Dream to Recall: Imagination-Guided Experience Retrieval for Memory-Persistent Vision-and-Language Navigation

Yunzhe Xu, Yiyuan Pan, Zhe Liu

Abstract—Vision-and-Language Navigation (VLN) requires agents to follow natural language instructions through environments, with memory-persistent variants demanding progressive improvement through accumulated experience. Existing approaches for memory-persistent VLN face critical limitations: they lack effective memory access mechanisms, instead relying on entire memory incorporation or fixed-horizon lookup, and predominantly store only environmental observations while neglecting navigation behavioral patterns that encode valuable decision-making strategies. We present Memoir, which employs imagination as a retrieval mechanism grounded by explicit memory: a world model imagines future navigation states as queries to selectively retrieve relevant environmental observations and behavioral histories. The approach comprises: 1) a language-conditioned world model that imagines future states serving dual purposes: encoding experiences for storage and generating retrieval queries; 2) Hybrid Viewpoint-Level Memory that anchors both observations and behavioral patterns to viewpoints, enabling hybrid retrieval; and 3) an experience-augmented navigation model that integrates retrieved knowledge through specialized encoders. Extensive evaluation across diverse memory-persistent VLN benchmarks with 10 distinctive testing scenarios demonstrates Memoir's effectiveness: significant improvements across all scenarios, with 5.4% SPL gains on IR2R over the best memory-persistent baseline, accompanied by 8.3× training speedup and 74% inference memory reduction. The results validate that predictive retrieval of both environmental and behavioral memories enables more effective navigation, with analysis indicating substantial headroom (73.3% vs 93.4% upper bound) for this imagination-guided paradigm. Code is available at https://github.com/xyz9911/Memoir.

Index Terms—Vision-and-language navigation, embodied intelligence, memory mechanisms, world models.

1 Introduction

Vision-and-Language Navigation (VLN) [1] represents a cornerstone challenge in embodied AI, requiring agents to interpret natural language instructions and navigate through environments to reach specified goals. The fundamental episodic nature of traditional VLN tasks [1], [2], [3], where agents operate independently across episodes without retaining experiential knowledge, limits their capacity for progressive improvement and environmental adaptation, constraining real-world applicability where sustained operation is essential. This limitation has motivated the development of memory-persistent navigation tasks [4], [5], [6] that evaluate agents' ability to accumulate and leverage experience across multiple navigation episodes. These tasks more accurately reflect practical application scenarios where robotic agents must continuously improve their navigation capabilities through environmental familiarity and learned behavioral patterns.

Recent advances in memory-persistent VLN have primarily focused on long-term memory mechanisms for progressive scene knowledge accumulation. Early approaches employed strategies such as episodic history stacking [5], but simply extending history suffers from redundancy-induced performance degradation. Subsequent work [8] addressed this by augmenting visual representations with broader spatial horizons rather than incorporating navigation histories, while OVER-NAV [7] leverages open-vocabulary detection to construct multimodal topological graphs that strengthen keyword-observation correspondence. Most recently, GR-DUET [6] enhanced the DUET architecture [9] with retained topological observation memory, achieving strong performance in VLN scene adaptation.

Despite these advances, existing approaches exhibit two critical limitations. **First**, current approaches lack effective memory access mechanisms, instead relying on either complete memory incorporation (leading to irrelevant information integration and computational overhead) or fixed-horizon spatial lookup (risking valuable experience loss). **Second**, navigation behavioral histories contain valuable decision-making patterns regarding how agents interpreted instructions and selected actions across different scenarios. However, existing memory-persistent VLN methods either ignore them entirely or, when attempted [5], fail to effectively leverage this information.

How can agents effectively determine which memories to access in order to leverage navigation experiences? Human navigators naturally engage in mental imagination of navigation routes [10] and future travel events [11], consulting experiences to finalize decisions based on mental simulations [12], highlighting that imagination serves as a query mechanism—agents can predict where they might navigate and retrieve relevant past experiences matching those predicted states. This paradigm differs from tradi-

• *The authors are with the School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, the Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240. E-mail: xyz9911@sjtu.edu.cn, pyy030406@sjtu.edu.cn, liuzhesjtu@sjtu.edu.cn.*

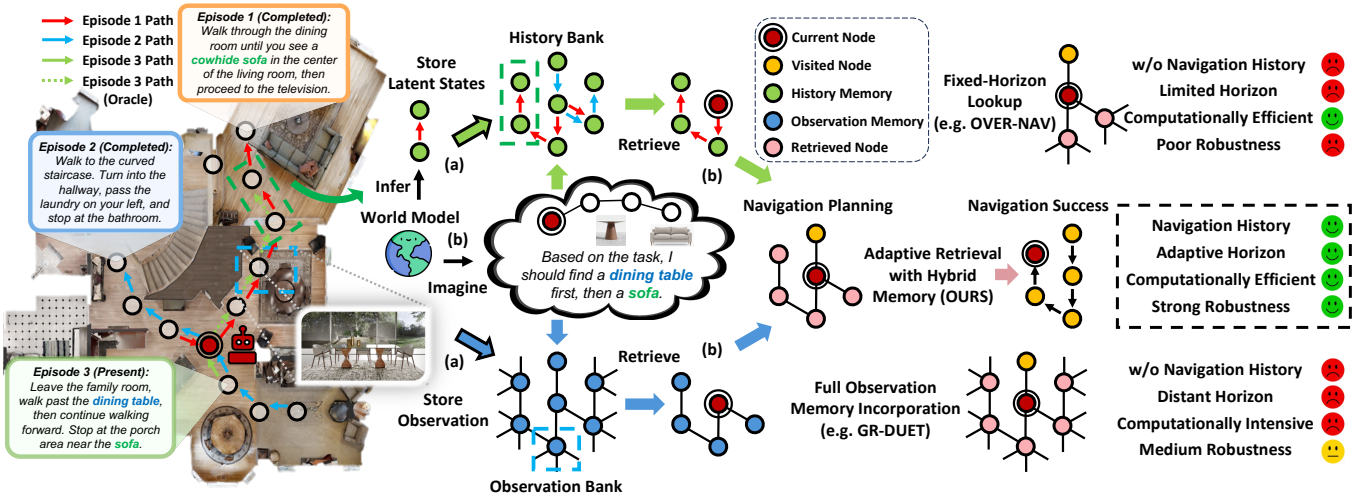


Fig. 1. Overview of Memoir's workflow for experience retrieval via imagination. **(a)** In previous episodes (1 and 2), the agent populates the history bank with latent states encoded by the world model, and fills the observation bank with observations. **(b)** In the current episode (3), the agent utilizes world model imagination to generate retrieval queries and retrieves memory from both memory banks at each viewpoint for navigation planning. Compared with GR-DUET [6] that incorporates all retained observation memory and OVER-NAV [7] that only applies fixed-horizon lookup, our approach adaptively retrieves both observation and histories for navigation planning through imagination.

tional imagine-planning approaches [13] that generate trajectories in isolation; instead, imagination is grounded by querying explicit long-term memory, ensuring retrieved experiences directly inform decision-making while avoiding hallucination. To this end, we propose **M**odel-based Hybrid Vi**e**wpoint-Level **M**em**o**ry for Experi**e**nce **R**etrieval (Memoir), an agent that employs predictive world modeling for memory retrieval at viewpoint granularity. Our approach addresses the aforementioned limitations through a unified framework. First, adaptive retrieval is achieved by using imagined future states as queries, avoiding both complete memory incorporation and fixed-horizon lookup. Second, behavioral pattern preservation is enabled by encoding navigation histories into latent states that capture decision-making strategies with viewpoint-level anchoring. Figure 1 illustrates Memoir's workflow.

Realizing this imagination-guided paradigm requires addressing three key challenges: how to generate predictive queries, what to store and retrieve, and how to integrate retrieved knowledge for navigation. Memoir tackles these through a unified framework. **1)** A language-conditioned world model learns to imagine future navigation states conditioned on instructions. These imagined states serve dual purposes: encoding current navigation experiences into latent representations for storage, and generating queries to retrieve similar past experiences through compatibility matching. **2)** Hybrid Viewpoint-Level Memory (HVM) maintains this accumulated knowledge by anchoring both environmental observations and behavioral patterns to viewpoints, enabling retrieval of not just what agents saw, but how they navigated. **3)** The navigation model then processes current observations alongside retrieved experiences through specialized encoders to make informed decisions. This enables adaptive memory access that preserves strategic knowledge across episodes.

We implement Memoir on various VLN methods, validating its effectiveness across memory-persistent benchmarks with 10 distinctive testing scenarios. Memoir demonstrates consistent improvements, achieving 5.4% improvement in SPL on IR2R [5], accompanied by 8.3× training speedup and 74% inference memory reduction. We also reveal substantial headroom (73.3% vs 93.4% upper bound) for this paradigm and illuminate future directions. Our key contributions include:

- We propose a novel paradigm where imagination serves as a retrieval mechanism grounded by explicit memory, addressing two fundamental limitations in memory-persistent VLN: ineffective memory access via complete incorporation or fixed-horizon lookup, and exclusive focus on environmental observations while neglecting behavioral histories that encode decision-making patterns. This enables adaptive filtering of both environmental and behavioral experiences based on navigation intent.
- We develop Memoir, a unified framework where a language-conditioned world model encodes navigation histories for storage and generates imagined trajectories as retrieval queries, Hybrid Viewpoint-Level Memory (HVM) maintains both observations and behavioral patterns at viewpoint granularity, and an experience-augmented navigation model integrates retrieved information for robust decision-making.
- Through extensive evaluation across 10 diverse memory-persistent scenarios, we demonstrate consistent improvements with significant efficiency benefits, while oracle analysis reveals substantial headroom, illuminating promising directions for advancing this paradigm.

2 Related Work

2.1 Vision-and-Language Navigation

Vision-and-Language Navigation (VLN) [1], [2], [3] requires agents to follow natural language instructions while navigating toward target destinations. Single-episode VLN research has evolved through data augmentation approaches

from speaker models [14], [15], [16] to synthetic data [17], [18] using Large Language Models (LLMs), and memory architectures progressing from historical representations [19], [20] to structured spatial systems, particularly topological observation memory [9], [21] which has been widely adopted. However, these single-episode approaches cannot accumulate knowledge across episodes. Memory-persistent VLN benchmarks [5], [6] address real-world requirements where agents should operate continuously and improve through accumulated experience. TourHAMT [5] extends historical memory by stacking complete navigation sequences, but suffers performance degradation from excessive redundancy. ESceme [8] enhances environmental observations with broader spatial contexts, while OVER-NAV [7] constructs omni-graphs with fixed-distance retrieval. MAP-CMA [5] builds global semantic maps augmented with fixed-horizon egocentric perception. GR-DUET [6] retains complete topological memory, achieving performance gains at computational cost. These approaches share fundamental limitations: reliance on complete memory incorporation or fixed-horizon lookup, and exclusive focus on environmental observations while neglecting navigation behavioral patterns that encode decision-making strategies across contexts. Our work addresses these limitations through imagination-guided memory retrieval that selectively accesses both environmental and behavioral histories.

2.2 Memory Mechanism

Memory mechanisms in navigation systems encompass two primary types that serve complementary roles in spatial reasoning. Navigation history memory captures temporal decision-making patterns and behavioral context through sequence representations [19], [22] or natural language expression [23], [24], preserving how agents make decisions across different scenarios. Environmental observation memory preserves spatial information through structured representations such as occupancy maps [25], [26], semantic maps [27], [28], bird's-eye view representations [29], [30], and topological memory [31], [32], [33], maintaining spatial layouts and visual features for scene understanding. In memory-persistent scenarios where agents must accumulate knowledge across diverse experiences, current approaches [5], [6] treat these information sources separately. Environmental observations alone cannot encode the behavioral reasoning underlying navigation decisions, while navigation histories without spatial anchoring cannot disambiguate similar patterns across different environments. This separation limits knowledge transfer across navigation scenarios, motivating our unified memory that leverage both spatial and temporal historical information.

