare	Infrequent events	Home/return tendency	Persistence control
Routine	Activity state machine	Chain-specific locations	Routine prediction
Personal	Placed/carried lifecycle	Owner placement habit	Long-term memory
Irregular	Affordance sampling	Affordance only	Uncertainty control

Table 7: Prediction and navigation protocols in EVOLVINGWORLD-BENCH.

Task	Decision signal	Primary measure
P Current-state prediction	Region/receptacle belief	R@k / NLL
N1 Predictive navigation	Event-filtered belief	First-Inspection SR
N2 Belief-guided search	Full belief + cost	Search SR / cost
N3 Evidence-aware replanning	Updated posterior	Recovery SR
N4 Online-dynamic navigation	Arrival-time belief	Online Recovery SR
N5 Cross-scene generalization	Any protocol above	Held-out SR

B.4 EPISODE VALIDITY AND LEAKAGE AUDIT

An episode is retained only if: (i) the target has at least one positive pre-query observation; (ii) the target state is known to the private evaluator and physically instantiated without collision; (iii) at least one valid viewpoint lies on the NavMesh; (iv) the start is on the NavMesh, at least 3 meters from the target, and does not reveal it; and (v) all public histories terminate at or before the query. A stop is successful only when the robot is within 1 meter geodesic distance of a valid viewpoint, the private evaluator measures target visible fraction $v_{\text{target}}^{\text{GT}} \geq \tau_{\text{success}} = 0.20$, and the agent identifies the correct target instance or category. This threshold is used only by the private evaluator and is never exposed to the agent.

The three coverage quantities have distinct roles: $\Delta C > 0.05$ triggers a new negative-evidence update, $C \geq 0.70$ defines a sufficiently covered inspection, and $v_{\text{target}}^{\text{GT}} \geq 0.20$ defines evaluation success only. The first two are estimated online from candidate geometry and depth; the third is private evaluator state.

We audit serialized examples for future timestamps, hidden event fields, filenames or indices correlated with answers, duplicate event groups across splits, and inconsistent paired variants. Static, routine, and random members of a pair must have identical public query fields. Invalid simulator episodes are reported separately and are not counted as ordinary failures.

B.5 MOBILITY-TYPE ASSIGNMENT

The frozen rule is

$$\text{type}(o) = \begin{cases} \text{stable}, & r_{\text{move}}(o) < \tau_{\text{move}}, \\ \text{activity}, & G_{\text{act}}(o) \geq \tau_{\text{act}}, \\ \text{personal}, & C_{\text{habit}}(o) \geq \tau_{\text{habit}}, \\ \text{irregular}, & \text{otherwise.} \end{cases} \quad (23)$$

All thresholds are selected on training data before navigation evaluation. When the generator specifies a mobility profile, the same statistics are used to verify that realized trajectories exhibit the intended

behavior. Every group must cover several categories, instances, and scenes, and no category may uniquely reveal one mobility type.

B.6 METRIC DEFINITIONS

Let S_i be episode success, L_i the executed path, and L_i^* the oracle shortest path to the query-time goal under the same start. Standard SPL is

$$\text{SPL} = \frac{1}{N} \sum_{i=1}^N S_i \frac{L_i^*}{\max(L_i, L_i^*)}. \quad (24)$$

First-Inspection SR records whether the first candidate to receive sufficient cumulative coverage contains the target; opportunistic evidence and replanning before that inspection are part of the policy. A sufficiently covered en-route view counts as the first inspection itself. Recovery SR conditions on an unsuccessful first inspection and measures whether the target is subsequently found. Excess path is $L_i - L_i^*$ for fixed-during-navigation episodes. For online-dynamic episodes, the oracle is trajectory aware and may intercept the target; we therefore report both success within the common budget and $L_i - L_{i,\text{dyn}}^*$. Dynamic SR is computed only on N4 episodes with at least one target transition between query and termination. Online Recovery SR conditions further on a transition that invalidates the currently selected destination and records eventual success after online evidence and replanning. Revisit Success is the percentage of dynamically reopened candidate visits that recover the target. N4 episodes in which evolution is enabled but no target transition is realized form a separate no-transition slice and are not mixed into these dynamic metrics.

Calibration metrics operate on the full candidate belief before navigation and after each update. ECE bins maximum confidence, Brier score evaluates all candidates, and true-state rank avoids hiding useful changes below Top-1. Search-cost curves plot success against allowed inspections and travelled distance rather than reporting one arbitrary budget.

C ADDITIONAL EXPERIMENTAL RESULTS

This section provides extended quantitative results omitted from the main paper for space constraints.

C.1 CROSS-ENVIRONMENT NAVIGATION

SR and SPL are reported in percent for all environments.

Table 8: Cross-environment navigation.

