RECALLWAM: UNIFYING RECALL AND FORESIGHT FOR OBJECT NAVIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Object navigation requires agents to find specified objects in unfamiliar environments through exploration, using past observations to avoid redundant search. World-action models (WAMs) exploit such visual experience by coupling action learning with future-observation prediction (Foresight), providing dense supervision for navigation. However, identical scene appearances can correspond to different exploration states and call for different search decisions. We therefore extend world-action learning to recover history-dependent exploration state alongside predicting future observations, making past coverage explicit for navigation decisions. To this end, we introduce RecallWAM, which augments foresight-based world-action learning with a recall capability that recovers exploration state from historical observations to guide navigation decisions. To teach the policy how recalled coverage should affect its actions, we construct balanced exploration trajectories that trade off shortest-path guidance with randomized branch exploration, providing history-dependent search experience while limiting unnecessary detours. Experiments on the benchmarks HM3D, MP3D, and OVON demonstrate state-of-the-art performance. Controlled ablations show that joint Recall and Foresight learning yields gains exceeding the sum of their individual gains.

1 INTRODUCTION

Object navigation (ObjectNav) requires agents to find semantically specified targets in unfamiliar environments (Batra et al., 2020). Since the target location is unknown, successful navigation depends not only on recognizing objects but also on deciding where to search (Chaplot et al., 2020). Past observations help distinguish previously observed regions from unexplored alternatives, avoiding redundant search while preserving necessary backtracking. The central challenge is therefore to turn exploration history into information that can guide subsequent decisions.

Existing navigation approaches organize this experience into explicit spatial representations or use visual histories as context for action prediction (Yokoyama et al., 2024a; Zhang et al., 2024a; 2026a). World-action models (WAMs) further exploit visual trajectories by coupling action learning with future-observation prediction, termed Foresight, providing dense visual supervision beyond action labels (Li et al., 2026b; Chen et al., 2025). Memory-augmented world models also demonstrate how historical views can support consistent predictions across viewpoint changes and revisits (Xiao et al., 2026; Wang et al., 2026). These capabilities provide a promising foundation for learning navigation from visual experience.

However, predicting scene appearance does not establish what the agent has already explored. The same view may be encountered before or after searching an adjacent area, calling for different decisions despite unchanged appearance. Future prediction alone does not explicitly resolve this distinction. **We extend world-action learning to recover history-dependent exploration state alongside predicting future observations, making past coverage explicit for present decisions.** To this end, we introduce RecallWAM, which augments foresight-based world-action learning with a recall capability (Figure 1). Recall recovers which traversable regions in the current view have previously been observed. This state serves as both a prediction target and an explicit policy input, rather than leaving its recovery entirely implicit in action prediction from visual history. Recall and Foresight are jointly learned with actions, coupling exploration-state recovery with the visual predictive representations used for navigation.

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

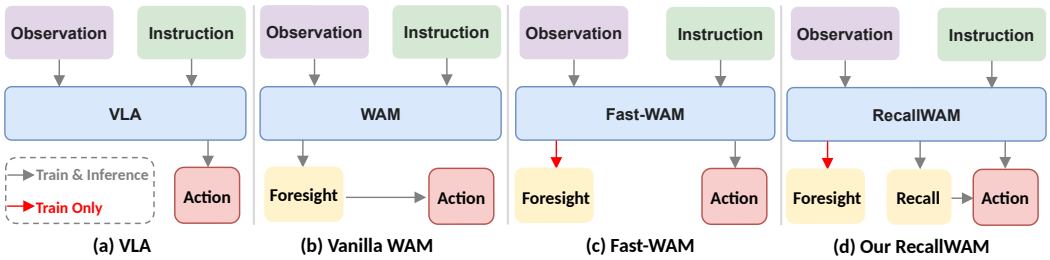

Figure 1: Comparison of navigation learning architectures. (a) VLA directly predicts actions. (b) WAM generates future observations to guide actions. (c) Fast-WAM retains Foresight for training but omits future generation at inference. (d) RecallWAM unifies Recall and Foresight, recovering exploration state as an explicit input to action prediction. Red arrows denote training-only computation; gray arrows denote computation used during both training and inference.

Learning to use Recall also requires demonstrations in which past exploration affects subsequent actions. Shortest-path demonstrations underrepresent such decisions, whereas random exploration can introduce excessive detours. We therefore construct balanced exploration trajectories that balance shortest-path guidance with randomized branch exploration and necessary backtracking. These trajectories expose how previously observed regions influence where to search next, providing experience for learning memory-guided decisions without sacrificing path efficiency to unrestricted exploration.

Our main contributions are:

- We reformulate world-action learning to unify future prediction and exploration-state recovery, making memory a state to recover for decisions rather than merely historical context.
- We propose RecallWAM, which combines temporally recent observations with spatially retrieved history to predict current-view exploration masks and explicitly conditions actions on the recovered state. Our balanced exploration trajectories combine shortest-path guidance with randomized branch exploration to teach efficient memory-guided search.
- A single RecallWAM model achieves state-of-the-art performance on the ObjectNav benchmarks HM3D, MP3D, and OVON. Controlled ablations show joint Recall–Foresight gains exceeding the sum of their individual gains.

2 Related Work

Object Navigation. Modular ObjectNav systems coordinate perception, exploration, and planning through predefined interfaces. Vision-language priors guide target search, but navigation decisions remain mediated by separately designed components (Yokoyama et al., 2024a; Wu et al., 2024; Yin et al., 2024).

Navigation foundation models following the vision-language-action (VLA) paradigm integrate visual-language understanding with learned action prediction (Zhang et al., 2024a; 2026a). They learn policies from visual trajectories and language goals, bringing foundation-model capabilities directly into navigation decisions. However, action supervision alone does not explicitly require modeling how observations evolve with movement.

Camera- and action-conditioned world models learn these visual changes (He et al., 2025; Bar et al., 2025). WAMs couple future-observation prediction (Foresight) with action learning, providing dense supervision beyond action labels (Li et al., 2026b; Chen et al., 2025). Fast-WAM (Yuan et al., 2026) retains Foresight supervision during training but omits future generation at inference, avoiding the latency of iterative video denoising. However, future prediction does not explicitly recover what the agent has already explored. RecallWAM extends this formulation with a Recall capability, jointly learned with Foresight and action prediction, that recovers history-dependent exploration state as an explicit policy input.