2.3 World Model

Predictive world models have demonstrated significant impact in model-based reinforcement learning through POMDP solutions via latent dynamics modeling [34], [35]. The Recurrent State-Space Model (RSSM) [34] represents the dominant architecture, with extensions to language conditioning [36] and large-scale pretraining [37]. Contrastive world models [38], [39] offer computational efficiency without observation reconstruction. In navigation domains, world models [40] serve diverse purposes: future observation synthesis for data augmentation [41], [42], trajectory planning through imagination [13], [43], [44], [45], and auxiliary task formulation [46], [47]. However, the integration of world models with memory for grounded imagination remains underexplored. MBEC [48] pioneered imagination-guided memory retrieval in episodic control, but is limited to episodic history during training rather than spatial-temporal retrieval during inference. Our approach extends this paradigm by unifying world modeling with experience retrieval for navigation reasoning, where the world model both encodes histories for storage and generates imagined states for spatial-temporal retrieval.

3 Preliminaries

3.1 VLN Formulation

Vision-and-Language Navigation (VLN) requires an agent to follow instructions and navigate towards target. The environment is represented as a connectivity graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ denotes navigable viewpoints and $\mathcal{E}$ represents traversable edges connecting adjacent viewpoints.

Single-Episode VLN Formulation. In the traditional episodic setting, an agent receives a natural language instruction ℓ and is initialized at a starting viewpoint $v_1 \in \mathcal{V}$. At each timestep t, the agent observes a panoramic observation $o_t = \{o_t^{(i)}\}_{i=1}^{36}$ comprising 36 directional views: 12 horizontal viewing angles, each captured at three elevation levels (upward, horizontal, downward). The agent's action space at viewpoint v_t includes navigation to any neighboring viewpoint $v_j \in \mathcal{N}(v_t)$ and a terminal stop action, where $\mathcal{N}(v_t) = \{v_j \in \mathcal{V} : (v_t, v_j) \in \mathcal{E}\}$ denotes the set of adjacent viewpoints. The episode terminates when the agent executes a stop action or reaches a maximum step limit $T_{\max}$.

Memory-Persistent VLN Formulation. While traditional VLN effectively evaluates basic instruction-following capabilities, it fails to capture the requirements of progressive improvement during persistent operation. Memory-persistent VLN addresses this limitation by introducing a persistent memory bank $\mathcal{M} = \{(\ell^{(k)}, \mathcal{G}^{(k)}, \mathcal{O}^{(k)}, \mathcal{A}^{(k)})\}_{k=1}^{N}$ that accumulates experiential knowledge across multiple episodes, where for k-th episode, $\ell^{(k)}$ is the instruction, $\mathcal{G}^{(k)} = (\mathcal{V}^{(k)}, \mathcal{E}^{(k)})$ is the observed subgraph after k episodes, $\mathcal{O}^{(k)} = \{o_t^{(k)}\}_{t=1}^{T^{(k)}}$ and $\mathcal{A}^{(k)} = \{a_t^{(k)}\}_{t=1}^{T^{(k)}}$ records observations and actions respectively. The bank $\mathcal{M}$ is incrementally updated in each episode and serves as a persistent repository for decisions, enabling progressive performance improvement through accumulated environmental familiarity and learned behavioral patterns.

3.2 Dual-Scale Graph Transformer (DUET)

DUET [9] enables topological navigation through topological mapping and global action planning.

Topological Mapping. The agent maintains an incrementally constructed topological representation $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ of the explored environment, where $\mathcal{G}_t \subseteq \mathcal{G}$ represents the observed subset after t navigation steps. The viewpoint set $\mathcal{V}_t$ is partitioned into three categories: visited viewpoints, frontier viewpoints (observable but unvisited neighbors), and the current viewpoint. At each timestep

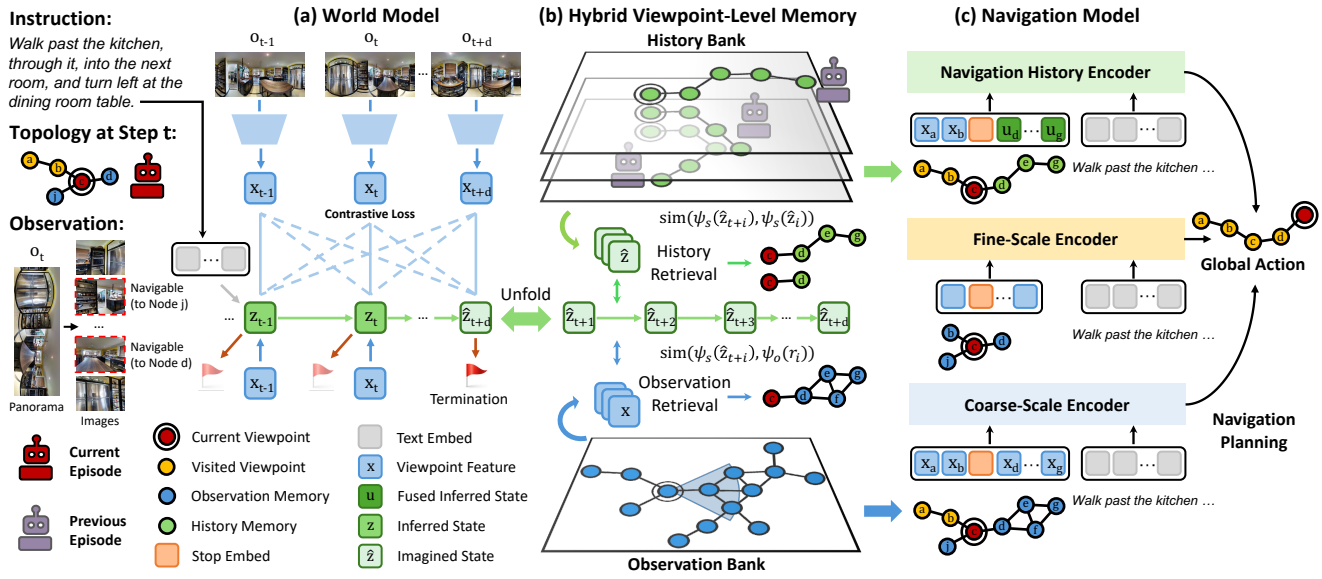


Fig. 2. Details of imagination-guided experience retrieval. **(a)** The world model learns state-observation compatibility through contrastive training (top). During navigation, it infers the current state from observations and instruction, then recursively imagines future states (bottom). **(b)** Imagined trajectories enable dual retrieval: histories via state sequence similarity matching, and observations via topological searching based on state-observation compatibility. **(c)** Three specialized encoders process retrieved navigation histories, local observations, and retrieved observations respectively to determine the final action.

t, the topological graph is updated by incorporating the current viewpoint v_t and its navigable neighbors $\mathcal{N}(v_t)$ into $\mathcal{V}_{t-1}$, with corresponding edge updates to $\mathcal{E}_{t-1}$. Visual representations $r_t = \{r_t^{(i)}\}_{i=1}^{36}$ are computed through an observation encoder applied to o_t. The visual representation of the current viewpoint x_t is obtained via average pooling of r_t, while each unvisited neighboring viewpoint $v_j \in \mathcal{N}(v_t)$ is represented by its corresponding directional embedding $r_t^{(i_j)}$ where i_j denotes the view index oriented toward v_j. For viewpoints observed from multiple locations, embeddings are averaged to maintain consistency.

Global Action Planning. DUET combines coarse-scale planning over the topological graph with fine-scale planning over immediate neighbors. The instruction ℓ is processed through a transformer to obtain textual representations $\hat{\ell}$. For coarse-scale planning, node representations x_j for viewpoints $v_j \in \mathcal{V}_t$ are augmented with a special stop token x_0. The coarse-scale encoder processes the instruction embedding $\hat{\ell}$ and viewpoint representations $X = [x_0, x_1, \ldots, x_{|\mathcal{V}_t|}]$ through cross-modal attention and Graph-Aware Self-Attention (GASA):

$$\text{GASA}(X) = \text{Softmax}\left(\frac{XW_q(XW_k)^T}{\sqrt{d}} + M\right)XW_v, \quad (1)$$

where the distance encoding matrix $M = EW_e + b_e$ incorporates the pairwise distance matrix E. For fine-scale planning, the fine-scale encoder processes the instruction $\hat{\ell}$ and panoramic features r_t to generate action scores for immediate neighbors $\mathcal{N}(v_t)$. The final navigation decision combines both scales through learned dynamic weighting, producing action scores for each candidate viewpoint.

3.3 Contrastive Variational World Model

World models provide latent representations of environment dynamics, enabling efficient inference about future states. Given an observation sequence $(o_1, o_2, \ldots, o_T)$, the world model operates on latent states z_t that capture environmental dynamics. The joint distribution factorizes as:

$$p(o, z) = \prod_{t=1}^{T} p(z_t \mid z_{t-1}) p(o_t \mid z_t). \quad (2)$$

To maximize the observation likelihood $p(o_{1:T})$, the model introduce a variational posterior $q(z_{1:T}|o_{1:T})$ and derive the evidence lower bound (ELBO) [49]:

$$\begin{aligned} \ln p(o) \geq \sum_{t=1}^{T} \Big(& \mathbb{E}_{q(z_t|o_{\leq t})}[\underbrace{\ln p(o_t \mid z_t)}_{\mathcal{J}_{\text{RECOVER}}}] \\ & - \mathbb{E}_{q(z_{t-1}|o_{\leq t})}[\underbrace{\text{KL}[q(z_t \mid o_{\leq t}) \parallel p(z_t \mid z_{t-1})]}_{\mathcal{J}_{\text{KL}}}] \Big). \end{aligned} \quad (3)$$

To empower the model with discriminative power while avoiding pixel-level reconstruction, the term $\mathcal{J}_{\text{RECOVER}}$ is replaced with a contrastive objective [35], [39]. Following the information-theoretic derivation, we can lower-bound $\mathcal{J}_{\text{RECOVER}}$ using noise-contrastive estimation (NCE) [50]:

$$\begin{aligned} \mathcal{J}_{\text{RECOVER}} & \geq \mathbb{E}_{q(z_t|\cdot)}\left[\ln p(z_t \mid o_t) - \ln \sum_{o' \in \mathcal{D}} p(z_t \mid o')\right] \\ & = \mathcal{J}_{\text{NCE}}, \end{aligned} \quad (4)$$

where $\mathcal{D}$ represents a mini-batch of negative samples. This contrastive objective trains the model to distinguish between correct state-observation pairs (z_t, o_t) and incorrect pairs (z_t, o'), effectively learning representations that capture environmental detail without explicit reconstruction.

4 MEMOIR

This section presents Memoir, a memory-persistent VLN agent that employs world model imagination for adaptive

TABLE 1
Key notation summary.

