Planning from Imagination: Episodic Simulation and Episodic Memory for Vision-and-Language Navigation

Yiyuan Pan[1], Yunzhe Xu[2], Zhe Liu[2]*, Hesheng Wang[1]

[1]Department of Automation, Shanghai Jiao Tong University, China
[2]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China
{pyy030406, xyz9911, liuzhesjtu, wanghesheng}@sjtu.edu.cn

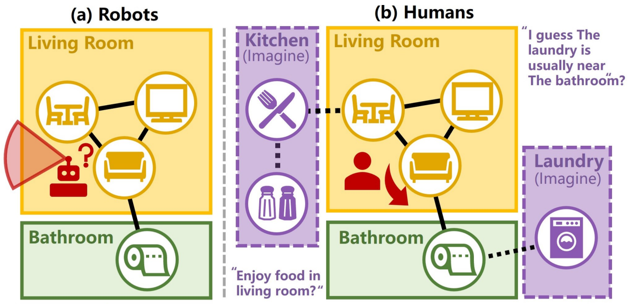
Abstract

Humans navigate unfamiliar environments using episodic simulation and episodic memory, which facilitate a deeper understanding of the complex relationships between environments and objects. Developing an imaginative memory system inspired by human mechanisms can enhance the navigation performance of embodied agents in unseen environments. However, existing Vision-and-Language Navigation (VLN) agents lack a memory mechanism of this kind. To address this, we propose a novel architecture that equips agents with a reality-imagination hybrid memory system. This system enables agents to maintain and expand their memory through both imaginative mechanisms and navigation actions. Additionally, we design tailored pre-training tasks to develop the agent's imaginative capabilities. Our agent can imagine high-fidelity RGB images for future scenes, achieving state-of-the-art results in a Success rate weighted by Path Length (`SPL`).

1 Introduction

Autonomous embodied agent navigation is advancing through Vision-and-Language Navigation (VLN) research (Anderson et al. 2018b; Qi et al. 2020). In VLN tasks, an agent needs to reach the target location given navigation instructions. This task, however, is more challenging when agents navigate in unseen environments, with performance degradation compared to seen environments.

Human navigation behaviors in unfamiliar environments provide valuable insights, particularly through the neuroscience concepts of episodic simulation and memory (Kühn and Gallinat 2014; Gomez, Rousset, and Baciu 2009), which may help address this challenge. These mechanisms enable humans to use memory to mentally simulate uncertain information or scenes, aiding decision-making. Humans can imagine fine-grained visual features (e.g., RGB images) and higher-level spatial structures (e.g., location distributions) in unseen environments (Tang, Yan, and Tan 2017). As illustrated in Figure 1, given an instruction like "go to the laundry room", humans might imagine a room with washing machines, likely adjacent to a bathroom encountered earlier, using such mental simulations to infer navigation directions.



Instruction: Step into the living room. **When you are near the sofa, turn to the** laundry **and stop next to the window.**

Figure 1: Humans utilize episodic memory and episodic simulation to recall past experiences and predict future outcomes in unfamiliar environments. In contrast, navigation agents often struggle in unseen environments due to their inability to construct and leverage such cognitive frameworks.

While existing approaches have incorporated mechanisms for imagining unseen environments, they lack the ability to integrate the imagined results into time-series persistent memory. Specifically, This prevents the formation of long-term episodic simulations, which is essential for effective navigation (Wang et al. 2023a; An et al. 2023). Current methods enable agents to predict potential object locations (An et al. 2024) and RGB features (Moudgil et al. 2021; Qi et al. 2021), but these predictions are transient and overwritten at each step, precluding durable memory representations. Such limitations hinder agents from fusing visual and contextual information to support future reasoning or decision-making. For instance, in Figure 1, an agent's inability to infer the presence of spice bottles from contextual cues such as a knife and fork illustrates the shortcomings of existing mechanisms. Consequently, these methods produce isolated and short-lived imaginative outputs, lacking the continuity and compositional depth characteristic of human episodic memory and simulation.

In this paper, we introduce the **S**pace-**A**ware **L**ong-term **I**maginer (SALI) agent, the first navigation agent explicitly designed to emulate human-like episodic simulation and memory. SALI leverages an imaginative memory system to capture both high-level spatial structures of the environ-

[*]Corresponding Author

ment and fine-grained RGB features. The generation of these imaginative outputs is achieved by integrating prior imagined results with historical information, forming a recurrent imagination module. This iterative process enables SALI to maintain a hybrid memory system that combines real and imagined representations, facilitating robust reasoning and navigation in complex and unseen environments. To implement this system, we model the agent's memory using a topological map (Chen et al. 2022), where nodes represent spatial locations enriched with RGB-D and semantic features generated by the imagination module or the navigation actions. Multimodal transformers then encode these nodes alongside natural language instructions, enabling reasonable navigation decisions at each step. The integration of imaginative and realistic information allows SALI to reason holistically about its environment while dynamically adapting to new scenarios. As a result, SALI achieves state-of-the-art (SoTA) performance on R2R (Anderson et al. 2018b) and REVERIE (Qi et al. 2020), highlighting its effectiveness.