Method	Standard Simulation Worlds				Photorealistic Simulation				Evolving Simulation Worlds	
	R2R-CE		HM3D-OVON		PointNav		SAGE-Bench		EVOLVINGWORLD-BENCH	
	SR	SPL	SR	SPL	SR	SPL	SR	SPL	SR	SPL
<i>Navigation-trained or task-adapted</i>										
MTU3D (Zhu et al., 2025)	8.8	6.5	40.8 ^{PR}	12.1 ^{PR}	75.8	72.7	25.9	15.6	35.2	30.7
NaVid (Zhang et al., 2024)	37.4 ^{PR}	35.9 ^{PR}	60.3	36.0	15.3	13.4	17.4	15.1	31.0	22.6
Uni-NaVid (Zhang et al., 2025a)	47.0 ^{PR}	42.7 ^{PR}	39.5 ^{PR}	19.8 ^{PR}	15.9	13.2	26.4	23.1	55.3	41.7
NaVILA (Cheng et al., 2024)	54.0 ^{PR}	49.0 ^{PR}	18.3	9.7	78.4	77.5	39.0	34.0	35.2	30.7
<i>Zero-shot navigation or task adapters</i>										
HSGM (Li et al., 2026a)	47.9 ^{PR}	32.8 ^{PR}	40.0	32.5	37.2	35.3	11.8	9.4	17.0	12.5
VLFM (Yokoyama et al., 2024)	12.0	9.6	35.2 ^{PR}	19.6 ^{PR}	53.0	42.2	24.4	18.8	35.0	25.8
AO-Planner (Chen et al., 2025)	25.5 ^{PR}	16.6 ^{PR}	30.9	22.0	76.6	74.7	27.4	19.7	37.0	29.3
TANGO (Ziliotto et al., 2025)	14.3	9.87	35.5 ^{PR}	19.5 ^{PR}	35.8	33.5	22.7	14.7	31.0	23.3
Uni-LaViRA (Ding et al., 2026)	60.7^{PR}	47.7^{PR}	60.0 ^{PR}	40.5 ^{PR}	81.3	77.7	35.6	28.7	65.2	39.4
MSGNav (Huang et al., 2026)	11.6	9.4	48.3 ^{PR}	27.0 ^{PR}	37.2	34.8	23.5	17.5	26.7	19.6
EVOLVINGNAV (ours)	55.8	43.1	64.2	51.7	86.2	79.8	42.4	32.1	86.6	68.2

^{PR} Paper-reported under the cited native protocol; unmarked baseline scores are our reproductions under the method-preserving protocol in Appendix A.7.

C.2 FULL PREDICTION RESULTS AND CALIBRATION

The temporal split contains 322 changed-only queries and LOSO contains 644 change-balanced queries. Ours reports mean $\pm$ standard deviation over five seeds; Top-1 is a percentage, while MRR and NLL are unscaled.

Figure 9 compares reliability (left) with the distribution of predictions across confidence bins (right), showing how evidence updating changes both calibration and confidence concentration.

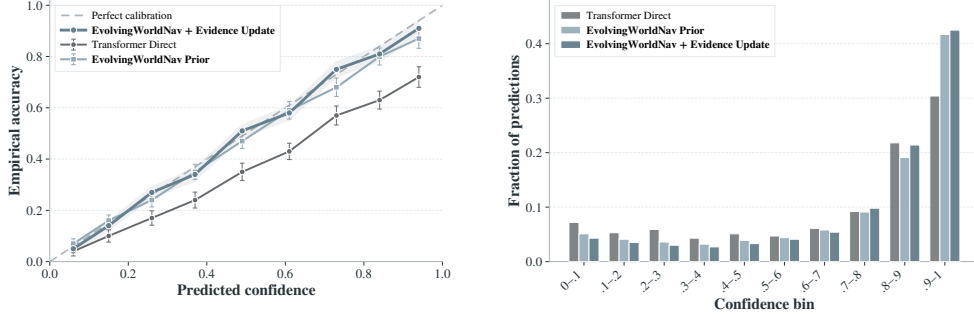


Figure 9: Calibration and confidence distributions of current-state beliefs before and after evidence updating.

Table 9: Full current-state prediction results.

Predictor	Temporal Top-1 $\uparrow$	MRR $\uparrow$	NLL $\downarrow$	LOSO Top-1 $\uparrow$	MRR $\uparrow$	NLL $\downarrow$
Uniform Random	9.32	0.2607	2.4509	11.96	0.3050	2.2980
Last Seen	0.00	0.1699	3.4011	50.00	0.5850	1.8825
Object Frequency	30.75	0.5476	1.7397	50.47	0.6441	1.8455
Time Frequency	33.85	0.5462	1.8400	49.53	0.6166	1.6455
Markov	18.94	0.4435	2.3772	49.38	0.6286	1.6982
Activity-Markov	25.16	0.4582	2.2436	49.84	0.6221	1.6183
Random Forest	20.50	0.4851	1.9181	58.07	0.7283	1.1441
Packed MLP	29.50	0.5225	2.1036	58.07	0.7173	1.3105
GRU	32.61	0.5471	1.8279	54.04	0.7016	1.2558
Direct Transformer	34.16	0.5586	1.7292	51.71	0.6884	1.3157
Shared semantic prior	36.02	0.5957	1.5766	59.47	0.7462	1.0448
Ours	38.24	0.5971	1.4925	61.18	0.7523	0.9778
	± 1.42	± 0.0291	± 0.0867	± 1.53	± 0.0176	± 0.0228