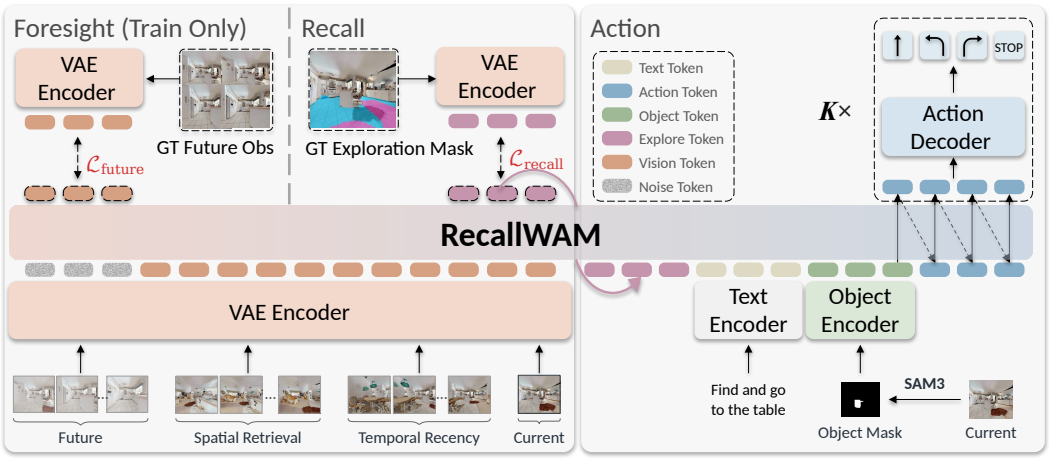

Figure 2: RecallWAM unifies Recall, Foresight, and action learning. Recall recovers current-view exploration state from spatiotemporal visual memory and supplies it to the navigation policy. Foresight provides training-only supervision for the shared world model.

Memory Representations in Navigation. Map-based memories accumulate occupancy or semantic evidence for exploration (Yokoyama et al., 2024a), while graph-based memories represent connectivity and spatial relations for structured search (Wu et al., 2024; Yin et al., 2024). Both provide explicit state for planning, but when they serve as the sole memory, visual evidence omitted during construction cannot be reconsidered later.

Visual-history approaches retain observations or features as navigation context (Fang et al., 2019). To manage redundant history, they use online token merging (Zhang et al., 2024a), context compression (Ren et al., 2025), representation caching (Zeng et al., 2026), and similarity-aware token selection (Xiang et al., 2025). Memory-conditioned world models also use historical evidence: WorldMem (Xiao et al., 2026) and Matrix-Game 3.0 (Wang et al., 2026) retrieve historical views based on camera field-of-view overlap to preserve scene consistency across viewpoint changes and revisits. RecallWAM adapts spatial retrieval to ObjectNav, combining spatially relevant history with the temporally recent window to recover an explicit exploration state for navigation decisions.

3 Method

3.1 Overview

RecallWAM unifies Recall and Foresight to model both what the agent has already explored and what it will see (Figure 2). The design couples three learning goals: recovering prior exploration as explicit state, predicting future observations, and using the learned representations and recovered state to make navigation decisions. We therefore jointly optimize

$$\mathcal{L} = \mathcal{L}_{\text{recall}} + \mathcal{L}_{\text{future}} + \mathcal{L}_{\text{action}}, \tag{1}$$

where $\mathcal{L}_{\text{recall}}$ supervises history-dependent exploration-state recovery, $\mathcal{L}_{\text{future}}$ provides Foresight supervision, and $\mathcal{L}_{\text{action}}$ trains the navigation policy.

Recall and Foresight share a world model whose visual representations are accessed by the policy. Recall additionally provides an explicit exploration-state input, rather than leaving the policy to infer prior coverage from visual features alone. Our implementation initializes the shared world model from Matrix-Game 3.0 (Wang et al., 2026) and uses its frozen VAE. The following subsections define the three objectives and the balanced exploration trajectories used to learn how recalled coverage affects search.

3.2 RECALLING EXPLORATION STATE

Recall recovers current-view exploration state from the current observation, target-category instruction, and spatiotemporal visual memory combining recent context with nonrecent latents retrieved by camera field-of-view overlap. Retrieval greedily maximizes additional current-view coverage under a fixed budget. The shared world model predicts an exploration mask in the frozen VAE's latent space: traversable regions are explored if visible in an earlier observation within 1 m of that observation's camera, and unexplored otherwise; non-traversable regions are invalid. The predicted exploration-state latents directly condition the action branch alongside shared visual features during both training and inference. To learn accurate state recovery that captures newly explored regions and relies on correct historical evidence, we optimize

$$\mathcal{L}_{\text{recall}} = \mathcal{L}_{\text{state}} + \mathcal{L}_{\text{new}} + \mathcal{L}_{\text{rank}}, \tag{2}$$

where the three terms supervise exploration-state reconstruction, newly explored region reconstruction, and history-conditioned error ranking, respectively.

Exploration-state reconstruction. We reconstruct the complete exploration state while upweighting explored regions to prevent their errors from being dominated by the remaining image area:

$$\mathcal{L}_{\text{state}} = \mathbb{E}_t \left[\sum_i \frac{w_{t,i}}{C \sum_j w_{t,j}} \left\| [R_\theta(o_t, \mathcal{H}_t; q_t) - E(e_t)]_i \right\|_2^2 \right], \tag{3}$$

where R_θ is the shared world model's Recall computation, conditioned on the current observation o_t, spatiotemporal visual history $\mathcal{H}_t$, and target-category instruction q_t. The frozen VAE encoder E encodes the ground-truth exploration mask e_t; i, j index latent locations, and C is the number of latent channels. The weight $w_{t,i}$ is 300 at locations corresponding to ground-truth explored regions and 1 elsewhere, including invalid regions.