Notation	Description
$\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$	Episodic graph at step t (viewpoints, edges)
$\mathcal{G}^{(k)} = (\mathcal{V}^{(k)}, \mathcal{E}^{(k)})$	Persistent graph accumulated over k episodes
ℓ, $\hat{\ell}$	Natural language instruction & embedding
$o_t = \{o_t^{(i)}\}_{i=1}^{36}$, $r_t = \{r_t^{(i)}\}_{i=1}^{36}$	Panoramic observation & features (36 views)
$x_t = \text{AvgPool}(r_t)$	Viewpoint feature via average pooling
γ_t, ϵ	Reward signal (distance to goal), stop threshold
z_t, $\hat{z}_t$	Inferred state & imagined state
ψ_s, ψ_o	State & observation embedding functions
$\tau_t = \{\hat{z}_{t+i}\}_{i=1}^{H_t}$	Imagined trajectory with horizon H_t
D	Overshooting dist. and max imagination horizon
$\mathcal{M}_o = (\mathcal{V}_o, \mathcal{X}_o)$	Observation bank (viewpoints, features)
$\mathcal{M}_h = (\mathcal{V}_h, \mathcal{Z}_h, \mathcal{T}_h)$	History bank (viewpoints, states, trajectories)
$c_{i,j}$, c_i	Compatibility scores for retrieval
W, P	Max width for obs. & max patterns for history
ρ_o, γ_o	Filter rate & decay factor for obs. retrieval
θ_h, γ_h	Base threshold & decay factor for history retrieval
σ_c, σ_f, σ_h	Fusion weights for navigation model encoders
$s_j^{(c)}$, $s_j^{(f)}$, $s_j^{(h)}$	Action scores (coarse, fine, history)

experience retrieval. As illustrated in Figure 2, our approach comprises three components: a language-conditioned contrastive world model that encodes histories and imagines future states as retrieval queries, a Hybrid Viewpoint-Level Memory (HVM) that stores both environmental observations and navigation histories for retrieval, and an experience-augmented navigation model integrating retrieved knowledge for navigation planning. To facilitate reading, we list the crucial notations in Memoir in Table 1.

Algorithm 1: Environmental Observation Retrieval

Input: W Max Width; ρ_o Filter Rate; γ_o Decay Factor; $\mathcal{G}^{(k)}$ Persistent Graph; v_t Current Viewpoint; τ_t Imagined States; $\mathcal{M}_o$ Observation Bank; $\mathcal{G}_t$ Episodic Graph

Initialize retrieval set $\mathcal{R} \leftarrow \emptyset$
for $i \leftarrow 1$ **to** $|\tau_t|$ **do**
  Initialize $\mathcal{R}_{\text{tmp}} \leftarrow \emptyset$
  Get i-th order neighbors $\mathcal{N}_i(v_t)$ from $\mathcal{G}^{(k)}$
  for *each viewpoint* $v_n \in \mathcal{N}_i(v_t)$ **do**
    Extract imagined state $\hat{z}_{t+i}$ from τ_t
    Compute compatibility $c_{i,n}$ via Equation (10)
    Add $(v_n, c_{i,n})$ to $\mathcal{R}_{\text{tmp}}$
  Sort $\mathcal{R}_{\text{tmp}}$ by score $c_{i,n}$ in descending order
  Retain top $(1 - \rho_o \cdot \gamma_o^{i-1})$ fraction of $\mathcal{R}_{\text{tmp}}$
  Keep top W nodes in $\mathcal{R}_{\text{tmp}}$
  for *each* $(v_n, c_{i,n}) \in \mathcal{R}_{tmp}$ **do**
    Find shortest path $P_{t,n}$ from v_t to v_n in $\mathcal{G}^{(k)}$
    Add path viewpoints: $\mathcal{R} \leftarrow \mathcal{R} \cup P_{t,n}$
for *each viewpoint* $v_n \in \mathcal{R}$ **do**
  Retrieve feature x_n from $\mathcal{M}_o$ for viewpoint v_n
  Retrieve edges E_n from $\mathcal{G}^{(k)}$ for v_n
  Update episodic graph: $\mathcal{G}_t$.update(v_n, E_n)
  Store feature x_n for viewpoint v_n
return updated episodic graph $\mathcal{G}_t$

4.1 Language-Conditioned World Model

To adapt the basic contrastive world model that focus solely on environmental dynamics [35] for VLN task, our approach explicitly incorporates instruction conditioning to leverage the strong prior knowledge inherent in VLN tasks. We extend the standard ELBO formulation by incorporating instruction ℓ and reward signal γ_t (indicating distance to goal):

$$\begin{aligned}\ln p(o, \gamma \mid \ell) \geq \sum_{t=1}^{T} \Big(& \mathbb{E}_{q(z_t \mid o_{\leq t}, \ell)} [\underbrace{\ln p(\gamma_t \mid z_t)}_{\mathcal{J}_{\text{REWARD}}}] \\ & + \underbrace{\mathbb{E}_{q(z_t \mid o_{\leq t}, \ell)} [\ln p(z_t \mid o_t) - \ln \sum_{o' \in \mathcal{D}} p(z_t \mid o')]}_{\mathcal{J}_{\text{NCE}}} \\ & - \mathbb{E}_{q(z_{t-1} \mid o_{\leq t}, \ell)} [\underbrace{\text{KL}[q(z_t \mid o_{\leq t}, \ell) \parallel p(z_t \mid z_{t-1})]}_{\mathcal{J}_{\text{KL}}}] \Big),\end{aligned} \tag{5}$$

where $\mathcal{J}_{\text{REWARD}}$ encourages accurate goal proximity prediction for imagination termination. The negative sample set $\mathcal{D}$ comprises observations from different timesteps and episodes within each training batch. The contrastive term $\mathcal{J}_{\text{NCE}}$ is implemented through a learnable compatibility function that measures semantic similarity between latent states and visual observations:

$$\begin{aligned} f(z_t, o_t) &= \frac{1}{\zeta} \text{sim}(\psi_s(z_t), \psi_o(x_t)) \\ p(z_t \mid o_t) &\propto \exp(f(z_t, o_t)), \end{aligned} \tag{6}$$

where x_t represents the visual feature extracted through DUET's observation encoder via averge pooling, ψ_s and ψ_o are learned embedding functions that map states and observations to a shared embedding space, $\text{sim}(a, b) = \frac{a^\top b}{\|a\| \|b\|}$ denotes cosine similarity, and ζ denotes temperature parameter. This formulation enables principled assessment of compatibility between imagined states and observations stored in long-term memory, providing a foundation for similarity-based memory retrieval.

To improve the model's long-horizon predictive capability and enhance memory retrieval quality, we extend the ELBO formulation with multi-step overshooting. The d-step overshooting objective encourages accurate prediction over extended horizons:

$$\begin{aligned}\mathcal{J}^{(d)} = \sum_{t=1}^{T} \Big(& \mathbb{E}_{p(z_t \mid z_{t-d+1}) q(z_{t-d+1} \mid \cdot)} [\underbrace{\ln p(\gamma_t \mid z_t)}_{\mathcal{J}_{\text{STOP}}}] + \\ & \underbrace{\mathbb{E}_{p(z_t \mid z_{t-d+1}) q(z_{t-d+1} \mid \cdot)} [\ln p(z_t \mid o_t) - \ln \sum_{o'} p(z_t \mid o')]}_{\mathcal{J}_{\text{NCE}}} - \\ & \mathbb{E}_{p(z_{t-1} \mid z_{t-d}) q(z_{t-d} \mid \cdot)} [\underbrace{\text{KL}[q(z_t \mid o_{\leq t}, \ell) \parallel p(z_t \mid z_{t-1})]}_{\mathcal{J}_{\text{KL}}}] \Big).\end{aligned} \tag{7}$$

With maximum overshooting distance D, the final optimization objective becomes:

$$\mathcal{J} = \mathcal{J}^{(1)} + \frac{1}{D-1} \sum_{d=2}^{D} \mathcal{J}^{(d)}. \tag{8}$$

To efficiently optimize the objective in Equation (8), we adopt the Recurrent State-Space Model (RSSM) architecture [34], comprising four components:

$$\begin{aligned} &\text{Inference Model:} && z_t \sim q(z_t \mid z_{t-1}, o_t, \ell) \\ &\text{Transition Model:} && \hat{z}_t \sim p(z_t \mid z_{t-1}) \\ &\text{Compatibility Model:} && p(z_t \mid o_t) \propto \exp(f(z_t, o_t)) \\ &\text{Reward Model:} && \hat{\gamma}_t \sim p(\gamma_t \mid z_t), \end{aligned} \tag{9}$$

where in practice the inference model takes x_t as input for observation, and $\hat{\ell}$ as input for instruction. The inference model encodes navigation histories into representations for storage, and the transition model generates imagined future states that facilitate similarity-based memory retrieval.

4.2 Hybrid Viewpoint-Level Memory (HVM)

Having established how our world model imagine and infer states, we now describe how the imagined states query long-term memory. We introduce a dual-bank memory architecture that maintains both environmental observations and navigation behavioral histories at viewpoint granularity. HVM comprises two complementary banks organized around a persistent graph $\mathcal{G}^{(k)} = (\mathcal{V}^{(k)}, \mathcal{E}^{(k)})$ accumulated over k episodes:

- **Observation Bank**: $\mathcal{M}_o = (\mathcal{V}_o, \mathcal{X}_o)$, where $\mathcal{V}_o = \{v_j\}$ represents the set of recorded viewpoints and $\mathcal{X}_o = \{x_j\}_{j=1}^{|\mathcal{V}_o|}$ contains corresponding viewpoint features extracted by DUET's observation encoder.
- **History Bank**: $\mathcal{M}_h = (\mathcal{V}_h, \mathcal{Z}_h, \mathcal{T}_h)$, where $\mathcal{V}_h = \{v_j\}$ denotes viewpoints with recorded navigation histories, $\mathcal{Z}_h = \{\{z_j^{(k)}\}_{k=1}^{N_j}\}_{j=1}^{|\mathcal{V}_h|}$ stores inferred agent states from past episodes, and $\mathcal{T}_h = \{\{\tau_j^{(k)}\}_{k=1}^{N_j}\}_{j=1}^{|\mathcal{V}_h|}$ contains corresponding imagined trajectory sequences, where N_j denotes the number of historical visits to viewpoint v_j.

At each timestep t, both memory banks are updated based on v_t: $\mathcal{M}_o$ receives the viewpoint feature x_t extracted from observation o_t, while $\mathcal{M}_h$ stores the inferred state z_t from the inference model and the imagined trajectory $\tau_t = \{\hat{z}_{t+i}\}_{i=1}^{H_t}$ generated by the transition model in Equation (9). H_t represents the imagination horizon, terminating when the predicted distance $\hat{\gamma}_{t+i}$ falls below threshold ϵ or reaches maximum horizon D.

Environmental Observation Retrieval. Given an imagined trajectory $\tau_t = \{\hat{z}_{t+i}\}_{i=1}^{H_t}$ at viewpoint v_t, we retrieve observations through topology-guided searching via state-observation compatibility. For imagined state $\hat{z}_{t+i}$ and stored feature x_j at viewpoint v_j from $\mathcal{M}_o$, we compute a compatibility score:

$$c_{i,j} = \frac{1}{2}(\text{sim}(\psi_s(\hat{z}_{t+i}), \psi_o(x_j)) + 1). \tag{10}$$

This scoring mechanism directly leverages the contrastive objective from Equation (6), ensuring consistency between training and retrieval. For each imagination step i and corresponding neighborhood order, the algorithm identifies all viewpoints in the i-th order neighborhood $\mathcal{N}_i(v_t) = \{v \in \mathcal{V}^{(k)} : d(v_t, v) = i\}$ and computes compatibility scores using Equation (10). Percentile-based filtering retains the top $(1 - \rho_o \cdot \gamma_o^{i-1})$ fraction of viewpoints ranked by score, followed by selecting the top-W viewpoints from the retained set. Finally, shortest paths from v_t to all selected viewpoints are added to the episodic graph $\mathcal{G}_t$. The complete procedure is detailed in Algorithm 1.