To summarize, our contributions are as follows.

- Our proposed SALI has human-like episodic memory and episodic simulation abilities, enhancing general navigation performance. Viewing from the perspective of human brains, we endow the agent with the ability to learn anticipatory knowledge through imagination.
- We established a recurrent, end-to-end imagination module that generates high-fidelity RGB scene representations. SALI dynamically fuses imagined scenes with real observations into a hybrid memory map, ensuring effective navigation decisions.
- SALI achieves SoTA performance on R2R and REVERIE, improving `SPL` by 8% and 4% in unseen scenarios, respectively.

2 Related Works

Vision-and-language Navigation. Guiding robot navigation using visual inputs and natural language instructions is a core task in embodied AI (Anderson et al. 2018b; Qi et al. 2020). A key challenge lies in effectively aligning multimodal visual-linguistic inputs (Lin et al. 2022). Existing VLN approaches (Lu et al. 2019; Qi et al. 2021; Hong et al. 2021) often falter in unseen environments, struggling to associate unfamiliar visual observations, such as novel materials and textures, with navigation instructions, leading to degraded performance. To address these limitations, we propose a novel VLN architecture inspired by human episodic simulation and memory, enhancing robustness and performance in unseen environments.

Memory Mechanism. Constructing memory for long-term decision support is critical for embodied agents. SLAM has been widely used to build memory maps in navigation (Chaplot et al. 2020; Temeltas and Kayak 2008). To reduce memory overhead, more methods such as 2D bird-eye-view maps (An et al. 2023; Liu et al. 2023) or ego-centric grid maps (Georgakis et al. 2022; Wang et al. 2023c) are investigated to help agents better understand spatial information. Additionally, some other navigation agents employ topological maps to further simplify and extract high-level features to learn underlying spatial knowledge (An et al. 2023). However, Integrating historical memory with future imagination is underexplored, limiting navigation robustness in unseen environments. Drawing inspiration from human cognition, we design an episodic simulation-based memory mechanism to enhance agents' foresight.

Prediction during Navigation. Inspired by video prediction (Mallya et al. 2020), prior works have adopted imaginary mechanisms to support navigation decisions by generating fine-grained visual information. They use RGB-D images to generate high-fidelity images (Moudgil et al. 2021; Wang et al. 2023a). Furthermore, there is also pioneering work that explores the generation of spatial structural information (Wang et al. 2018b; Li et al. 2023) with the help of depth cameras (Shah et al. 2023; Cui et al. 2019). However, existing imagination mechanisms operate in isolation, without integration into persistent memory. We propose a memory-based framework to bridge transient predictions with long-term memory, enabling reusable navigation information. Our adaptive mechanism integrates imagination into memory for durable, context-aware navigation.

3 Method

Problem Formulation. The VLN task is set in a discrete environment $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ (Anderson et al. 2018b), where $\mathcal{V}$ is navigation nodes and $\mathcal{E}$ represents the connectivity relations. The VLN dataset provides instruction set I and corresponding ground-truth navigation paths. It requires guiding a robot to accomplish navigation tasks with given instructions. At each navigation step t, the agent can acquire the RGB image r_t and depth image d_t via the camera and the current position information p_t via the GPS sensor. Following (Irshad et al. 2022), an agent is equipped with a visual classifier to acquire semantic images s_t. The agent needs to learn a policy to make actions based on observations $O_t = \{r_t, d_t, s_t, p_t\}$ and instruction.

Method Overview. As shown in Figure 2, SALI has a human-like episodic simulation and episodic memory mechanism. At each navigation step t, the agent will maintain a topological map as its memory to store both realistic and imaginative information and make navigation decisions based solely on the memory (Section 3.1). Then, the agent will use its memory to imagine future information of both high-level spatial knowledge and low-level image features and merge the imagination into the memory (Section 3.2). We conclude this part by presenting our approach to training and inference in Section 3.3.

3.1 Real-Imaginary Hybrid Memory

To provide the agent with the ability of global action planning, we construct a mixed-granularity memory. As shown in Figure 2, at navigation step t, the agent will update the topological memory G_t based on the current observation O_t and historical information.