Newly explored region reconstruction. Full-state reconstruction does not specifically prioritize regions whose labels change as exploration history accumulates. We therefore add a reconstruction term focused on regions that change from unexplored to explored:

$$\mathcal{L}_{\text{new}} = \mathbb{E}_t \left[\frac{1}{C|\Omega_t|} \sum_{i \in \Omega_t} \left\| [R_\theta(o_t, \mathcal{H}_t; q_t) - E(e_t)]_i \right\|_2^2 \right], \tag{4}$$

where Ω_t denotes latent locations corresponding to unexplored-to-explored label changes within the common valid region of a near-viewpoint revisit pair from the same episode. This term is evaluated only when Ω_t is nonempty.

History-conditioned error ranking. Reconstruction alone does not explicitly require correct retrieved history to improve state recovery. We therefore encourage correct history to yield lower reconstruction error than missing or unrelated history, holding the current observation, recent context, and ground-truth exploration mask fixed:

$$\mathcal{L}_{\text{rank}} = \frac{1}{2} \mathbb{E}_t \left[[\alpha d_t^+ - d_t^0]_+ + [\alpha d_t^+ - d_t^-]_+ \right], \tag{5}$$

where d_t^+, d_t^0, and d_t^- are the per-sample mean latent reconstruction errors with correct retrieved history, removed retrieved history, and retrieved history replaced by latents from another episode, respectively. All three errors use the same nonempty region Ω_t and the channel and spatial averaging in Eq. 4. The operator $[x]_+ = \max(x, 0)$ denotes the positive part, and $\alpha = 1.25$ controls the relative margin.

3.3 FORESIGHT AND ACTION LEARNING

Foresight supervises the visual representations shared with Recall and the policy, while Recall additionally supplies explicit exploration state for navigation decisions. We retain standard flow matching on future latents (Li et al., 2026b; Chen et al., 2025):

$$\mathcal{L}_{\text{future}} = \mathbb{E} \Big[w(\sigma) \left\| v_\theta\big(z_{t,\sigma}^+, \sigma \mid c_t\big) - (\epsilon - z_t^+) \right\|_2^2 \Big], \tag{6}$$

where z_t^+ denotes future VAE latents and $z_{t,\sigma}^+ = (1-\sigma)z_t^+ + \sigma\epsilon$ mixes them with Gaussian noise ϵ at noise level $\sigma \in [0,1]$. The velocity predictor v_θ is conditioned on observed visual context and the target-category instruction, collectively denoted by c_t; $w(\sigma)$ weights noise levels.

The policy learns to predict actions from shared visual representations, target-object masks, and Recall-predicted exploration-state latents through action imitation:

$$\mathcal{L}_{\text{action}} = -\mathbb{E}_t\left[\frac{1}{K}\log \pi_{\theta,\phi}\left(\mathbf{a}_t^\star \mid c_t, m_t^{\text{obj}}, \hat{z}_t^e\right)\right], \tag{7}$$

where $\mathbf{a}_t^\star = (a_t^\star, \ldots, a_{t+K-1}^\star)$ is the demonstrated action sequence with $K = 4$, m_t^{obj} is the target-object mask, and $\hat{z}_t^e = R_\theta(o_t, \mathcal{H}_t; q_t)$ denotes the Recall-predicted exploration-state latents that directly condition the policy. Action supervision updates both the shared world-model parameters θ and the policy-specific parameters ϕ. Following Fast-WAM (Yuan et al., 2026), observed-context and action tokens cannot attend to future tokens, allowing Foresight to supervise the shared representations during training without requiring future generation for action prediction. At inference, the policy predicts K actions, executes only the first, and replans after receiving the next observation.

3.4 Balanced Exploration Trajectories

Learning to use Recall requires demonstrations in which prior exploration affects subsequent search. Shortest-path demonstrations provide limited such experience, whereas random exploration can incur excessive detours. We therefore combine shortest-path guidance with randomized branch exploration (Figure 3). Each trajectory supplies exploration-state targets encoding coverage before the current observation, demonstrated actions, and future observations for joint learning.



(a) Shortest Path (b) Randomized Branch (c) Balanced Exploration

Figure 3: Trajectory-generation strategies: (a) shortest-path imitation, (b) randomized branch exploration, and (c) balanced exploration with shortest-path guidance and structured backtracking.

To limit detours without eliminating exploratory choices, we upweight the branch preferred by shortest-path guidance:

$$P(u \mid v) = \frac{1 + (\beta_{\text{sp}} - 1)\mathbf{1}[u = u^\star]}{\beta_{\text{sp}} + |C(v)| - 1}, \tag{8}$$

where $C(v)$ is the nonempty set of undiscovered neighbors of the current node v, and $u \in C(v)$ is a candidate. The preferred candidate $u^\star$ receives weight $\beta_{\text{sp}} \geq 1$, while alternatives receive weight 1; $\mathbf{1}[\cdot]$ denotes the indicator function. The preferred branch minimizes its traversal cost plus the remaining distance to a valid target viewpoint:

$$u^\star \in \arg\min_{u \in C(v)}\left[\ell(v,u) + \min_{q \in Q_g} d_G(u,q)\right], \tag{9}$$

where $\ell(v,u)$ is the branch length, Q_g contains graph anchors of valid target viewpoints, and d_G is the weighted shortest-path distance in the exploration graph G. Setting $\beta_{\text{sp}} = 1$ yields uniform branch sampling; as $\beta_{\text{sp}} \to \infty$, the preferred available branch is always selected.

We apply this rule through depth-first traversal of a Voronoi graph derived from the simulator navigation mesh, backtracking when no undiscovered neighbor remains. Each selected tree edge is traversed at most once per direction during exploration, avoiding repeated search of completed branches. Once the target becomes visible, the demonstrator follows a shortest path to a reachable target viewpoint and stops. Simulator geometry and ground-truth target information are used only for offline data generation.

Table 1: **Performance on object navigation.** Comparison on HM3D Val, MP3D Val, and OVON Val-Unseen benchmarks. Baselines are grouped by their primary memory type. The best and second-best results across all methods are denoted by **bold** and underline, respectively. "–" indicates results that are not reported.