Algorithm 2: Navigation History Retrieval

Input: P Max Patterns; v_t Current Viewpoint; θ_h Threshold; τ_t Imagined States; γ_h Decay Factor; $\mathcal{M}_h$ History Bank; $\mathcal{G}^{(k)}$ Persistent Graph; $\mathcal{G}_t$ Episodic Graph

Retrieve all patterns Q from $\mathcal{M}_h$ for viewpoint v_t
for *each* $(z', \tau') \in Q$ **do**
  Initialize $L \leftarrow \min(|\tau_t|, |\tau'|)$, scores $\mathcal{C} \leftarrow \emptyset$
  for $i \leftarrow 1$ **to** L **do**
    Get imagined state $\hat{z}_{t+i}$, $\hat{z}'_i$ from τ_t, τ'
    Compute compatibility c_i via Equation (11)
    if $c_i < \theta_h \cdot \gamma_h^{i-1}$ **then**
      break
    Append score: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_i\}$
  Store pattern with score: $(z', \tau', \mathcal{C}) \in Q$
Sort Q in descending order (by length and score)
Retain top P patterns as Q
for *each* $(z', \tau', \mathcal{C}) \in Q$ **do**
  Trace subsequent trajectory from $\mathcal{M}_h$: $\{z'_i, v_i\}_{i=1}^{|\mathcal{C}|}$
  for $i \leftarrow 1$ **to** $|\mathcal{C}|$ **do**
    Retrieve feature x_i from $\mathcal{M}_o$ for viewpoint v_i
    Retrieve edges E_i from $\mathcal{G}^{(k)}$ for v_i
    Update episodic graph: $\mathcal{G}_t$.update(v_i, E_i)
    Store state z'_i, score c_i and x_i for viewpoint v_i
return updated episode graph $\mathcal{G}_t$

Navigation History Retrieval. History retrieval identifies stored historical navigation patterns that exhibit similar imagined trajectories to the current agent's imagination. This process leverages the insight that agents with similar future expectations likely share comparable strategies and should benefit from each other's experiences. For a stored trajectory $\tau' = \{\hat{z}'_i\}_{i=1}^{H'}$ at viewpoint v_t from the history bank $\mathcal{M}_h$, we perform sequential similarity matching based on imagined trajectory $\tau_t = \{\hat{z}_{t+i}\}_{i=1}^{H_t}$. The compatibility between imagined states at step i is computed as:

$$c_i = \frac{1}{2}(\text{sim}(\psi_s(\hat{z}_{t+i}), \psi_s(\hat{z}'_i)) + 1). \tag{11}$$

For each imagination step i in trajectory, we continue matching until either reaching the minimum of the two trajectory lengths, or encountering a compatibility score below the step-dependent threshold $\theta_h \cdot \gamma_h^{i-1}$. The compatibility scores up to the matching termination are stored as $\mathcal{C}$.

We rank stored trajectories using a two-stage criterion: matching length (longer matches preferred), and the minimum compatibility score among matched steps. The top-P trajectory patterns are selected. For each selected pattern, we retrieve the subsequent $|\mathcal{C}|$ viewpoints and their associated inferred states $\{z'_i, v_i\}_{i=1}^{|\mathcal{C}|}$, incorporating the corresponding subgraph structure into $\mathcal{G}_t$ along with state representations and compatibility scores. The complete procedure is outlined in Algorithm 2.

4.3 Navigation Model

At each timestep t, the agent imagines future states τ_t and retrieve environmental observation and navigation history according to v_t. The retrieved information is then integrated into the episodic topological graph $\mathcal{G}_t$ maintained by topological mapping. Now, we extend DUET [9] with specialized processing encoders that integrate retrieved experiential knowledge into navigation decisions. Our model processes these retrieved information through dedicated encoders: global observations, local observations and navigation behavioral patterns. The navigation model comprises three branches:

Coarse-Scale Encoder. The coarse-scale encoder incorporates retrieved observations by expanding viewpoint representations X with an additional type—retrieved viewpoints. The full viewpoint representations $X = [x_0, x_1, \ldots, x_{|\mathcal{V}_t|}]$ containing retrieved observations are processed through the coarse-scale encoder for $\hat{X}$. Global action scores are computed as $s_j^{(c)} = \text{FFN}(\hat{x}_j)$ for viewpoint v_j, providing high-level navigation preferences.

Fine-Scale Encoder. The fine-scale encoder processes immediate panoramic feature r_t for $\hat{r}_t$. Local action scores $s_j^{(f)} = \text{FFN}(\hat{r}_t^{(i_j)})$ are computed for each neighbor $v_j \in \mathcal{N}(v_t)$ and converted to the global action space:

$$s_j^{(f')} = \begin{cases} s_{\text{back}}, & \text{if } v_j \in \mathcal{V}_t \setminus \mathcal{N}(v_t) \\ s_j^{(f)}, & \text{otherwise,} \end{cases} \tag{12}$$

where i_j denotes the view index oriented toward v_j and s_{back} aggregates scores for all visited neighbors of viewpoint v_t to encourage backtracking when necessary.

Navigation-History Encoder. The navigation-history encoder processes retrieved behavioral patterns by fusing historical states with current viewpoint representations. The node set $\mathcal{V}_h$ includes all viewpoints processed by this branch, comprising both currently visited locations and nodes retrieved from the history bank. For each viewpoint $v_j \in \mathcal{V}_h$ with retrieved states $Z_j = [z_j'^{(1)}, z_j'^{(2)}, \ldots, z_j'^{(N_j)}]$ and compatibility scores $C_j = [c_j^{(1)}, c_j^{(2)}, \ldots, c_j^{(N_j)}]$, where N_j denotes the number of retrieved states at v_j, we compute:

$$u_j = \left(\text{softmax}\left(\frac{C_j}{\zeta}\right)\right)^\top Z_j + x_j. \tag{13}$$

For visited nodes without retrieved historical states, we simply use the observation $u_j = x_j$. The fused state representations $U = [u_1, u_2, \ldots, u_{|\mathcal{V}_h|}]$ are processed through a transformer to produce history-informed action scores $s_i^{(h)}$, which are then mapped to the global action space:

$$s_i^{(h')} = \begin{cases} s_0, & \text{if } v_i \in \mathcal{V}_t \setminus \mathcal{V}_h \\ s_i^{(h)}, & \text{otherwise.} \end{cases} \tag{14}$$

Dynamic Fusion. We implement a learned dynamic fusion mechanism that automatically balances contributions from the three branches based on current situational factors. The fusion weights are computed through:

$$[\sigma_f, \sigma_c, \sigma_h] = \text{Softmax}(\text{FFN}([\hat{r}_0; \hat{x}_0; \hat{u}_0])), \tag{15}$$

Algorithm 3: Memoir Navigation Loop

Input:

$T_{\max}$	Max Step limit	$\mathcal{M}_o$	Observation Bank
D	Imagination Horizon	$\mathcal{M}_h$	History Bank
ϵ	Stop threshold	$\mathcal{G}^{(k)}$	Persistent Graph

Initialize episodic graph $\mathcal{G}_0 \leftarrow \emptyset$
Receive initial observation o_1, viewpoint v_1
for *step* $t = 1$ **to** $T_{\max}$ **do**
  Update topological graphs $\mathcal{G}_t$ and $\mathcal{G}^{(k)}$
  Infer current state $z_t \sim q(z_t \mid z_{t-1}, o_t, \ell)$
  Initialize imagined trajectory $\tau_t \leftarrow \emptyset$
  for $i = 1$ **to** D **do**
    Imagine next state $\hat{z}_{t+i} \sim p(z_{t+i} \mid z_{t+i-1})$
    Predict reward $\hat{\gamma}_{t+i} \sim p(\gamma_{t+i} \mid z_{t+i})$
    Update trajectory $\tau_t \leftarrow \tau_t \cup \{\hat{z}_{t+i}\}$
    if $\hat{\gamma}_{t+i} < \epsilon$ **then**
      break
  $\mathcal{G}_t \leftarrow \text{ObsRetrieval}(\mathcal{M}_o, v_t, \tau_t, \mathcal{G}_t, \mathcal{G}^{(k)})$
    `// Algorithm 1`
  $\mathcal{G}_t \leftarrow \text{HistoryRetrieval}(\mathcal{M}_h, v_t, \tau_t, \mathcal{G}_t, \mathcal{G}^{(k)})$
    `// Algorithm 2`
  Extract viewpoint feature x_t from o_t
  $\mathcal{M}_o$.add(v_t, x_t)
  $\mathcal{M}_h$.add(v_t, z_t, τ_t)
  Compute score s_j for each candidate node v_j
  Select action $a_t \leftarrow \text{argmax}_j\, s_j$
  if $a_t =$ *stop* **then**
    break
  Receive o_{t+1}, $v_{t+1} \leftarrow$ env.step(a_t)

where $\hat{r}_0$, $\hat{x}_0$, and $\hat{u}_0$ represent the encoded stop token representations from fine-scale, coarse-scale, and navigation-history encoders respectively, and $[;]$ denotes concatenation. The final navigation scores integrate all three branches:

$$s_j = \sigma_f s_j^{(f')} + \sigma_c s_j^{(c)} + \sigma_h s_j^{(h')}. \tag{16}$$

As presented in Algorithm 3, Memoir realizes imagination-guided memory retrieval by generating imagined trajectories as queries to adaptively access relevant observations and behavioral histories from persistent memory. The navigation model integrates retrieved experiences, enabling informed decisions grounded by historical evidence while continuously updating memory banks for progressive improvement across episodes.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate Memoir on two established memory-persistent VLN benchmarks that provide complementary evaluation perspectives. Iterative Room-to-Room (IR2R) [5] extends the foundational Room-to-Room (R2R) dataset [1] to multi-episode scenarios through structured tours, containing 183 training tours with an average length of 76.6 episodes. The validation splits comprise seen environments (159 tours, average 6.4 episodes) and unseen environments

(33 tours, average 71.2 episodes). General Scene Adaptation (GSA-R2R) [6] incorporates 150 Habitat-Matterport3D (HM3D) scenes [51] with 600 paths per scene, providing 90,000 total episodes across 10 evaluation scenarios covering residential and non-residential environments with various instructions including basic navigational commands, scene-specific descriptions, and user-personalized instructions.

Implementation Details. We implement Memoir on three foundational models: DUET [9] and ScaleVLN [16] representing traditional VLN models, and GR-DUET [6] representing the memory-persistent approaches. All models utilize pretrained weights from their respective pretraining phases without task-specific fine-tuning. For rigorous comparison, we retrain all baseline models with identical hyperparameters and experimental conditions, including synchronized episode ordering in GSA-R2R. Our world model implementation employs two architectural variants: GRU and Transformer. Both variants utilize textual embeddings and share the observation encoder with the navigation model. World model pretraining occurs on R2R and augmented trajectories [15] for 5,000 iterations with batch size 32 and learning rate 5e-5, followed by imitation learning at learning rate 1e-5. Results are reported over 3 separate runs.

Evaluation Metrics. We employ standard VLN metrics [52] for navigation performance evaluation. To quantify the effectiveness of long-term memory retrieval, we introduce four complementary metrics that evaluate both observation retrieval and history retrieval quality. The metrics include:

Trajectory Length (TL): predicted path length in meters.

Navigation Error (NE): distance between agent's final position to target in meters.

Success Rate (SR): the percentage of final positions less than 3 meters away from the target location.

Success Rate penalized by Path Length (SPL): SR normalized by the ratio between the length of the shortest path and the predicted path.

Normalized Dynamic Time Warping (nDTW): dynamic time warping normalized between predicted and expert paths.

Tour-normalized Dynamic Time Warping (T-nDTW): the overall navigation consistency across complete tours.