C.3 REAL-WORLD SEARCH CASE

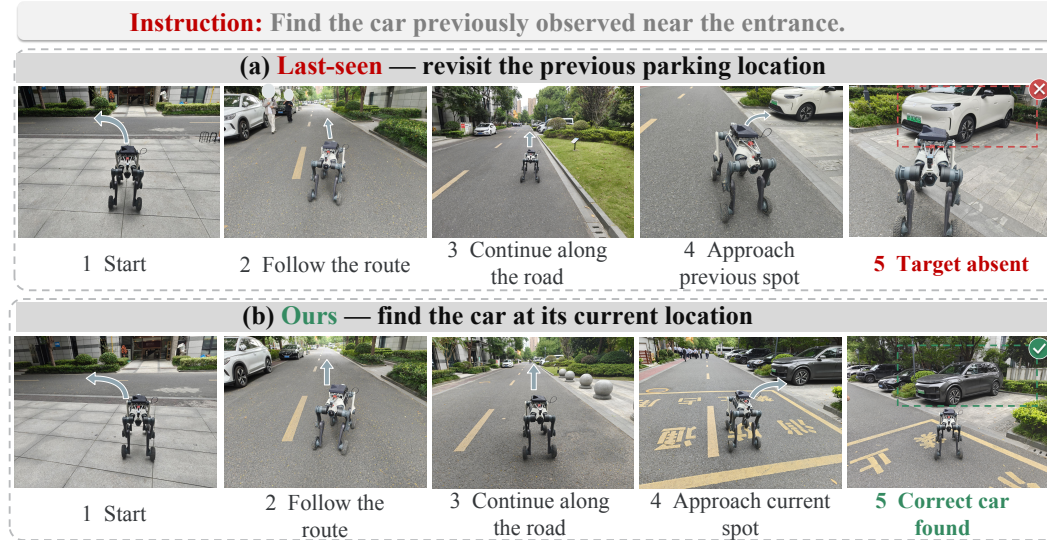


Figure 10: Outdoor car search under stale and predictive memories.

C.4 EVOLVINGWORLD-BENCH DETAILED RESULTS

Table 10 complements the main benchmark table with search cost and recovery after an unsuccessful first inspection; lower is better for the two cost columns. Table 11 tests whether paired navigation

gains persist across different VLMs under matched tools, prompts, episodes, and action budgets. In the VLM study, SR is reported in percent and SPL is unscaled.

Table 10: EvolvingWorld-Bench routine-world search cost and recovery.

Method	Inspect. ↓	Distance (m) ↓	Recovery SR (%) ↑
<i>Navigation and persistent memory</i>			
VLfM (Yokoyama et al., 2024)	4.85	18.72	22.41
LagMemo GLUE (Zhou et al., 2025)	3.92	13.24	45.62
HOV-SG (Werby et al., 2024)	3.36	12.08	51.24
DynaMem (Liu et al., 2024a)	2.51	11.59	63.93
<i>Predictive world and state modeling</i>			
SLaTe-PRO (Patel et al., 2023)	4.40	14.58	29.00
SGM+NEP (Kurenkov et al., 2023)	4.08	7.42	27.56
FlowMaps (Argenziano et al., 2026)	4.66	16.51	20.27
PredictiveGraphs (Saavedra-Ruiz et al., 2026)	3.71	10.48	17.96
EVOLVINGNAV (ours)	2.18	8.36	68.42

EVOLVINGNAV requires the fewest inspections (2.18) and achieves the highest Recovery SR (68.42%), improving over DynaMem by 4.49 points after an unsuccessful first inspection. SGM+NEP travels the shortest average distance (7.42 m), while EVOLVINGNAV remains close at 8.36 m. Together with the First-Inspection, Search SR, and SPL results in Table 1, these complementary metrics show that the gain is not only higher eventual success: retained uncertainty and evidence updating reduce redundant inspections and support recovery when the initial prediction is wrong.

Table 11: VLM robustness on EvolvingWorld-Bench.