Method	Venue	HM3D Val		MP3D Val		OVON Val-Unseen	
		SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
Graph-based							
PixNav (Cai et al., 2024)	ICRA'24	37.9	20.5	–	–	–	–
SG-Nav (Yin et al., 2024)	NeurIPS'24	54.0	24.9	40.2	16.0	–	–
VoroNav (Wu et al., 2024)	ICML'24	42.0	26.0	–	–	–	–
CogNav (Cao et al., 2025)	ICCV'25	72.5	26.2	46.6	16.1	–	–
UniGoal (Yin et al., 2025)	CVPR'25	54.5	25.1	41.0	16.4	–	–
Map-based							
GAMap (Yuan et al., 2024)	NeurIPS'24	53.1	26.0	–	–	–	–
InstructNav (Long et al., 2024)	CoRL'24	58.0	20.9	–	–	–	–
TriHelper (Zhang et al., 2024b)	IROS'24	56.5	25.3	–	–	–	–
VLFM (Yokoyama et al., 2024a)	ICRA'24	52.5	30.4	36.4	17.5	35.2	19.6
ApexNav (Zhang et al., 2025)	RA-L'25	<u>76.2</u>	38.0	39.2	17.8	–	–
BeliefMapNav (Zhou et al., 2026)	NeurIPS'25	61.4	30.6	37.3	17.6	–	–
MTU3D (Zhu et al., 2025)	ICCV'25	–	–	–	–	40.8	12.1
VLMNav (Goetting et al., 2024)	ICNS'25	50.4	21.0	–	–	–	–
WMNav (Nie et al., 2025)	IROS'25	72.2	33.3	45.4	17.2	–	–
CompassNav (Li et al., 2026a)	ICLR'26	61.6	27.8	42.0	17.5	43.5	21.6
DRIVENav (Gao et al., 2026)	RA-L'26	72.4	<u>41.3</u>	41.8	<u>22.6</u>	50.2	<u>32.6</u>
MerNav (Qi et al., 2026)	arXiv'26	74.8	37.6	50.8	19.5	–	–
OpenFrontier (Padilla-Cerdio et al., 2026)	RSS'26	**77.3**	35.6	40.7	17.8	39.0	20.1
Visual History							
AstraNav-World (Chen et al., 2025)	arXiv'25	–	–	–	–	45.7	28.7
Nav-R1 (Liu et al., 2025)	arXiv'25	–	–	–	–	42.2	20.1
Nav-R^2 (Xiang et al., 2025)	arXiv'25	–	–	–	–	44.0	18.0
Uni-NaVid (Zhang et al., 2024a)	RSS'25	73.7	37.1	–	–	39.5	19.8
ABot-N0 (Chu et al., 2026)	arXiv'26	–	–	–	–	<u>54.0</u>	30.5
JanusVLN (Zeng et al., 2026)	ICLR'26	–	–	–	–	44.9	31.7
NavFoM (Zhang et al., 2026a)	ICLR'26	–	–	–	–	45.2	31.9
OVSegDT (Zemskova et al., 2025)	CVPR'26	–	–	–	–	44.7	20.6
Qwen-RobotNav (Zhang et al., 2026b)	arXiv'26	75.6	30.6	<u>52.2</u>	16.0	53.1	20.9
RecallWAM (Ours)	–	**77.3**	**41.8**	**59.3**	**26.0**	**62.3**	**36.9**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

We evaluate RecallWAM on the HM3D and MP3D validation sets for ObjectNav (Ramakrishnan et al., 2021; Chang et al., 2017) and HM3D-OVON (OVON) for open-vocabulary ObjectNav (Yokoyama et al., 2024b). Success rate (SR) and success weighted by path length (SPL) (Anderson et al., 2018) measure task completion and path efficiency, respectively. Scores are percentages; differences are percentage points.

For the main benchmark results, we jointly train a single RecallWAM model on the training splits of HM3D, MP3D, and OVON. The agent receives single-view RGB images resized to 256×288 and odometry, with target-object masks generated by frozen SAM3 (Carion et al., 2026). All benchmarks share four actions: FORWARD (0.25 m), LEFT and RIGHT (15° turns), and STOP.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Object navigation. RecallWAM achieves state-of-the-art performance on HM3D and MP3D across graph-based, map-based, and visual-history baselines (Table 1). On HM3D, it matches Open-

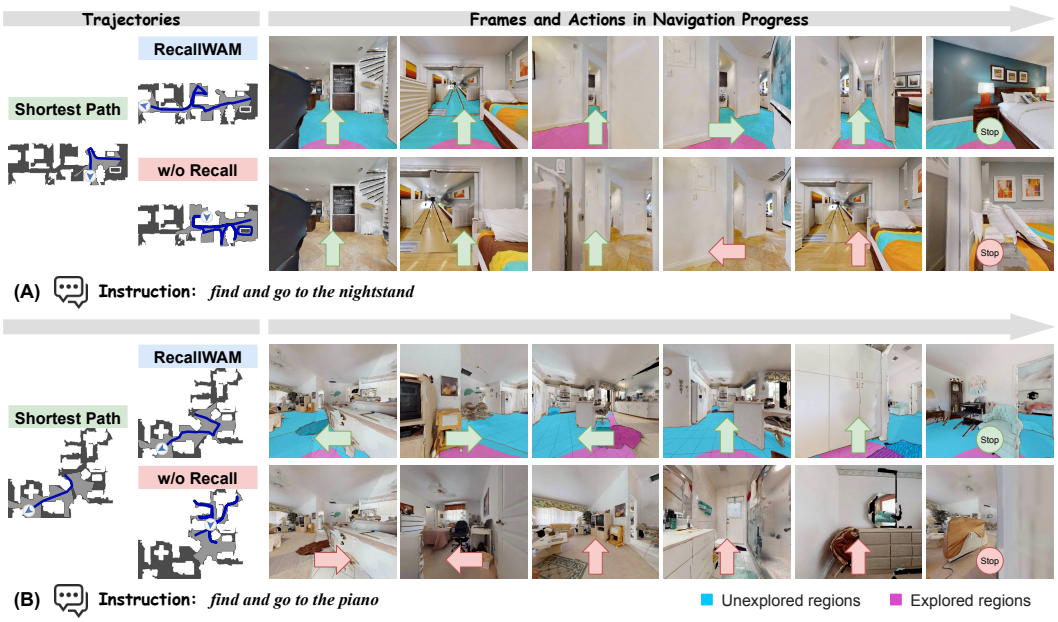

Figure 4: RecallWAM versus w/o Recall on nightstand (A) and piano (B) search. Left: shortest-path references and model trajectories. Right: observations, actions, and Recall-predicted exploration masks. Cyan/magenta denote unexplored/explored traversable regions; green/red arrows indicate correct/incorrect executed actions.