Memory Map Representation The topological map memory is represented as $G_t = \{N_t, E_t\}$, where E_t records the Euclidean distance between neighboring nodes, and N_t represents the nodes containing visual inputs V_t, position

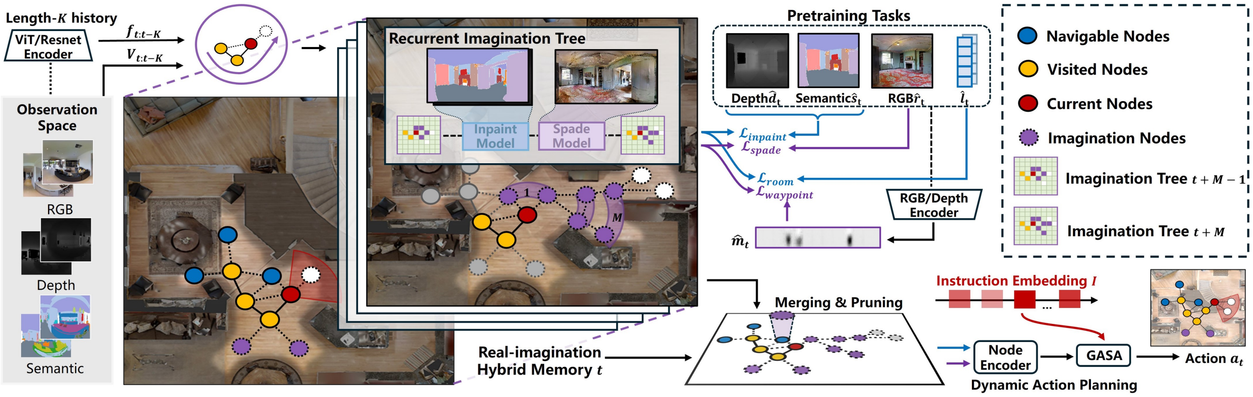


Figure 2: We propose building a hybrid imagination-reality memory for long-term navigation decisions. Based on the navigation observation and trajectories, the agent will imagine future scenes for unvisited environments. The imagination will then be fused into its hybrid memory to aid further decision-making. The figure also illustrates a series of pre-training tasks that we propose for the imagination module.

information p_t, and node feature f_t. Visual inputs $V_t = \{r_t, d_t, s_t\}$, stored in equirectangular panorama form, are used to generate fine-grained images. We use a pre-trained Vision Transformer (ViT) (Lu et al. 2019) and ResNet encoder to get the node feature f_t from visual inputs V_t.

Nodes N_t can be classified into four categories: visited nodes 🟡, current nodes 🔴, navigable nodes 🔵, and imagination nodes 🟣. Each node type stores different information. For V_t, visited and current nodes store complete and fixed image information, while the navigable node's V_t is represented by partial visual inputs observed from neighboring real nodes and contains potential information from imagination nodes. The V_t of the imagination node is completely generated from imagined images. For p_t, the agent can only reach the potential position associated with navigable nodes, as locations corresponding to 🟣, 🔴, or 🟡 are not accessible via action a_t.

To manage memory efficiency, pruning operations are applied to newly imagined nodes. If two nodes are determined to represent the same node, the pruning operation retains a single node, and its feature f_t is updated using the average pooling of the features from both nodes. The new node remains an imagination node when both are imagination nodes, otherwise is labeled as a navigable node. In addition, we set an upper bound of $\bar{N}$ on the number of imagination nodes. The criteria of pruning are based on the feature cosine similarity and the negative position mean square error (MSE) between the imagination node N_i and another node N_j:

$$\text{Criterion}(N_i, N_j) = \frac{f_i f_j}{||f_i|| ||f_j||} - \text{MSE}(p_i, p_j). \quad (1)$$

Dynamic Action Planning As shown in Figure 2, the topological memory G_t at time t is processed by a multi-modal transformer to obtain contextual representations. this transformer contains pre-trained encoders for real nodes (🟡, 🔵, 🔴), fine-tuned encoders for imagination nodes (🟣), and a pre-trained instruction encoder. The real node encoder and the imagination node encoder share the same structure, unified as the node encoder for simplicity.

Node Embedding. The input to generate node embeddings $\hat{V}_t$ includes three elements: node feature f_t, location encoding, and navigation step encoding. The latter marks visited nodes with their last visited time (0 for 🟣, 🔵). This encoding reflects the temporal structure of memory. Node embeddings $\hat{V}_t$ are categorized into $\hat{V}_t^r$ and $\hat{V}_t^i$ based on whether the nodes are real or imagined. Furthermore, a 'stop' node is added to the memory to represent a stop action and is connected to all other nodes.

Instruction and Node Encoders. Each word embedding in instruction I is augmented with a positional embedding and a type embedding. A multi-layer transformer processes these tokens to generate contextual representations, denoted as instruction embeddings $\hat{I}$.