Different VLMs	Last-seen Agent			EVOLVINGNAV			Δ Search SR ↑
	First-Inspection SR ↑	Search SR ↑	SPL ↑	First-Inspection SR ↑	Search SR ↑	SPL ↑	
GPT-4o	31.88	48.94	0.3215	59.72	71.86	0.6027	+22.92
GPT-5.5	42.36	51.31	0.3732	63.48	73.57	0.6531	+22.26
GPT-5.6-Luna	46.92	57.43	0.3970	63.87	83.36	0.6639	+25.93
Qwen2.5-VL-3B	16.26	26.55	0.1502	38.27	45.02	0.3727	+18.47
Qwen2.5-VL-32B	18.84	31.30	0.1591	42.44	46.21	0.3713	+14.91
Average gain							+20.90

C.5 FINE-GRAINED ABLATION ANALYSIS

The main-paper ablation isolates the five decisions that define the complete agent. Here we decompose implementation choices within temporal memory, evidence updating, and the predictor while preserving the same perception, candidate set, inspection viewpoints, and low-level navigator.

C.5.1 TEMPORAL AND MEMORY ABLATIONS

SR and SPL are reported in percent throughout this analysis.

Table 12: Temporal and memory ablations.

Group	Variant	First-Inspection SR ↑	Search SR ↑	SPL ↑	Inspect. ↓
Temporal encoding	w/o elapsed-time encoding	48.37	80.28	64.82	4.36
	w/o calendar context	47.75	76.11	66.20	4.05
	w/o both temporal cues	40.84	76.03	64.98	4.12
Historical evidence	w/o instance-specific history	57.57	80.77	66.61	4.30
	w/o negative history	45.86	79.53	63.83	4.44
History capacity	$K = 8$	39.55	63.77	56.00	4.57
	$K = 16$	45.88	69.48	65.77	4.37
	$K = 32$	56.25	79.60	68.40	4.15
	$K = 64$ (Full)	60.23	84.63	70.34	3.98

Elapsed-time and calendar ablations distinguish irregular observation gaps from daily and weekly periodic context. Removing both cues reduces First-Inspection SR by 19.39 points relative to the full model. Instance-specific trajectories provide smaller but consistent gains, whereas removing negative history costs 14.37 First-Inspection points and 6.51 SPL points. Increasing capacity from $K = 8$ to $K = 64$ raises First-Inspection SR by 20.68 points, Search SR by 20.86 points, and SPL

by 14.34 points while reducing the average inspection count from 4.57 to 3.98. These trends show that performance benefits from temporally structured evidence over a sufficiently long history, rather than merely the most recent observations.

C.5.2 EVIDENCE UPDATE ABLATIONS

Figure 11 illustrates one deliberate-inspection event in the same stepwise filter used for opportunistic views. Causal memory initially assigns probability 0.62 to the last-seen coffee table. The frozen logistic calibrator maps the online view features $\mathbf{f}_{j,i}$ to $\hat{r}_{j,i} = 0.931$; when the target is not detected, the likelihood at that state is therefore 0.069. The posterior consequently shifts from (0.62, 0.25, 0.13) to (0.10, 0.59, 0.31) and the planner selects the kitchen island next. The inspected state retains nonzero mass, reflecting residual perceptual uncertainty rather than an irreversible deletion.

Top-1 is reported as a percentage in Table 13.

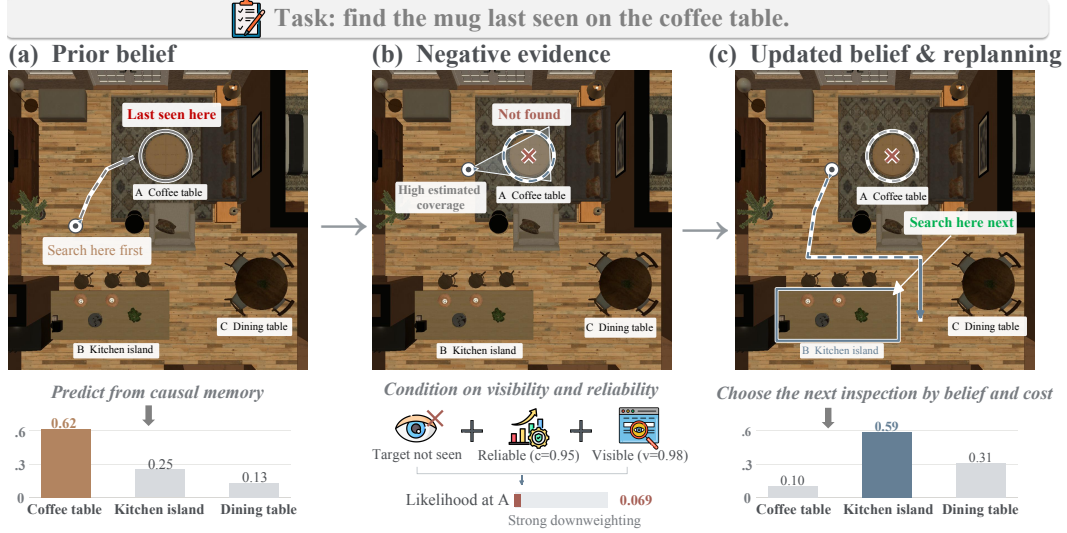


Figure 11: Evidence-aware belief updating and replanning after a negative inspection.