Frontier's 77.3 SR while improving SPL by 6.2 points to 41.8, increasing path efficiency at the same success rate. On MP3D, it reaches 59.3 SR and 26.0 SPL, exceeding the strongest prior results by 7.1 SR points over Qwen-RobotNav and 3.4 SPL points over DRIVENav.

Open-vocabulary object navigation. RecallWAM leads on OVON Val-Unseen with 62.3 SR and 36.9 SPL, exceeding the strongest prior results by 8.3 SR points over ABot-N0 and 4.3 SPL points over DRIVENav. It also surpasses AstraNav-World (Chen et al., 2025), a WAM jointly learning Foresight and actions, by 16.6 SR and 8.2 SPL points. These are full-system comparisons; controlled ablations in Section 4.4 separately examine Recall's contribution and complementarity with Foresight. Complete OVON results appear in Appendix A.1.

4.3 Qualitative Analysis

We examine how recovered exploration state guides actions on OVON Val-Unseen. The w/o Recall variant removes the Recall objective and exploration-state input while retaining the same spatiotemporal history configuration and Foresight–Action learning.

In Figure 4, RecallWAM reaches the nightstand by turning right into an unexplored passage (A) and the piano by turning left away from the explored kitchen side (B). Without Recall, the agent repeatedly traverses the hallway in (A) and revisits rooms in (B), failing both tasks. These examples illustrate how explicit exploration-state recovery turns prior coverage into guidance toward unexplored alternatives, reducing unproductive revisits beyond access to visual history alone.

4.4 Ablation Studies

We examine spatiotemporal visual memory and its explicit exploration-state input, joint Recall–Foresight learning, Recall objectives, and the exploration–efficiency trade-off in balanced exploration trajectories. All variants are trained on OVON only and evaluated on OVON Val-Unseen.

Memory-guided exploration. Table 2 separates two aspects of Recall: assembling historical evidence and providing the recovered state to the policy. Foresight supervision remains enabled

throughout these comparisons. In (a), combining four spatially retrieved latents with four recent temporal latents improves SR by 2.71 points over spatial retrieval alone and 4.34 points over recent context alone. The two sources serve complementary roles: retrieval based on camera field-of-view overlap supplies spatially relevant evidence beyond the recent window, while recent observations preserve local visual continuity.

Because the combined setting also increases context size, (b) compares spatial–temporal allocations under a fixed eight-latent budget. The $(4, 4)$ allocation performs best among the tested configurations, exceeding $(2, 6)$ and $(6, 2)$ by 1.39 and 1.43 SR points, respectively. These results support combining temporal continuity with spatially retrieved evidence within a bounded context, rather than concentrating the budget on either source.

In (c), we fix the $(4, 4)$ configuration and remove only the explicit exploration-state input. Unlike w/o Recall in Section 4.3, this variant retains Recall supervision and visual cross-attention. Nevertheless, SR and SPL decrease by 3.25 and 2.46 points, respectively. The benefit of Recall therefore extends beyond auxiliary supervision and access to historical visual features: explicitly supplying the recovered exploration state improves navigation decisions. This supports our formulation of exploration memory as both a state to recover and an explicit input to navigation decisions.

Table 2: Spatiotemporal visual memory and exploration-state input on OVON Val-Unseen. Spatial and Temporal count nonrecent latents retrieved by camera field-of-view overlap and recent latents, respectively. (a) History composition. (b) Spatial–temporal allocation under an eight-latent budget. (c) Explicit exploration-state input with $(4, 4)$ history. Recall and Foresight supervision remain enabled throughout. Exploration-state input is enabled in (a, b); (c) removes only this input while retaining visual cross-attention.

(a) History composition

Spatial	Temporal	SR↑	SPL↑
0	4	52.85	32.73
4	0	54.48	34.14
4	4	**57.19**	**34.49**

(b) Fixed-budget allocation

Spatial	Temporal	SR↑	SPL↑
2	6	55.80	33.34
4	4	**57.19**	**34.49**
6	2	55.76	33.36

(c) State input

State input	SR↑	SPL↑
×	53.94	32.03
✓	**57.19**	**34.49**

Joint Recall–Foresight learning. Table 3(a) examines whether exploration-state recovery and Foresight provide complementary benefits, with action learning enabled in all variants. Foresight alone changes SR and SPL by -0.20 and $+0.69$ points, respectively, whereas Recall alone improves them by 2.47 and 1.04 points. When combined with Recall, Foresight provides an additional 2.57 SR and 2.30 SPL points, reaching 57.19 SR and 34.49 SPL. Joint learning therefore improves on the action-only baseline by 5.04 SR and 3.34 SPL points, exceeding the sum of the individual gains by 2.77 and 1.61 points, respectively. This empirical complementarity supports extending world-action learning to recover prior exploration alongside predicting future observations, rather than relying on Foresight alone for history-dependent search.

Recall objectives. Beyond full-state reconstruction ($\mathcal{L}_{\text{state}}$), adding $\mathcal{L}_{\text{new}}$ and then $\mathcal{L}_{\text{rank}}$ raises exploration mask IoU from 62.61 to 71.97 and 76.10, respectively, on an exploration-state set from OVON Val-Unseen (Table 3(b)). These gains support supervising exploration changes and their historical evidence for more accurate history-dependent state recovery.

Balanced exploration trajectories. We vary the shortest-path preference β_{sp} to examine the trade-off between path efficiency and history-dependent search experience in Section 3.4. At $\beta_{\text{sp}} = 1$, candidate branches are sampled uniformly; as $\beta_{\text{sp}} \to \infty$, the preferred available branch is always selected within the structured traversal.