Observation Accuracy (OA): the precision of retrieved observations from the observation bank $\mathcal{M}_o$ across the episode:

$$\text{OA} = \frac{|\bigcup_{t=1}^{T}(\mathcal{R}_t \cap \mathcal{V}^o_{\text{gt},t})|}{|\bigcup_{t=1}^{T}\mathcal{R}_t|}, \tag{17}$$

where $\mathcal{R}_t$ denotes viewpoints retrieved from $\mathcal{M}_o$ at timestep t (as detailed in Algorithm 1), and $\mathcal{V}^o_{\text{gt},t}$ represents the ground truth viewpoints on the teacher trajectory within D steps that exist in the observation bank, where $\mathcal{V}_o$ denotes all viewpoints stored in $\mathcal{M}_o$ and T is episode length.

Observation Recall (OR): the coverage of relevant environmental observations across the episode:

$$\text{OR} = \frac{|\bigcup_{t=1}^{T}(\mathcal{R}_t \cap \mathcal{V}^o_{\text{gt},t})|}{|\bigcup_{t=1}^{T}\mathcal{V}^o_{\text{gt},t}|}. \tag{18}$$

History Accuracy (HA): the precision of retrieved navigation patterns from the history bank $\mathcal{M}_h$ across the episode:

$$\text{HA} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{|Q_t|}|\mathcal{V}^h_{\text{traj},t,j} \cap \mathcal{V}^h_{\text{gt},t,j}|}{\sum_{t=1}^{T}\sum_{j=1}^{|Q_t|}|\mathcal{V}^h_{\text{traj},t,j}|}, \tag{19}$$

where $|Q_t|$ denotes the number of retrieved navigation history patterns at timestep t (as detailed in Algorithm 2), $\mathcal{V}^h_{\text{traj},t,j} = \{v_1^{(j)}, v_2^{(j)}, \ldots, v_{|\mathcal{C}_j|}^{(j)}\}$ represents the sequence of viewpoints in the j-th retrieved navigation history trajectory, and $\mathcal{V}^h_{\text{gt},t,j}$ represents the viewpoints on the teacher trajectory that exist in the original history trajectory.

History Recall (HR): the coverage of relevant navigation patterns across the episode:

$$\text{HR} = \frac{\sum_{t=1}^{T}\sum_{j=1}^{|Q_t|}|\mathcal{V}^h_{\text{traj},t,j} \cap \mathcal{V}^h_{\text{gt},t,j}|}{\sum_{t=1}^{T}\sum_{j=1}^{|Q_t|}|\mathcal{V}^h_{\text{gt},t,j}|}. \tag{20}$$

5.2 Quantitative Analysis

5.2.1 Iterative Room-to-Room (IR2R)

Table 2 presents a comparison of Memoir against both traditional and memory-persistent methods on the IR2R benchmark. When applied to traditional VLN models, Memoir demonstrates substantial performance improvements: 11.1% SPL enhancement for DUET-based implementations and 5.6% for ScaleVLN-based implementations on unseen scenarios. These results prove that incorporating retrieved information from long-term memory serves as an effective prior for robust navigation decisions, even for models not originally designed for memory persistence. When compared against memory-persistent approaches, Memoir significantly outperforms GR-DUET, achieving 5.4% improvement in SPL on unseen scenarios (73.3% versus 67.9%) and 11.6% improvement on seen scenarios. This superior performance validates our hypothesis that incorporating complete memory information introduces excessive noise that degrades navigation decisions and reduces flexibility in scenarios with limited experience availability. Our adaptive retrieval approach effectively addresses these limitations.

Discussion. While achieving exceptional performance on unseen scenarios, memory-persistent variants often exhibit reduced performance on seen scenarios compared to their traditional counterparts. For instance, DUET achieves 74.5% SPL compared to GR-DUET's 55.1% on seen environments, with our method experiencing approximately 2% SPL degradation. This phenomenon stems from: (1) difference in tour lengths between validation splits, seen tours average only 6.4 episodes compared to 71.2 episodes in unseen tours, limiting accumulated experience; (2) regularization effects where long-term memory integration prevents overfitting to training environments by encouraging broader contextual reasoning rather than environmental detail memorization. Memoir substantially reduces this performance gap compared to GR-DUET, demonstrating more balanced memory utilization.

5.2.2 General Scene Adaptation (GSA-R2R)

Tables 3, 4, and 5 present Memoir's performance across diverse scene adaptation scenarios. Table 3 compares against adaptation-based methods and memory-persistent approaches across five distinct user instruction styles. Tables 4 and 5 evaluate performance across different environmental characteristics and instruction expressions. Memoir consistently outperforms both adaptation-based and memory-based methods, achieving an average 2.38% SR increase

TABLE 2
Comparison of navigation performance between Memoir and various VLN methods on the IR2R benchmark.

Methods	PH	TH	PHI	IW	Val Seen TL ↓	NE ↓	nDTW ↑	SR ↑	SPL ↑	t-nDTW ↑	Val Unseen TL ↓	NE ↓	nDTW ↑	SR ↑	SPL ↑	t-nDTW ↑
HAMT [20]					10.1 ±0.1	4.2 ±0.1	71 ±1	63 ±1	61 ±1	58 ±1	**9.4 ±0.1**	4.7 ±0.0	66 ±0	56 ±0	54 ±0	50 ±0
TourHAMT [5]	✓	✓	✓	✓	**9.4 ±0.4**	5.8 ±0.1	59 ±0	45 ±1	43 ±1	45 ±0	10.0 ±0.2	6.2 ±0.1	52 ±0	39 ±1	36 ±0	32 ±1
	✓	✓	✓		10.5 ±0.3	6.0 ±0.2	58 ±1	45 ±2	43 ±2	42 ±1	10.9 ±0.2	6.8 ±0.2	51 ±1	38 ±1	34 ±1	31 ±1
	✓	✓			10.6 ±0.3	6.0 ±0.1	58 ±1	45 ±1	42 ±1	42 ±1	10.3 ±0.3	6.7 ±0.2	50 ±1	38 ±1	34 ±1	29 ±1
	✓				10.9 ±0.3	6.1 ±0.1	58 ±1	45 ±1	42 ±1	41 ±0	11.0 ±0.6	6.7 ±0.1	51 ±0	38 ±0	34 ±0	28 ±1
OVER-NAV [7]					9.9 ±0.1	3.7 ±0.1	73 ±1	65 ±1	63 ±1	62 ±0	**9.4 ±0.1**	4.1 ±0.1	69 ±0	60 ±1	57 ±0	**55 ±1**
Comparison with Traditional VLN Models:																
VLN models pretrained with default protocol:																
DUET [9]					12.5 ±0.4	2.2 ±0.1	79.8 ±1.1	79.8 ±0.7	74.5 ±0.9	69.1 ±1.7	14.4 ±0.1	3.5 ±0.0	65.0 ±0.1	69.2 ±0.3	58.0 ±0.1	47.0 ±0.8
+Memoir (Ours)					11.5 ±0.1	2.6 ±0.2	78.9 ±0.9	77.1 ±0.5	72.8 ±0.5	68.0 ±0.8	**11.0 ±0.0**	**2.8 ±0.1**	**75.2 ±0.0**	**75.4 ±0.2**	**69.1 ±0.3**	**58.8 ±0.4**
VLN models pretrained with environmental augmentation:																
ScaleVLN [16]					12.8 ±0.0	2.2 ±0.0	79.6 ±0.4	79.5 ±0.5	74.1 ±0.6	67.0 ±0.2	13.5 ±0.0	2.7 ±0.0	71.6 ±0.1	76.2 ±0.1	66.5 ±0.2	53.4 ±0.2
+Memoir (Ours)					11.6 ±0.2	2.5 ±0.1	78.7 ±0.1	76.1 ±0.5	72.3 ±0.1	67.1 ±0.0	**10.9 ±0.2**	**2.6 ±0.0**	**77.2 ±0.6**	**77.4 ±0.2**	**72.1 ±0.4**	**62.2 ±0.6**
Comparison with Memory-Persistent VLN Models:																
VLN models pretrained with full navigation graph:																
GR-DUET [6]					12.5 ±0.7	4.3 ±0.2	65.2 ±0.7	61.1 ±1.4	55.1 ±0.1	49.0 ±0.5	11.0 ±0.2	3.1 ±0.0	74.5 ±0.5	72.7 ±0.5	67.9 ±0.1	54.8 ±0.1
+Memoir (Ours)					**11.8 ±0.4**	**3.0 ±0.0**	**74.1 ±0.3**	**72.2 ±0.2**	**66.7 ±0.5**	**61.9 ±0.3**	**10.2 ±0.2**	**2.5 ±0.0**	**79.2 ±0.2**	**77.6 ±0.5**	**73.3 ±0.1**	**66.9 ±0.7**

* PH: previous history integration; TH: trainable history encoder; PHI: previous history identifier; IW: inflection weighting.

TABLE 3
Comparison of navigation performance on the GSA-R2R benchmark with user instructions.

Methods	Child SR ↑	Child SPL ↑	Keith SR ↑	Keith SPL ↑	Moira SR ↑	Moira SPL ↑	Rachel SR ↑	Rachel SPL ↑	Sheldon SR ↑	Sheldon SPL ↑
TourHAMT [5]	14.6 ±0.2	12.0 ±0.2	15.1 ±0.2	12.3 ±0.1	13.9 ±0.1	11.3 ±0.1	15.3 ±0.1	12.5 ±0.1	14.4 ±0.1	11.8 ±0.1
OVER-NAV [7]	20.9 ±0.1	16.1 ±0.2	20.5 ±0.1	16.4 ±0.1	19.5 ±0.2	15.4 ±0.2	20.6 ±0.3	16.2 ±0.2	20.5 ±0.1	16.2 ±0.1
DUET [9]	54.3	44.1	56.0	46.3	52.3	43.3	56.3	46.4	54.0	44.4
+MLM [15]	54.5 ±0.2	44.7 ±0.2	56.4 ±0.3	46.8 ±0.3	53.8 ±0.3	43.6 ±0.4	56.8 ±0.5	46.6 ±0.6	54.5 ±0.4	44.2 ±0.3
+MRC [15]	54.4 ±0.2	44.2 ±0.1	56.0 ±0.1	46.3 ±0.1	52.3 ±0.2	43.3 ±0.1	56.0 ±0.1	46.2 ±0.2	53.7 ±0.2	44.2 ±0.4
+BT [53]	57.5 ±0.7	54.0 ±0.9	61.2 ±0.3	57.9 ±0.1	57.3 ±0.5	54.0 ±0.6	61.6 ±0.8	58.1 ±0.7	57.6 ±0.5	54.3 ±0.5
+TENT [54]	54.3 ±0.2	41.7 ±0.1	55.4 ±0.2	43.8 ±0.2	51.7 ±0.2	41.0 ±0.1	55.0 ±0.2	43.2 ±0.2	53.0 ±0.2	41.9 ±0.1
+SAR [55]	54.5 ±0.5	41.5 ±0.4	54.9 ±0.3	43.1 ±0.2	51.0 ±0.4	40.3 ±0.6	55.3 ±0.5	43.0 ±0.6	52.9 ±0.2	41.4 ±0.4
VLN models pretrained with full navigation graph:										
GR-DUET [6]	65.2 ±0.1	59.7 ±0.1	66.7 ±0.1	62.0 ±0.1	60.9 ±0.2	56.2 ±0.2	67.1 ±0.1	62.2 ±0.1	63.9 ±0.1	58.9 ±0.1
GR-DUET* [6]	64.9 ±0.5	60.5 ±0.4	65.1 ±0.3	61.4 ±0.4	60.5 ±0.3	56.6 ±0.2	65.7 ±0.5	61.7 ±0.4	63.0 ±0.4	59.0 ±0.4
+Memoir (Ours)	**66.5 ±0.5**	**61.3 ±0.5**	**68.0 ±0.1**	**63.6 ±0.2**	**62.5 ±0.3**	**57.5 ±0.4**	**68.2 ±0.1**	**63.6 ±0.3**	**65.3 ±0.1**	**60.4 ±0.3**

* Results reproduced under aligned experimental conditions (episode ordering, training iterations, batch size, learning rate and dropout rate).