Table 13: Belief quality after negative-evidence updates.

Update rule	Rank ↓	Top-1 ↑	NLL ↓	ECE ↓
No Update	5.82	62.98	2.838	0.172
Hard Removal	10.09	42.69	11.514	0.204
Bayesian Update	5.25	64.71	1.497	0.151
Bayesian + Calibration (Ours)	4.96	66.93	1.530	0.130

No Update leaves the prior unchanged, while Hard Removal assigns zero mass after every unsuccessful inspection. This irreversible rule performs worst across all four metrics, showing that an apparent miss is insufficient evidence when the view is partially occluded or perception is uncertain. Bayesian updating lowers NLL from 2.838 to 1.497 without discarding hypotheses outright. Adding calibrated view-conditioned likelihoods produce the best mean rank (4.96), Top-1 accuracy (66.93%), and ECE (0.130), while its NLL remains within 0.033 of the uncalibrated Bayesian update. The result supports viewpoint-conditioned negative evidence as the most reliable input to the observe–update–replan loop.

C.5.3 PREDICTOR ARCHITECTURE

Table 14: Predictor architecture on changed-only queries.

Predictor	Top-1 ↑	MRR ↑	NLL ↓
GRU	32.61	0.5471	1.8279
Direct Transformer	34.16	0.5586	1.7292
Structured Transformer (ours)	38.24 ± 1.42	0.5971 ± 0.0291	1.4925 ± 0.0867

All three predictors use the same observation tokens and candidate encoder. The Direct Transformer normalizes candidate logits directly, whereas the structured model separates persistence from relo-

cation. This controlled view therefore isolates architectural factorization from memory content and action selection.

C.6 EVOLUTION-AWARE DIAGNOSTICS

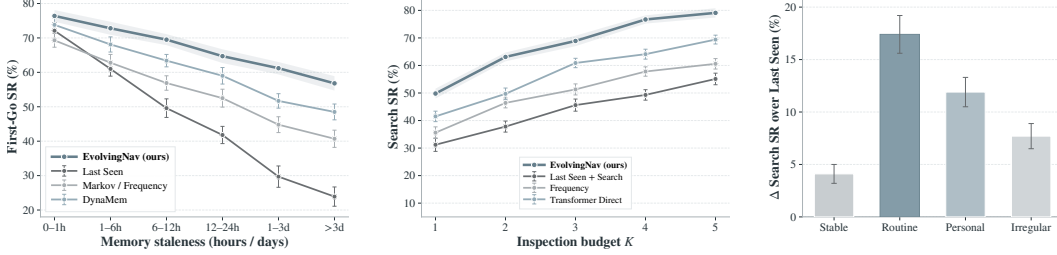


Figure 12: Predictive diagnostics under memory staleness and search budgets.

Paired-world control. Table 15 holds scenes and queries fixed while changing only the world transition mechanism. Last Seen reaches 84.38 SR in static worlds but falls to 38.49 under routine evolution, exposing the failure of stale memory. In contrast, P4D-NAV reaches 60.85 routine-world SR, improving by 22.36 points over Last Seen and by 12.03 points over Transformer Direct. Under matched random motion, the gain over Last Seen contracts to 6.93 points and P4D-NAV trails Transformer Direct by 1.27 points. This contrast localizes its main advantage to learnable temporal regularity rather than a scene-frequency or generic search shortcut.

Table 15: Paired-world SR and routine gain (%).

Method	Static SR	Routine SR	Random SR	Routine gain over Last Seen
Last Seen	84.38	38.49	23.19	—
Markov Transition	80.22	32.57	20.64	−5.92
Direct Transformer	69.79	48.82	31.39	+10.33
P4D-NAV	82.19	60.85	30.12	+22.36

Effect of memory staleness and inspection budget. Figure 12 complements these aggregate tables by showing how performance changes with temporal staleness, inspection budget, and hidden mobility type. These episode-level diagnostics are not inferred from aggregate success rates.

Behavior-shift and trace-grounding diagnostic. Figure 13 tests whether the learned regularities are limited to the benchmark generator’s nominal timing. Online Recovery SR falls for all methods under phase offsets, return delays, destination shifts, and their combination, but EVOLVINGNAV remains above Last Seen and PredictiveGraphs in every condition. The duration CDF further shows that the generated transition durations track the broad temporal scale of held-out real traces without exactly reproducing them. This diagnostic supports robustness to moderate behavior shift and temporal grounding beyond a single generator setting; it is not presented as evidence of unrestricted out-of-distribution generalization.

D REAL-WORLD EVALUATION

This section reports the physical evaluation, first describing the robot platforms and then disaggregating the measured M20 results by environment, temporal condition, and qualitative case.