Table 3(c) shows that $\beta_{\text{sp}} = 12$ performs best among the tested settings, achieving 57.19 SR and 34.49 SPL. Increasing the preference to 16 reduces performance to 49.96 SR and 30.04 SPL, while a value of 32 yields 48.57 SR and 31.38 SPL. Stronger shortest-path guidance therefore does not necessarily produce better navigation policies. These results support balancing path efficiency with randomized branch exploration, consistent with the need to preserve demonstrations in which prior coverage affects subsequent search rather than concentrating training on target-directed progress.

Table 3: Joint Recall–Foresight learning, Recall objectives and balanced exploration trajectories on OVON Val-Unseen. (a) Contributions of Recall and Foresight; disabling Recall removes both its objective and the exploration-state input. (b) Recall loss ablation, reporting exploration mask IoU (%). (c) Shortest-path preference β_{sp} in balanced exploration trajectories; larger values favor the preferred branch over randomized alternatives.

(a) Recall–Foresight learning

Recall	Foresight	SR↑	SPL↑
×	×	52.15	31.15
×	✓	51.95	31.84
✓	×	54.62	32.19
✓	✓	**57.19**	**34.49**

(b) Exploration mask prediction

$\mathcal{L}_{\text{state}}$	$\mathcal{L}_{\text{new}}$	$\mathcal{L}_{\text{rank}}$	IoU (%)↑
✓	×	×	62.61
✓	✓	×	71.97
✓	✓	✓	**76.10**

(c) Shortest-path preference

β_{sp}	SR↑	SPL↑
4	55.45	33.55
8	56.36	33.39
12	**57.19**	**34.49**
16	49.96	30.04
32	48.57	31.38

4.5 Failure Analysis

We analyze unsuccessful episodes from the main-comparison checkpoint on HM3D Val, MP3D Val, and OVON Val-Unseen, categorizing them as simulator-annotation, stopping, perception, or exploration failures (Figure 5). A substantial fraction involve stops at semantically correct target instances missing from simulator annotations. Among the remaining failures, premature stopping is most frequent, followed by false-positive perception errors.

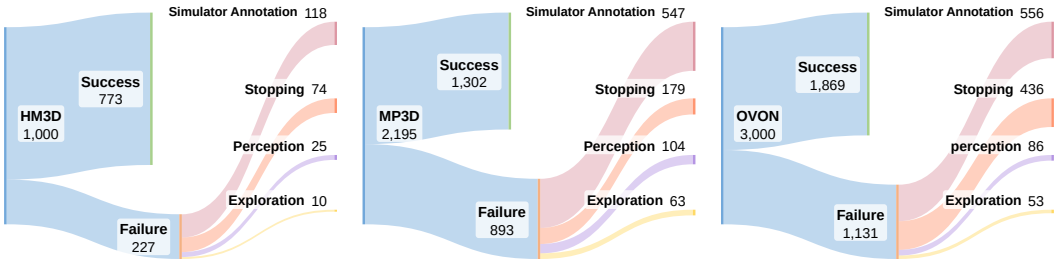

Figure 5: Failure categories on HM3D Val, MP3D Val, and OVON Val-Unseen. Exploration failures denote episodes that exhaust the step budget without encountering the target.

Under this categorization, exploration is the least frequent failure mode: 10/227, 63/893, and 53/1,131 failed episodes on HM3D, MP3D, and OVON, respectively, exhaust the step budget without encountering the target. Together, these account for 126 of 2,251 failures, or 5.6%. Most failures therefore involve annotation, stopping, or perception rather than search-budget exhaustion without target encounter. This pattern is consistent with the qualitative results in Section 4.3, where explicit exploration-state recovery helps redirect search toward unexplored alternatives and avoid unproductive revisits. The comparison between SAM3-predicted and ground-truth target-object masks in Appendix A.2 further examines the role of perception in the remaining performance gap.

5 Conclusion and Future Work

We presented RecallWAM, which unifies Recall and Foresight for ObjectNav. Recall converts spatiotemporal visual memory into an explicit state for navigation decisions, while balanced exploration trajectories teach history-dependent search. A single model achieves state-of-the-art performance on HM3D, MP3D, and OVON, with joint Recall–Foresight gains exceeding the sum of their individual gains.

Our evaluation follows the standard simulated ObjectNav benchmark setting adopted by prior work, with frozen target perception. Remaining perception and stopping failures, together with reliance on odometry for spatial retrieval, motivate stronger target verification, uncertainty-aware stopping, robustness to localization errors, and real-world deployment.

AI USE STATEMENT

Large Language Models (LLMs) were used to assist with language editing and manuscript presentation, including grammar correction, sentence rephrasing, and improving clarity and readability. The technical content, methodological decisions, experimental design, analysis, and conclusions were determined and verified by the authors. The authors take full responsibility for all content presented in the manuscript.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed method, training objectives, trajectory construction, and evaluation protocol in the main paper. The appendix further reports additional experimental results. We evaluate RecallWAM on the standard HM3D, MP3D, and HM3D-OVON benchmarks using established object-navigation metrics and evaluation protocols. These details are provided to facilitate reproduction and verification of our results.

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.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801. IEEE, 2025.

Dhruv Batra, Aaron Gokaslan, Aniruddha Kembhavi, Oleksandr Maksymets, Roozbeh Mottaghi, Manolis Savva, Alexander Toshev, and Erik Wijmans. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171*, 2020.

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

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

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Devendra Singh Chaplot, Dhiraj Prakashchand Gandhi, Abhinav Gupta, and Russ R Salakhutdinov. Object goal navigation using goal-oriented semantic exploration. *Advances in neural information processing systems*, 33:4247–4258, 2020.

Jintao Chen, Junjun Hu, Haochen Bai, Minghua Luo, Xinda Xue, Botao Ren, Chengyu Bai, Shichao Xie, Ziyi Chen, Fei Liu, et al. Astranav-world: World model for foresight control and consistency. *arXiv preprint arXiv:2512.21714*, 2025.

Zedong Chu, Shichao Xie, Xiaolong Wu, Yanfen Shen, Minghua Luo, Zhengbo Wang, Fei Liu, Xiaoxu Leng, Junjun Hu, Mingyang Yin, et al. Abot-n0: Technical report on the vla foundation model for versatile embodied navigation. *arXiv preprint arXiv:2602.11598*, 2026.