TABLE 4
Comparison of navigation performance on the GSA-R2R benchmark with scene instructions.

Methods	Test-N-Scene TL ↓	NE ↓	SR ↑	SPL ↑	nDTW ↑
TourHAMT [5]	7.3 ±0.1	8.1 ±0.1	9.7 ±0.1	8.0 ±0.1	32.3 ±0.1
OVER-NAV [7]	11.8 ±0.1	7.6 ±0.2	16.7 ±0.4	12.6 ±0.2	34.6 ±0.3
DUET [9]	14.9	6.4	39.6	30.1	40.9
+MLM [15]	14.3 ±0.1	6.5 ±0.1	39.8 ±0.1	30.5 ±0.1	41.1 ±0.1
+MRC [15]	14.9 ±0.1	6.4 ±0.1	39.7 ±0.1	30.2 ±0.1	40.9 ±0.1
+BT [53]	8.4 ±0.0	6.3 ±0.2	41.2 ±1.5	38.2 ±1.2	51.3 ±1.2
+TENT [54]	16.4 ±0.1	6.3 ±0.1	40.6 ±0.2	28.9 ±0.2	38.9 ±0.2
+SAR [55]	16.3 ±0.5	6.0 ±0.2	41.4 ±0.6	29.1 ±0.3	39.0 ±0.3
VLN models pretrained with full navigation graph:					
GR-DUET [6]	10.1 ±0.0	5.5 ±0.0	48.1 ±0.1	42.8 ±0.1	53.7 ±0.1
GR-DUET* [6]	**9.9 ±0.3**	5.5 ±0.0	47.1 ±0.5	42.2 ±0.8	54.1 ±0.6
+Memoir (Ours)	10.3 ±0.4	**5.1 ±0.0**	**50.2 ±0.3**	**44.8 ±0.4**	**56.2 ±0.6**

* Results reproduced under aligned experimental conditions.

and 1.59% SPL improvement compared to GR-DUET across eight distinct testing scenarios with aligned experimental configurations. The improvements demonstrate that hybrid memory provides critical context absent in traditional approaches: by accessing past episodes where agents successfully processed similar expressions and executed corresponding actions, Memoir learns from historical patterns that GR-DUET's observation memory cannot capture.

Discussion. Memoir consistently outperforms GR-DUET, though with smaller gains than against IR2R. This reduced improvement stems from memory density differences, with GSA-R2R accumulating 600 episodes on average, increasing the topological completeness for GR-DUET.

5.3 Qualitative Analysis.

Figure 3 demonstrates memory retrieval effectiveness in challenging scenarios where DUET and GR-DUET fail. Given the task of locating a "massage table" with two potential candidates barely visible from the hallway, the DUET agent incorrectly approaches the wrong target without observing the actual target, while the GR-DUET agent becomes confused among numerous candidate locations and produces incorrect decisions. Our model succeeds through

TABLE 5
Comparison of navigation performance on the GSA-R2R benchmark with basic instructions.

Methods	Test-R-Basic					Test-N-Basic				
	TL ↓	NE ↓	SR ↑	SPL ↑	nDTW ↑	TL ↓	NE ↓	SR ↑	SPL ↑	nDTW ↑
TourHAMT [5]	11.6 ±0.1	7.4 ±0.1	14.9 ±0.1	12.2 ±0.1	34.7 ±0.1	9.4 ±0.1	7.7 ±0.1	11.0 ±0.2	8.6 ±0.2	32.2 ±0.1
OVER-NAV [7]	14.1 ±0.1	6.7 ±0.0	22.3 ±0.3	16.8 ±0.2	37.1 ±0.1	11.4 ±0.1	7.1 ±0.1	16.6 ±0.2	13.0 ±0.1	35.0 ±0.2
DUET [9]	13.1	4.2	57.7	47.0	55.6	14.8	5.3	48.1	37.3	45.9
+MLM [15]	13.1 ±0.1	4.1 ±0.1	57.9 ±0.2	47.3 ±0.1	55.9 ±0.2	13.1 ±0.2	5.3 ±0.1	48.3 ±0.5	38.8 ±0.5	48.4 ±0.3
+MRC [15]	13.1 ±0.1	4.2 ±0.1	57.7 ±0.1	47.0 ±0.1	55.6 ±0.1	14.7 ±0.1	5.3 ±0.1	48.1 ±0.1	37.3 ±0.1	45.9 ±0.1
+BT [53]	8.0 ±0.1	3.8 ±0.1	61.3 ±0.6	57.7 ±0.3	70.1 ±0.5	7.9 ±0.0	5.2 ±0.1	49.5 ±0.8	46.0 ±0.8	59.4 ±0.9
+TENT [54]	14.6 ±0.0	4.2 ±0.0	57.2 ±0.4	44.2 ±0.4	52.9 ±0.1	16.2 ±0.1	5.4 ±0.1	46.5 ±0.4	33.7 ±0.2	42.6 ±0.3
+SAR [55]	13.8 ±0.8	4.0 ±0.1	57.6 ±0.2	44.6 ±0.2	53.0 ±0.2	16.5 ±0.0	5.4 ±0.0	44.6 ±1.5	31.5 ±1.6	40.6 ±1.3
VLN models pretrained with full navigation graph:										
GR-DUET [6]	9.4 ±0.0	3.1 ±0.0	69.3 ±0.2	64.3 ±0.1	71.4 ±0.1	8.9 ±0.0	4.4 ±0.0	56.6 ±0.1	51.5 ±0.1	61.0 ±0.1
GR-DUET* [6]	**8.6 ±0.2**	3.2 ±0.1	67.6 ±0.5	63.6 ±0.6	71.9 ±0.5	**8.7 ±0.4**	4.4 ±0.0	55.3 ±0.2	50.4 ±0.3	60.8 ±0.4
+Memoir (Ours)	9.3 ±0.0	**3.0 ±0.0**	**69.8 ±0.2**	**64.9 ±0.4**	**73.3 ±0.2**	9.3 ±0.2	**4.2 ±0.0**	**57.7 ±0.1**	**52.0 ±0.1**	**61.9 ±0.4**

* Results reproduced under aligned experimental conditions (episode ordering, training iterations, batch size, learning rate and dropout rate).

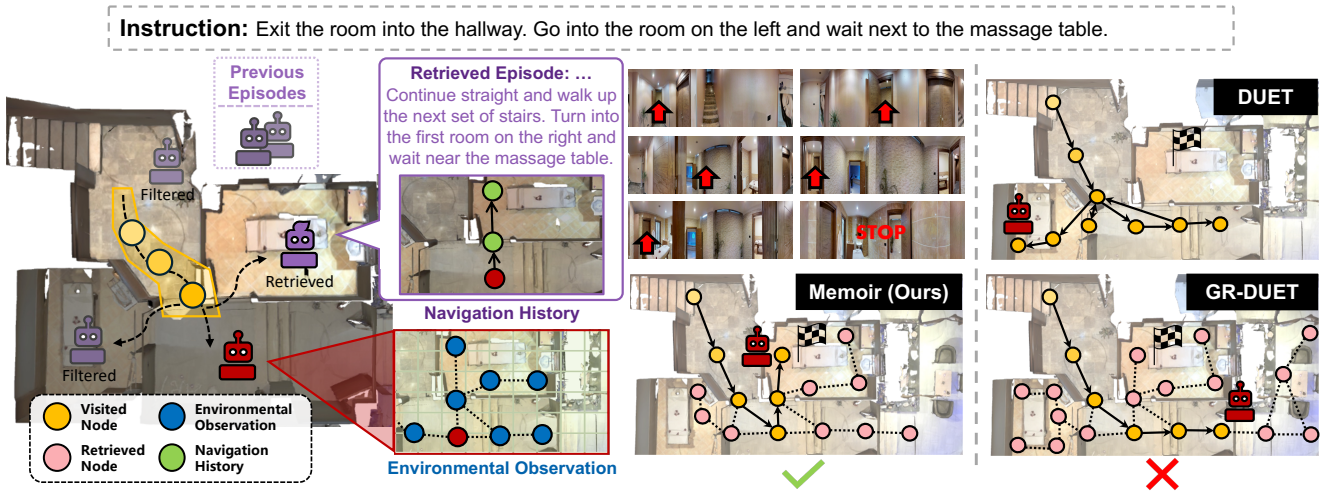
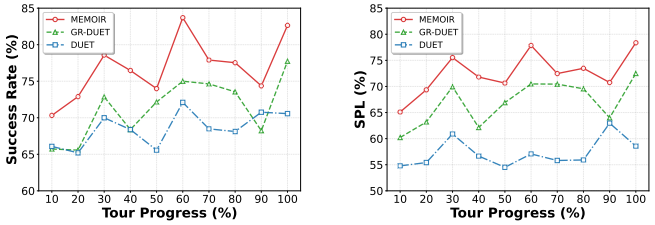


Fig. 3. Visualization of Memoir's memory retrieval from environmental observation bank and navigation history bank as well as the panoramic trajectory visualization. We compare the navigation result between DUET , GR-DUET and ours. The goal location is indicated by checkered flag.

TABLE 6
Computational efficiency comparison (batch size = 4).

Methods	Training		Inference	
	Memory ↓	Latency ↓	Memory ↓	Latency ↓
DUET [9]	7.2 GB	0.15s	2.2 GB	0.13s
GR-DUET [6]	29.4 GB	4.39s	9.9 GB	0.25s
Memoir (Ours)	**13.1 GB (-55%)**	**0.53s (-88%)**	**2.6 GB (-74%)**	0.31s (+28%)



(a) SR v.s. Tour Progress (b) SPL v.s. Tour Progress

Fig. 4. Performance scaling across tour progression on IR2R.

the combination of observation retrieval, which identifies promising paths toward relevant locations while controlling redundancy, and history retrieval, which matches similar past episodes targeting "massage room" objectives, prompting the agent to the destination.

5.4 Ablation Studies & Analyses

Computational Efficiency. Table 6 demonstrates Memoir's computational advantages over the memory-persistent baseline. While DUET operates with minimal memory overhead, GR-DUET's complete memory retention strategy dramatically increases resource requirements (29.4GB training, 9.9GB inference) due to processing all accumulated observations simultaneously. Our retrieval mechanism achieves substantial efficiency gains: 55% reduction in training memory and 88% reduction in training latency, representing an 8.3× speedup. During inference, memory usage decreases by 74%, approaching DUET's efficiency while maintaining memory-persistent capabilities. The slight inference latency increase (0.31s vs 0.25s) reflects the overhead of imagination-guided retrieval, suggesting opportunities for future optimization through caching or parallel processing. These results establish Memoir as the first memory-persistent VLN

TABLE 7
Ablation of components for memory retrieval.

Observation				History				IR2R Val Unseen								
RETR	RAND	FULL	PERF	RETR	RAND	FULL	PERF	TL ↓	NE ↓	SR ↑	SPL ↑	nDTW ↑	OR ↑	OA ↑	HR ↑	HA ↑
The upper-bound of long-term memory retrieval																
			✓				✓	9.77	0.51	95.44	93.40	93.68	100	100	100	100
								12.24	2.81	72.33	63.97	70.35	0.0	0.0	0.0	0.0
		✓				✓		10.44	2.86	74.67	69.98	76.29	100	9.81	100	19.33
	✓				✓			10.97	2.76	75.82	70.34	76.03	59.05	21.31	36.65	22.11
✓					✓			10.77	2.58	76.63	71.70	78.08	97.05	23.33	36.24	21.94
	✓			✓				10.80	2.61	76.63	71.03	76.98	58.83	21.72	98.36	22.40
✓				✓				**10.32**	**2.53**	**78.03**	**73.46**	**79.46**	96.49	**24.58**	96.52	**24.21**

* RETR: retrieval via imagination; RAND: random sampling; FULL: complete incorporation; PERF: oracle retrieval.