D.1 ROBOT PLATFORMS

We use three commercial DEEP Robotics platforms: LYNX M20, X30, and Lite3. Platform-specific perception and low-level control are abstracted behind a common interface, while the memory, prediction, belief update, and high-level search components remain unchanged.

Table 16: Robot platforms for real-world evaluation.

Platform	Morphology	Standing size	Mass	Endurance / range
LYNX M20	Wheel-legged	820 × 430 × 570 mm	~35 kg	3 h / 15 km unloaded
X30	Quadruped	1000 × 695 × 470 mm	56 kg	2.5–4 h / ≥10 km
Lite3 (LiDAR)	Quadruped	610 × 370 × 496 mm	13.5 kg	1.5–2 h / 2.7 km

Manufacturer-rated endurance and range are descriptive specifications, not experimental outcomes. Absolute completion time is compared only between methods executed on the same platform.

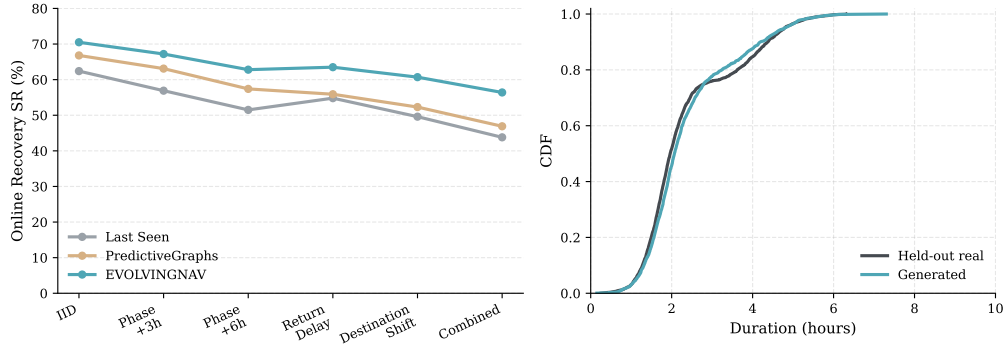


Figure 13: Recovery under behavior shift and temporal grounding to held-out traces.

D.2 ENVIRONMENT BREAKDOWN

Table 17 separates indoor and outdoor trials. The outdoor setting has longer routes and lower success for all methods, while the direction of the EVOLVINGNAV gain is consistent across both environments.

Table 17: LYNX M20 results by environment.

Method	Indoor			Outdoor		
	First-Inspection	Search SR	Dist.	First-Inspection	Search SR	Dist.
Last Seen + Search	18.8	31.3	28.6	15.6	21.9	83.8
Time Frequency	25.0	34.4	27.2	18.8	28.1	78.6
Retrieval + Reasoning	31.3	40.6	25.1	21.9	34.4	75.5
EVOLVINGNAV	40.6	53.1	22.4	28.1	43.8	65.2

D.3 TEMPORAL CONDITION ANALYSIS

Temporal conditions. Table 18 shows that predictive memory is most useful when recurring changes provide a usable historical signal. Last Seen remains strong when nothing changes, whereas all methods degrade as regularity weakens. The condition-wise view separates benefits from routine modeling from improvements that could be explained by generic search alone.

Table 18: LYNX M20 Search SR by temporal condition (%).

Method	Unchanged	Routine-consistent	Weak-routine	Broken-routine
Last Seen + Search	68.8	18.8	12.5	6.3
Time Frequency	50.0	50.0	18.8	6.3
Retrieval + Reasoning	62.5	43.8	25.0	18.8
EVOLVINGNAV	62.5	68.8	37.5	25.0

Recovery and cross-platform transfer. After failed first inspections, EVOLVINGNAV recovered 9/37 M20 episodes (24.3%). Table 19 reports the matched 32-block transfer comparison for X30 and Lite3 alongside the corresponding M20 subset; the primary M20 result still uses all 64 trials. This evaluates transfer, not platform parity.

Table 19: Cross-platform navigation transfer.

Platform	Method	First-Inspection SR	Search SR	Dist. (m)
LYNX M20	Last Seen + Search	18.8	25.0	55.4
	EVOLVINGNAV	37.5	50.0	44.6
X30	Last Seen + Search	18.8	25.0	54.1
	EVOLVINGNAV	31.3	43.8	46.8
Lite3	Last Seen + Search	18.8	31.3	51.8
	EVOLVINGNAV	25.0	43.8	45.9

D.4 QUALITATIVE CASES

The outdoor comparison in Figure 10 isolates stale memory in a car-search task: Last Seen revisits the previous entrance-side parking location and observes that the target is absent, whereas EVOLVINGNAV reaches its current location. The two-row indoor comparison in Figure 14 shows the same failure mode for a cup: EVOLVINGNAV routes to the predicted office location, while Last Seen first inspects the now-empty pantry rack. These traces visualize representative outcomes and do not add observations to the aggregate success-rate measurements.