Kuan Fang, Alexander Toshev, Li Fei-Fei, and Silvio Savarese. Scene memory transformer for embodied agents in long-horizon tasks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 538–547. IEEE, 2019.

Maoguo Gao, Zejun Zhu, Zhiming Sun, Zhengwei Ma, Longze Yuan, Zhongjing Ma, Zhigang Gao, Jinhui Zhang, and Suli Zou. Drive-nav: Directional reasoning, inspection, and verification for efficient open-vocabulary navigation. *arXiv preprint arXiv:2603.28691*, 2026.

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *arXiv preprint arXiv:2411.05755*, 2024.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *International Conference on Learning Representations*, volume 2025, pp. 100433–100464, 2025.

LinFeng Li, Jian Zhao, Yuan Xie, Xin Tan, and Xuelong Li. Compassnav: Steering from path imitation to decision understanding in navigation. In *International Conference on Learning Representations*, volume 2026, pp. 44744–44768, 2026a.

Yingyan Li, Shuyao Shang, Weisong Liu, Bing Zhan, Haochen Wang, Yuqi Wang, Yuntao Chen, Xiaoman Wang, Yasong An, Chufeng Tang, et al. Drivevla-w0: World models amplify data scaling law in autonomous driving. In *International Conference on Learning Representations*, volume 2026, pp. 7890–7911, 2026b.

Qingxiang Liu, Ting Huang, Zeyu Zhang, and Hao Tang. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*, 2025.

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.

Dujun Nie, Xianda Guo, Yiqun Duan, Ruijun Zhang, and Long Chen. Wmnav: Integrating vision-language models into world models for object goal navigation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2392–2399. IEEE, 2025.

Esteban Padilla-Cerdio, Boyang Sun, Marc Pollefeys, and Hermann Blum. Openfrontier: General navigation with visual-language grounded frontiers. *arXiv preprint arXiv:2603.05377*, 2026.

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

Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021.

Botao Ren, Junjun Hu, Xinda Xue, Minghua Luo, Jintao Chen, Haochen Bai, Liangliang You, and Mu Xu. Astranav-memory: Contexts compression for long memory. *arXiv preprint arXiv:2512.21627*, 2025.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

Pengying Wu, Yao Mu, Bingxian Wu, Yi Hou, Ji Ma, Shanghang Zhang, and Chang Liu. Voronav: Voronoi-based zero-shot object navigation with large language model. *arXiv preprint arXiv:2401.02695*, 2024.

Wentao Xiang, Haokang Zhang, Tianhang Yang, Zedong Chu, Ruihang Chu, Shichao Xie, Yujian Yuan, Jian Sun, Zhining Gu, Junjie Wang, et al. Nav-R^2 dual-relation reasoning for generalizable open-vocabulary object-goal navigation. *arXiv preprint arXiv:2512.02400*, 2025.

Zeqi Xiao, Yushi Lan, Yifan Zhou, Wenqi Ouyang, Shuai Yang, Yanhong Zeng, and Xingang Pan. Worldmem: Long-term consistent world simulation with memory. *Advances in Neural Information Processing Systems*, 38:49632–49652, 2026.

Hang Yin, Xiuwei Xu, Zhenyu Wu, Jie Zhou, and Jiwen Lu. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems*, 37:5285–5307, 2024.

Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Unigoal: Towards universal zero-shot goal-oriented navigation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19057–19066. IEEE, 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, 2024a.

Naoki Yokoyama, Ram Ramrakhya, Abhishek Das, Dhruv Batra, and Sehoon Ha. Hm3d-ovon: A dataset and benchmark for open-vocabulary object goal navigation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5543–5550. IEEE, 2024b.

Shuaihang Yuan, Hao Huang, Yu Hao, Congcong Wen, Anthony Tzes, and Yi Fang. Gamap: Zero-shot object goal navigation with multi-scale geometric-affordance guidance. *Advances in Neural Information Processing Systems*, 37:39386–39408, 2024.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Tatiana Zemskova, Aleksei Staroverov, Dmitry Yudin, and Aleksandr Panov. Ovsegdt: segmenting transformer for open-vocabulary object goal navigation. *arXiv preprint arXiv:2508.11479*, 2025.

Shuang Zeng, Dekang Qi, Xinyuan Chang, Feng Xiong, Shichao Xie, Xiaolong Wu, Shiyi Liang, Mu Xu, and Xing Wei. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. In *International Conference on Learning Representations*, volume 2026, pp. 33001–33026, 2026.

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. *arXiv preprint arXiv:2412.06224*, 2024a.

Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. In *International Conference on Learning Representations*, volume 2026, pp. 127293–127322, 2026a.

Jiazhao Zhang, Gengze Zhou, Hale Yin, Yiyang Huang, Zixing Lei, Qihang Peng, Haoqi Yuan, Jie Zhang, Xudong Guo, Xiaoyue Chen, et al. Qwen-robotnav technical report: A scalable navigation model designed for an agentic navigation system. *arXiv preprint arXiv:2606.18112*, 2026b.

Lingfeng Zhang, Qiang Zhang, Hao Wang, Erjia Xiao, Zixuan Jiang, Honglei Chen, and Renjing Xu. Trihelper: Zero-shot object navigation with dynamic assistance. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 10035–10042. IEEE, 2024b.

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*, 2025.

Zibo Zhou, Yue Hu, Lingkai Zhang, Zonglin Li, and Siheng Chen. Beliefmapnav: 3d voxel-based belief map for zero-shot object navigation. *Advances in Neural Information Processing Systems*, 38:83008–83036, 2026.

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.

A Supplementary Experiments and Analysis

A.1 Complete Results on OVON

Table 4: Complete results on the three OVON validation splits. Baselines are grouped by their primary memory type. Bold and underlining denote the best and second-best reported results, respectively. "–" indicates unreported results.