TABLE 8
Ablation of world model variants.

MODEL	DIST	IR2R Val Unseen SR ↑	SPL ↑	OR ↑	OA ↑	HR ↑	HA ↑
GRU	5	76.12	71.48	97.54	16.90	30.21	30.78
+overshoot	5	76.71	72.30	96.21	18.60	98.24	21.65
Transformer	1	73.44	67.24	31.25	40.33	26.26	30.74
	3	75.18	70.21	76.71	26.05	29.11	32.29
	5	77.14	72.11	96.44	17.86	28.58	32.32
+overshoot	1	73.09	67.26	31.15	43.22	59.43	29.24
	3	76.63	71.99	77.04	28.48	93.40	25.06
	5	**78.03**	**73.46**	96.49	24.58	96.52	24.21

TABLE 9
Ablation of navigation history integration.

HIST ENCODER	EMBED TYPE	IR2R Val Unseen SR ↑	SPL ↑	nDTW ↑	t-nDTW ↑
	VP + State	76.59	72.35	78.60	65.57
✓	VP	76.54	71.48	78.24	65.66
✓	State	77.35	72.83	78.70	66.37
✓	VP + State	**78.03**	**73.46**	**79.46**	**66.44**

TABLE 10
Ablation of expert policies.

EXPERT POLICY	IR2R Val Unseen SR ↑	SPL ↑	nDTW ↑	GSA Test-R-Basic SR ↑	SPL ↑	nDTW ↑
SPL	76.20	71.71	78.03	67.34	62.22	71.30
+random sample	**78.03**	**73.46**	**79.46**	**69.64**	**64.91**	**73.29**

approach that achieves both practical efficiency and state-of-the-art performance, making continuous learning viable for resource-constrained deployment while advancing navigation capabilities beyond existing methods.

Performance Scaling. Figure 4 demonstrates the scaling performance as agents complete episodes within IR2R tours. Memoir exhibits the most pronounced improvement with sustained high performance till the tour's end. In contrast, DUET maintains relatively stable but lower performance (65-72% SR, 55-63% SPL) without learning from accumulated experience. GR-DUET shows inconsistent scaling as its performance notably deteriorates at 90% progression, falling below DUET's baseline. This comparison validates that imagination-guided memory retrieval enables more consistent and substantial performance gains from accumulated experience compared to simple memory incorporation.

Memory Components. Table 7 presents results validating the effectiveness of memory components. The "perfect" variant directly incorporates memories leading to target locations, simulating ideal world model behavior with perfect retrieval capabilities, achieving 93.40% SPL on unseen environments, highlighting the necessity of retrieval and serving as an performance upper bound. The variant with both observation and history components disabled yields the lowest performance, followed by complete long-term memory incorporation. Random memory selection enhances navigation performance by 0.36% SPL, while substituting random selection with our retrieval approach improves navigation performance for both memory types, demonstrating that both observations and histories contain valuable signals-provided they are accessed selectively. Our retrieval variant achieves optimal navigation performance and retrieval accuracy (24.58% OA, 24.21% HA). Conversely, complete memory incorporation decreases accuracy, and random selection significantly reduces recall (OR and HR), demonstrating that predictive imagination enable precise memory filtering unavailable to uninformed and exhaustive approaches. The substantial gap relative to the ideal world model reveals current challenges in the world model's ability to capture environmental dynamics accurately. The world model would benefit from data scaling and advanced architectures to further enhance future performance.

World Model. Table 8 evaluates world model variants comparing GRU and Transformer architectures with and without overshooting. Transformer variant outperform GRU variant by 1.16% SPL, 7.68% observation retrieval accuracy (OA), and 2.56% history retrieval accuracy (HA) under 5-step overshooting distance. With increased imagination steps, retrieval effectiveness generally increases as the exploration horizon broadens, as recall improves for observations (OR) and histories (HR), enabling more informed navigation decisions. However, retrieval accuracy decreases due to exponentially increasing candidates in topological graphs with greater distances. The overshooting objective significantly enhances OA and HR, contributing to robust navigation performance (+1.36% SPL). These results validate that effective retrieval requires both powerful predictive models (Transformer over GRU) and grounded training (overshooting) to

TABLE 11
Ablation of world model pretraining.

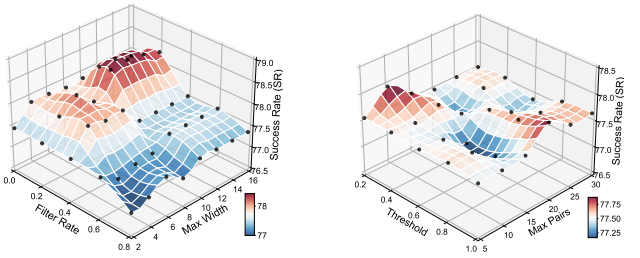
	IR2R Val Unseen			GSA Test-R-Basic		
PRETRAIN	SR ↑	SPL ↑	nDTW ↑	SR ↑	SPL ↑	nDTW ↑
	74.93	71.55	78.12	64.62	61.52	71.30
✓	**78.03**	**73.46**	**79.46**	**69.64**	**64.91**	**73.29**

TABLE 12
Ablation of observation completion.

	IR2R Val Unseen			GSA Test-R-Basic		
NEIGHBOR OBS	SR ↑	SPL ↑	nDTW ↑	SR ↑	SPL ↑	nDTW ↑
Partial	77.05	72.15	78.22	68.35	63.90	72.89
Completion	**78.03**	**73.46**	**79.46**	**69.64**	**64.91**	**73.29**

TABLE 13
Ablation of neighborhood incorporation.

	IR2R Val Unseen			GSA Test-R-Basic		
RETRIEVE	SR ↑	SPL ↑	nDTW ↑	SR ↑	SPL ↑	nDTW ↑
VP only	76.46	70.91	77.45	67.44	63.02	71.83
VP + neighbors	**78.03**	**73.46**	**79.46**	**69.64**	**64.91**	**73.29**



(a) Param Search (Obs.) (b) Param Search (Hist.)

Fig. 5. Study of hyper-parameters of retrieval on IR2R val-unseen. Left: Environmental observation retrieval. Right: Navigation history retrieval.

balance exploration breadth with query precision.

History Integration. Table 9 compares strategies for incorporating retrieved histories. Concatenating viewpoint features with state features in the dedicated encoder yields optimal performance, achieving 1.98% higher SPL than viewpoint features alone and 0.63% higher than state features alone. This indicates that state features carry crucial pattern information, while still benefiting from observation representation enhancement. Without incorporating history encoder, where historical representation is concatenated with coarse-scale encoder inputs, performance decreases by 1.11% SPL, demonstrating the necessity of separating duties across three distinct encoders.

Expert Policy Strategy. Table 10 compares training strategies when memory retrieval dynamically expands the available action space beyond immediate neighbors. Random sampling among multiple optimal paths during training (achieving 73.46% SPL) outperforms deterministic SPL-based expert selection (71.71% SPL) by providing better policy regularization and robustness to navigation choices.

World Model Pretraining. Table 11 demonstrates the importance of proper world model initialization. Pretraining the world model components on navigation trajectories before joint training improves performance by 1.91% SPL on IR2R and 3.39% SPL on GSA-R2R, indicating that randomly initialized world models provide poor retrieval signals.

Observation Completion. Table 12 demonstrates that completing partial observations at non-retrieved viewpoints using stored features from $\mathcal{M}_o$ significantly enhances environmental understanding. When a viewpoint in the episodic graph lacks complete visual information, retrieving its stored panoramic feature enables more informed decision-making.

Neighbor Incorporation. Table 13 studies the retrieval strategy that incorporates adjacent viewpoints of retrieved nodes during observation retrieval. Including immediate neighbors provides richer spatial context about connectivity and surrounding environment, enabling the coarse-scale encoder to make better-informed planning decisions. This approach improves SPL by 2.55% and 1.89% on respective benchmark.

Parameter Study. Figure 5 analyzes the impact of key retrieval hyperparameters. For observation retrieval, SR improves with reduced filter rates and increased search width, peaking at $\rho_o = 0.2$ and $W = 12$ before degrading as excessive context introduces noise. This indicates incorporating a broader range of viewpoint observations facilitates more informed navigation decisions. For navigation history retrieval, the model prioritizes precision over recall, achieving optimal performance at $\theta_h = 0.2$ with $P = 10$. A secondary optimum occurs at threshold $\theta_h = 1.0$ and max patterns $P = 20$ (decay factor $\gamma_h = 0.8$), where highly restrictive similarity thresholds compensate through increased pattern acceptance.

5.5 Failure Analysis

Figure 6 presents a scenario where Memoir fails despite functioning as designed. In Episode A, the agent must navigate to a bedroom absent from previous episodes. Observation retrieval identifies two distracting bedroom entrances as candidates, while history retrieval surfaces Episodes B (failed) and C (succeeded), both targeting a different bedroom.

Retrieval Limitations. The retrieved observations prefer incorporating abundant promising candidates as discovered in Figure 5(a), highlighting both bedroom entrances as semantically relevant but failing to discriminate the critical spatial feature—"nearest to the desk." Retrieved histories similarly cannot distinguish Episodes B and C despite different goals. In Episode B, premature imagination termination after one step limits retrieval to only the nearest entrances, preventing correct target discovery. These failures reveal world model deficiencies in predictive retrieval for both memory types.

Exploration-Exploitation Trade-off. Episode A fails when both retrieval types converge on the same incorrect location. The agent prioritizes high-similarity histories as discovered in Figure 5(b), defaulting to exploitation over exploration even when retrieval fails to cover the true goal. It also fails to distinguish task outcomes, treating Episodes B and C equally rather than learning from success. This highlights a

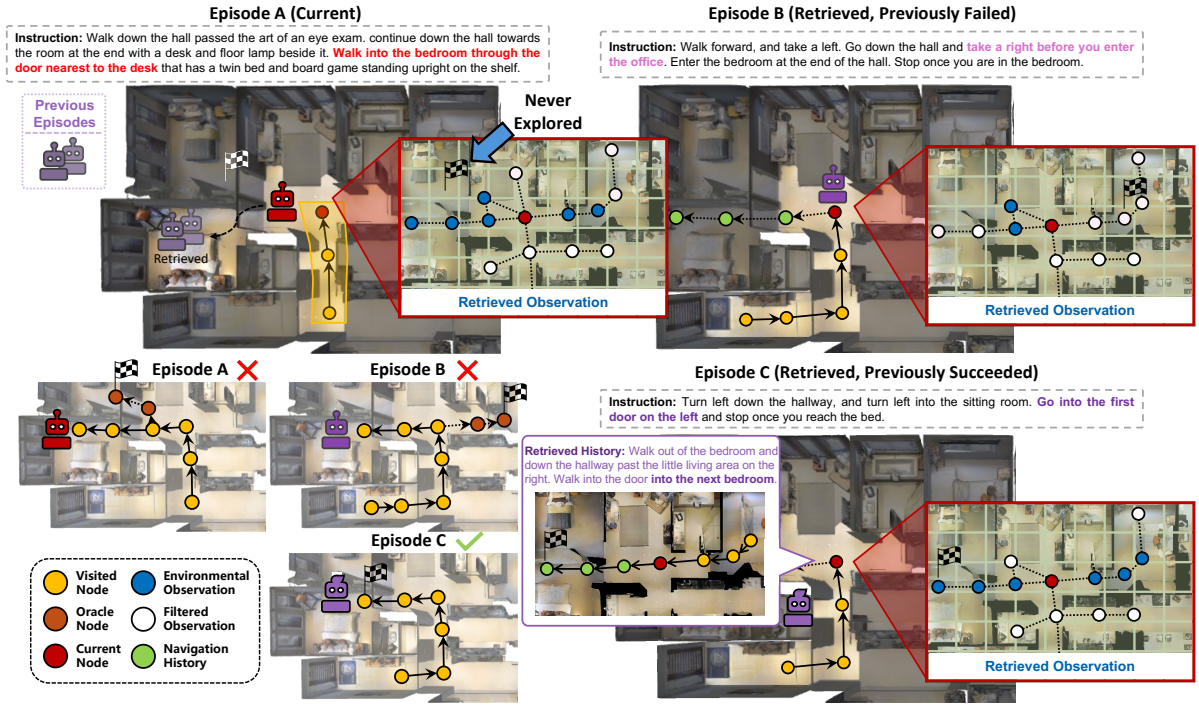


Fig. 6. Visualization of failure modes in imagination-guided memory retrieval. Episode A (current) retrieves experiences from Episodes B (previously failed) and C (previously succeeded) at a critical decision point. Key instruction differences are highlighted in bold.

new challenge: determining when accumulated experience should be trusted versus when novel alternatives warrant investigation.