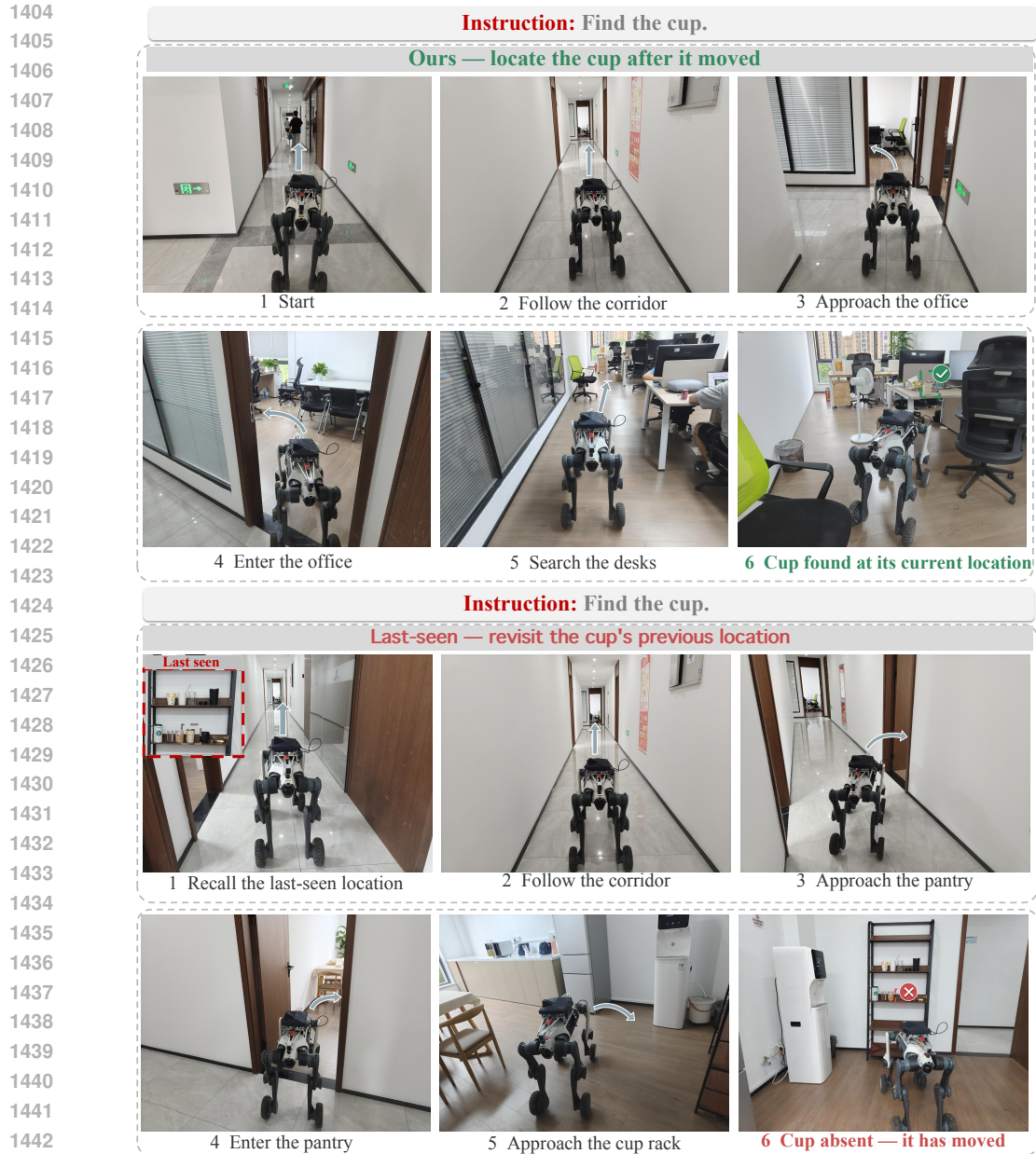


Figure 14: Indoor cup search with predictive belief (top) and stale Last Seen memory (bottom).

Additional system execution cases. Beyond predictive target search, Figures 15 and 16 document two longer-horizon executions. The indoor case combines navigation, person verification, document collection, and delivery. Because the LYNX M20 has no manipulator, after the person is verified, a human places the documents on the robot; the robot then autonomously transports them to the meeting room. The outdoor case shows obstacle detection and route replanning during shared-bike search. They demonstrate broader system orchestration but are not used as quantitative evidence for the predictive-belief module.