Method	Venue	Val-Seen		Val-Synonyms		Val-Unseen	
		SR↑	SPL↑	SR↑	SPL↑	SR↑	SPL↑
Map-based							
VLFM (Yokoyama et al., 2024a)	ICRA'24	35.2	18.6	32.4	17.3	35.2	19.6
MTU3D (Zhu et al., 2025)	ICCV'25	55.0	23.6	45.0	14.7	40.8	12.1
CompassNav (Li et al., 2026a)	ICLR'26	–	–	–	–	43.5	21.6
DRIVENav (Gao et al., 2026)	RA-L'26	–	–	–	–	50.2	<u>32.6</u>
OpenFrontier (Padilla-Cerdio et al., 2026)	RSS'26	–	–	–	–	39.0	20.1
Visual History							
AstraNav-World (Chen et al., 2025)	arXiv'25	–	–	–	–	45.7	28.7
Nav-R1 (Liu et al., 2025)	arXiv'25	<u>58.4</u>	26.3	48.1	23.1	42.2	20.1
Nav-R^2 (Xiang et al., 2025)	arXiv'25	45.6	21.0	45.9	21.1	44.0	18.0
Uni-NaVid (Zhang et al., 2024a)	RSS'25	41.3	21.1	43.9	21.8	39.5	19.8
ABot-N0 (Chu et al., 2026)	arXiv'26	55.3	<u>32.1</u>	55.4	<u>33.2</u>	<u>54.0</u>	30.5
JanusVLN (Zeng et al., 2026)	ICLR'26	–	–	–	–	44.9	31.7
NavFoM (Zhang et al., 2026a)	ICLR'26	40.1	27.1	45.4	32.6	45.2	31.9
OVSegDT (Zemskova et al., 2025)	CVPR'26	43.6	20.1	40.1	17.9	44.7	20.6
Qwen-RobotNav (Zhang et al., 2026b)	arXiv'26	57.7	24.4	<u>60.1</u>	25.1	53.1	20.9
RecallWAM (Ours)	–	**64.5**	**38.9**	**61.3**	**36.8**	**62.3**	**36.9**

Table 4 extends the main-comparison model's results to all three OVON validation splits. RecallWAM achieves the highest reported SR and SPL on each split. On Val-Seen, it exceeds the strongest baseline results by 6.1 SR points over Nav-R1 and 6.8 SPL points over ABot-N0. On Val-Synonyms, the corresponding improvements are 1.2 SR points over Qwen-RobotNav and 3.6 SPL points over ABot-N0. On Val-Unseen, RecallWAM improves on ABot-N0 by 8.3 SR points and DRIVENav by 4.3 SPL points. These results extend the main finding across all three splits: the full RecallWAM framework supports both successful and efficient open-vocabulary search. The controlled ablations in Section 4.4 separately examine the contributions of exploration-state recovery and joint Recall–Foresight learning.

A.2 Navigation with Predicted and Ground-Truth Target-Object Masks

We examine how target perception affects memory-guided navigation using RecallWAM trained on OVON only and evaluated on Val-Unseen. The comparison uses SAM3-predicted or ground-truth (GT) target-object masks. Both settings retain Recall-predicted exploration states: GT refers to target-object masks, not exploration masks.

Table 5: Effect of target-object mask quality on OVON Val-Unseen. RecallWAM is trained on OVON only. GT denotes ground-truth target-object masks; exploration states are predicted by Recall in both settings.

Target-object masks	SR↑	SPL↑
SAM3	57.19	34.49
GT	**82.52**	**55.06**

With GT target-object masks, RecallWAM reaches 82.52 SR and 55.06 SPL, improving over SAM3 by 25.33 and 20.57 points, respectively (Table 5). The large gap indicates that target perception remains an important bottleneck even with explicit exploration-state recovery. Recall makes prior

coverage available for deciding where to search, but does not replace reliable evidence for recognizing the target and deciding when to stop. This distinction complements the failure analysis in Section 4.5: improving exploration memory does not eliminate perception-related limitations, and stronger target evidence can substantially improve navigation performance.

B Additional Qualitative Results

B.1 History-Conditioned Exploration-State Recovery

We examine whether Recall uses spatially retrieved evidence beyond recent temporal context to recover prior exploration. In Figure 6, the current RGB observation, recent context, and model checkpoint remain fixed; only the spatially retrieved history is removed or restored. Both predictions are compared against the same ground-truth exploration mask. Recall remains active in both conditions, distinguishing this comparison from the w/o Recall variant in Section 4.3.

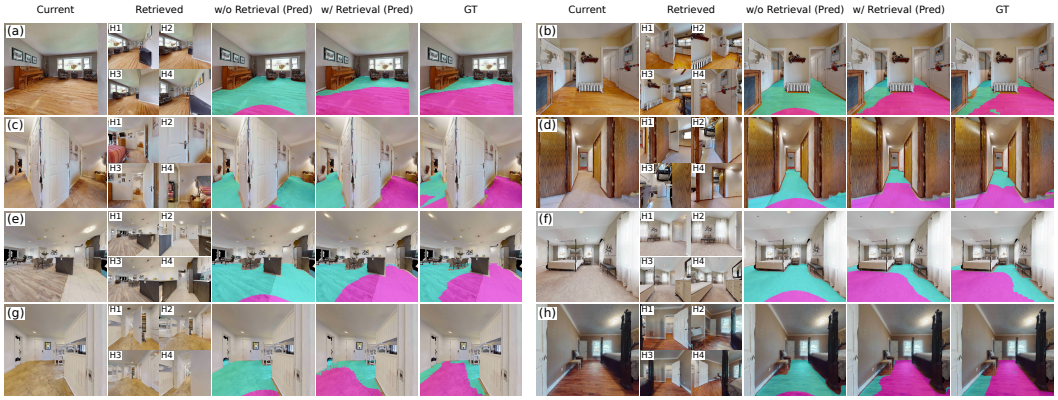

Figure 6: Effect of spatially retrieved history on exploration-state recovery. The current observation, recent context, and checkpoint are fixed. Restoring retrieved evidence helps Recall recover previously observed regions that are missed without it. Cyan/magenta denote unexplored/explored traversable regions.

In the shown examples, removing spatially retrieved history causes previously observed regions to be predicted as unexplored. Restoring this evidence recovers the explored regions and improves agreement with the ground-truth mask despite unchanged current appearance. These examples support the role of spatiotemporal visual memory: spatial retrieval supplies evidence beyond recent context for history-dependent exploration-state recovery. Together with the exploration-state input ablation in Table 2(c), they support the two roles of exploration memory in RecallWAM: a state recovered from historical evidence and an explicit input used for navigation decisions.