Future Work. These failure modes suggest two potential research directions for advancing imagination-guided memory retrieval. First, *enhanced world modeling capability* through larger-scale pretraining and explicit spatial relationship modeling could address both retrieval inaccuracies in distinguishing spatially distinct targets and premature imagination horizons that affect retrieval scope. Second, *confidence-aware retrieval* to determine when retrieved experience should be trusted, requiring retrieval confidence estimation to dynamically balance exploitation against exploration. The substantial performance gap between our method (73.46% SPL) and the oracle retrieval upper bound (93.40% SPL in Table 7) demonstrates significant room for improvement in these directions.

6 Conclusion

This work introduces Memoir, a memory-persistent VLN agent employing predictive world modeling for adaptive experience retrieval. Unlike traditional imagine-planning that generates trajectories in isolation, we ground imagination with explicit memory through a language-conditioned world model, Hybrid Viewpoint-Level Memory (HVM) storing observations and behavioral patterns, and an experience-augmented navigation model. Extensive experiments demonstrate 5.4% SPL improvement on IR2R with 8.3× training speedup and 74% inference memory reduction, validating that predictive retrieval of both environmental and behavioral memories enables more effective navigation. The oracle retrieval performance (93.4% SPL) demonstrates the potential of imagination-guided retrieval. Future work should explore enhanced world modeling and confidence-aware exploration mechanisms to narrow this gap, establishing a principled framework connecting predictive simulation with explicit memory for embodied AI.

References

[1] P. Anderson *et al.*, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *CVPR*, 2018, pp. 3674–3683.

[2] Y. Qi *et al.*, "Reverie: Remote embodied visual referring expression in real indoor environments," in *CVPR*, 2020, pp. 9982–9991.

[3] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," in *EMNLP*, 2020, pp. 4392–4412.

[4] S. Wani, S. Patel, U. Jain, A. Chang, and M. Savva, "Multion: Benchmarking semantic map memory using multi-object navigation," in *NeurIPS*, vol. 33, 2020, pp. 9700–9712.

[5] J. Krantz *et al.*, "Iterative vision-and-language navigation," in *CVPR*, 2023, pp. 14 921–14 930.

[6] H. Hong, Y. Qiao, S. Wang, J. Liu, and Q. Wu, "General scene adaptation for vision-and-language navigation," in *ICLR*, 2025.

[7] G. Zhao, G. Li, W. Chen, and Y. Yu, "Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation," in *CVPR*, 2024, pp. 16 296–16 306.

[8] Q. Zheng, D. Liu, C. Wang, J. Zhang, D. Wang, and D. Tao, "Esceme: Vision-and-language navigation with episodic scene memory," *IJCV*, pp. 1–21, 2024.

[9] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Think global, act local: Dual-scale graph transformer for vision-and-language navigation," in *CVPR*, 2022, pp. 16 537–16 547.

[10] M. Seeber *et al.*, "Human neural dynamics of real-world and imagined navigation," *Nat. Hum. Behav.*, vol. 9, no. 4, pp. 781–793, 2025.

[11] M. Karl, F. Kock, B. W. Ritchie, and J. Gauss, "Affective forecasting and travel decision-making: An investigation in times of a pandemic," *Ann. Tour. Res.*, vol. 87, p. 103139, 2021.

[12] Y. W. Li and L. C. Wan, "Inspiring tourists' imagination: How and when human presence in photographs enhances travel mental simulation and destination attractiveness," *Tour. Manag.*, vol. 106, p. 104969, 2025.

[13] H. Wang, W. Liang, L. Van Gool, and W. Wang, "Dreamwalker: Mental planning for continuous vision-language navigation," in *ICCV*, 2023, pp. 10 873–10 883.

[14] D. Fried *et al.*, "Speaker-follower models for vision-and-language navigation," in *NeurIPS*, vol. 31, 2018.

[15] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *CVPR*, 2020, pp. 13 137–13 146.

[16] Z. Wang *et al.*, "Scaling data generation in vision-and-language navigation," in *ICCV*, 2023, pp. 12 009–12 020.

[17] J. Zhang *et al.*, "Navid: Video-based vlm plans the next step for vision-and-language navigation," in *RSS*, 2024.

[18] Y. Xu, Y. Pan, Z. Liu, and H. Wang, "Flame: Learning to navigate with multimodal llm in urban environments," in *AAAI*, vol. 39, 2025, pp. 9005–9013.

[19] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "Vln bert: A recurrent vision-and-language bert for navigation," in *CVPR*, 2021, pp. 1643–1653.

[20] S. Chen, P.-L. Guhur, C. Schmid, and I. Laptev, "History aware multimodal transformer for vision-and-language navigation," in *NeurIPS*, vol. 34, 2021, pp. 5834–5847.

[21] H. Wang, W. Wang, W. Liang, C. Xiong, and J. Shen, "Structured scene memory for vision-language navigation," in *CVPR*, 2021, pp. 8455–8464.

[22] E. Parisotto and R. Salakhutdinov, "Neural map: Structured memory for deep reinforcement learning," in *ICLR*, 2018.

[23] X. Dong *et al.*, "Se-vln: A self-evolving vision-language navigation framework based on multimodal large language models," *arXiv preprint arXiv:2507.13152*, 2025.

[24] C. Wang, S. Wei, and J. Qi, "Matchnav: Llm-based enhanced description and instruction matching in vision-and-language navigation," *Inf. Fusion*, vol. 125, p. 103444, 2026.

[25] J. F. Henriques and A. Vedaldi, "Mapnet: An allocentric spatial memory for mapping environments," in *CVPR*, 2018, pp. 8476–8484.

[26] S. K. Ramakrishnan, Z. Al-Halah, and K. Grauman, "Occupancy anticipation for efficient exploration and navigation," in *ECCV*, 2020, pp. 400–418.

[27] V. Cartillier, Z. Ren, N. Jain, S. Lee, I. Essa, and D. Batra, "Semantic mapnet: Building allocentric semantic maps and representations from egocentric views," in *AAAI*, vol. 35, 2021, pp. 964–972.

[28] C. Huang, O. Mees, A. Zeng, and W. Burgard, "Visual language maps for robot navigation," in *ICRA*, 2023, pp. 10 608–10 615.

[29] Z. Teng *et al.*, "360bev: Panoramic semantic mapping for indoor bird's-eye view," in *WACV*, 2024, pp. 373–382.

[30] R. Liu, X. Wang, W. Wang, and Y. Yang, "Bird's-eye-view scene graph for vision-language navigation," in *ICCV*, 2023, pp. 10 968–10 980.

[31] D. S. Chaplot, R. Salakhutdinov, A. Gupta, and S. Gupta, "Neural topological slam for visual navigation," in *CVPR*, 2020, pp. 12 875–12 884.

[32] N. Kim, O. Kwon, H. Yoo, Y. Choi, J. Park, and S. Oh, "Topological semantic graph memory for image-goal navigation," in *CoRL*, 2023, pp. 393–402.

[33] A. Werby, C. Huang, M. Büchner, A. Valada, and W. Burgard, "Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation," in *RSS*, 2024.

[34] D. Hafner *et al.*, "Learning latent dynamics for planning from pixels," in *ICML*, 2019, pp. 2555–2565.

[35] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *ICLR*, 2020.

[36] J. Lin *et al.*, "Learning to model the world with language," in *ICML*, vol. 235, 2024, pp. 29 992–30 017.

[37] J. Wu, H. Ma, C. Deng, and M. Long, "Pre-training contextualized world models with in-the-wild videos for reinforcement learning," in *NeurIPS*, vol. 36, 2024.

[38] X. Ma, S. Chen, D. Hsu, and W. S. Lee, "Contrastive variational reinforcement learning for complex observations," in *CoRL*, 2021, pp. 959–972.

[39] M. Okada and T. Taniguchi, "Dreaming: Model-based reinforcement learning by latent imagination without reconstruction," in *ICRA*, 2021, pp. 4209–4215.

[40] A. Bar, G. Zhou, D. Tran, T. Darrell, and Y. LeCun, "Navigation world models," in *CVPR*, 2025, pp. 15 791–15 801.

[41] J. Li and M. Bansal, "Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation," in *NeurIPS*, vol. 36, 2023, pp. 21 878–21 894.

[42] X. Yao, J. Gao, and C. Xu, "Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments," *arXiv preprint arXiv:2506.23468*, 2025.

[43] J. Y. Koh, H. Lee, Y. Yang, J. Baldridge, and P. Anderson, "Pathdreamer: A world model for indoor navigation," in *ICCV*, 2021, pp. 14 738–14 748.

[44] G. Georgakis, K. Schmeckpeper, K. Wanchoo, S. Dan, E. Miltsakaki, D. Roth, and K. Daniilidis, "Cross-modal map learning for vision and language navigation," in *CVPR*, 2022, pp. 15 460–15 470.

[45] Y. Pan, Y. Xu, Z. Liu, and H. Wang, "Planning from imagination: Episodic simulation and episodic memory for vision-and-language navigation," in *AAAI*, vol. 39, 2025, pp. 6345–6353.

[46] J. Li and M. Bansal, "Improving vision-and-language navigation by generating future-view image semantics," in *CVPR*, 2023, pp. 10 803–10 812.

[47] S. Wang *et al.*, "Monodream: Monocular vision-language navigation with panoramic dreaming," *arXiv preprint arXiv:2508.02549*, 2025.

[48] H. Le, T. Karimpanal George, M. Abdolshah, T. Tran, and S. Venkatesh, "Model-based episodic memory induces dynamic hybrid controls," in *NeurIPS*, vol. 34, 2021, pp. 30 313–30 325.

[49] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2013.

[50] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[51] S. K. Ramakrishnan *et al.*, "Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai," in *NeurIPS D&B*, 2021.

[52] P. Anderson *et al.*, "On evaluation of embodied navigation agents," *arXiv preprint arXiv:1807.06757*, 2018.

[53] H. Wang, W. Wang, T. Shu, W. Liang, and J. Shen, "Active visual information gathering for vision-language navigation," in *ECCV*, 2020, pp. 307–322.

[54] D. Wang, E. Shelhamer, S. Liu, B. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021.

[55] S. Niu *et al.*, "Towards stable test-time adaptation in dynamic wild world," in *ICLR*, 2023.