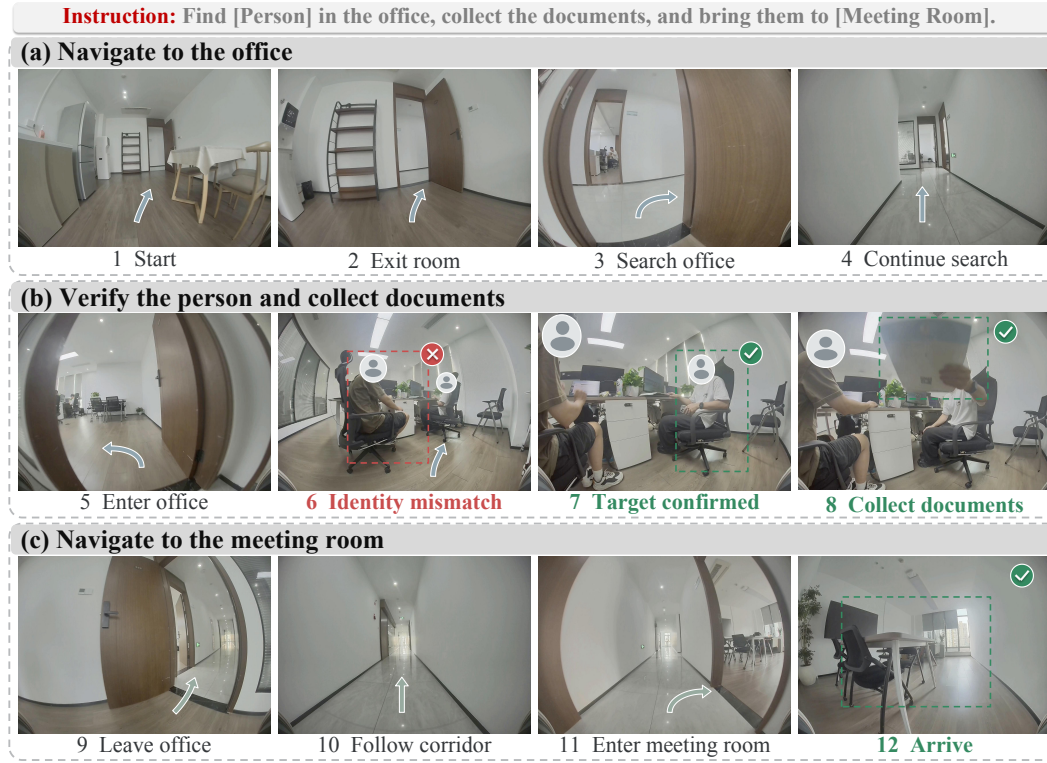


Figure 15: Long-horizon indoor document-delivery execution.

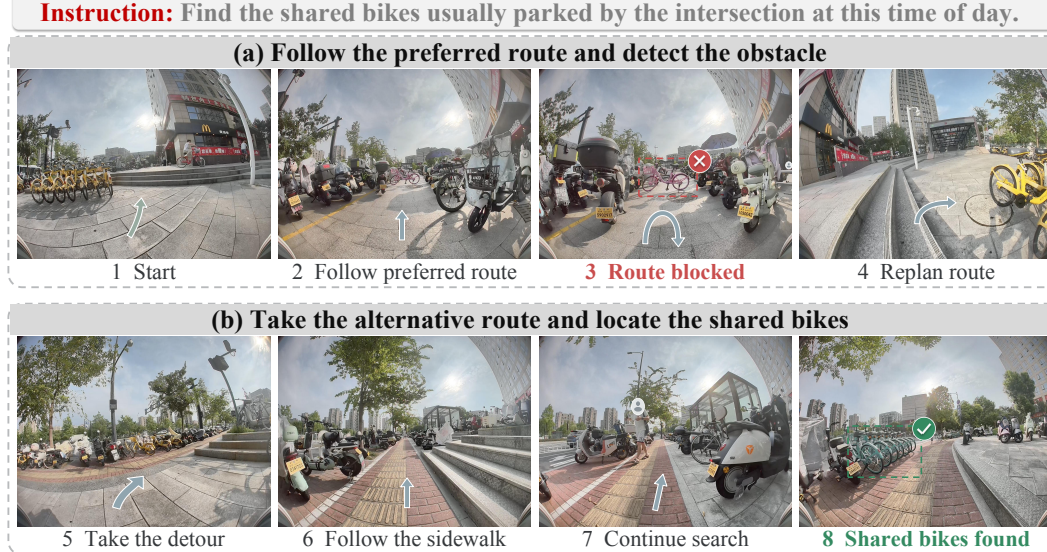


Figure 16: Long-horizon outdoor shared-bike search.

E REPRODUCIBILITY AND RESPONSIBLE AI

Ethics. Persistent visual memory may capture people or private spaces, while an incorrect belief may induce unnecessary motion or unsafe behavior. Deployments should obtain consent, enforce retention and access controls, preserve provenance, and retain platform-specific collision avoidance and emergency stops outside the learned model. Our benchmark compiles existing records into simulated HSSD episodes and introduces no new human-subject study.

1512 **AI use.** Generative AI tools assisted manuscript editing, LaTeX checks, and local engineering.
1513 They did not generate benchmark observations, ground-truth labels, or reported scores. The authors
1514 reviewed all assisted material and remain responsible for the content and artifacts.
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565