Node and word embeddings $\hat{V}_t$ and $\hat{I}$ are fed into a multi-layer cross-modal transformer. Following (Chen et al. 2022), we endow the transformer with a graph-aware self-attention (GASA) layer to capture the environment layout. The network outputs navigation scores s_i^r and s_i^i for each navigable and imagination node as follows, where FFN is a two-layer feed-forward network:

$$s_t^k = \text{FFN}(\text{GASA}(\hat{V}_t^k)), k = i, r. \quad (2)$$

Action Fusion Policy. We propose an adaptive framework for navigation decision-making, which involves the dynamic integration of navigation scores obtained from both navigable nodes and imagination nodes. Initially, navigation scores from imagination nodes will be multiplied by a fusion factor γ_t and added to the scores of the nearest navigable nodes in Euclidean distance, since they are not reachable. The fusion factor is generated by concatenating the imagination nodes and real nodes and then feeding it into an FFN layer. The

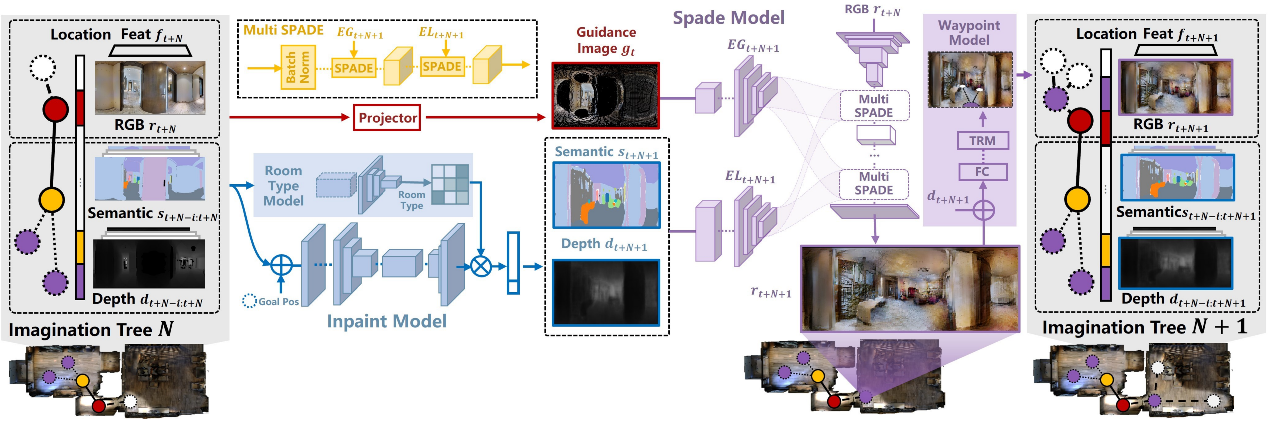


Figure 3: The imagination model includes four pre-trained models (inpaint model, spade model, room-type model, and waypoint model). We propose an end-to-end architecture that allows local imaginary trees to continuously update and maintain themselves, producing high-quality depth images, semantic images, and RGB images.

scores for next-node selection are obtained by:

$$\gamma_t = \text{Sigmoid}(\text{FFN}(\left[\hat{V}_t^r, \hat{V}_t^i\right])), \tag{3}$$

$$\hat{s_t} = s_t^r + \sum\nolimits_{s_t^i \in \mathcal{S}(i)} \gamma_t s_t^i, \tag{4}$$

where $\mathcal{S}(i)$ represents the set of all imagination nodes that need to be added for navigable node N_i.

3.2 Recurrent Imagination Tree

We put forth a recurrent imagination module with a tree structure. Drawing inspiration from (Koh et al. 2021; Mallya et al. 2020), SALI can generate high-resolution future images. At step t, a length-K history information $H_t = \{d_{t-K:t}, s_{t-K:t}, p_{t-K:t}\}$, the RGB image r_t, and the neighboring position of the current position p_t^g—all extracted from navigation memory G_t—are utilized to initialize imaginary tree $T_t^0 = \{H_t, r_t, p_t^g\}$. The imaginary tree T_t^M is then generated iteratively and finally integrated with the G_t as mentioned in Section 3.1.1.

Recurrent Imagination Mechanism The imagination tree grows iteratively. Without loss of generality, Figure 3 illustrates the process of expanding a recurrent imagination tree of imagination step $N+1$ ($N+1 \leq M$) from step $t+N$. We first use the input queue T_t^N to generate a list of structured label images s_t^{N+1} and d_t^{N+1} at imagination step $N+1$ by the inpaint model for each navigable position $p_t^{g,N}$. Subsequently, the guidance image g_t^N is generated through a point cloud projector. Finally, the RGB image r_t^{N+1} and the target position $p_t^{g,N+1}$ will be generated by the spade model with the guidance image g_t^N, the semantic image s_t^{N+1}, and the depth image d_t^{N+1} as inputs. The short-term memory is then updated to the state T_t^{N+1}. It is noteworthy that all images are uniformly oriented with 0 headings to prevent the generation of different images for the same point due to variations in agent action orientations.

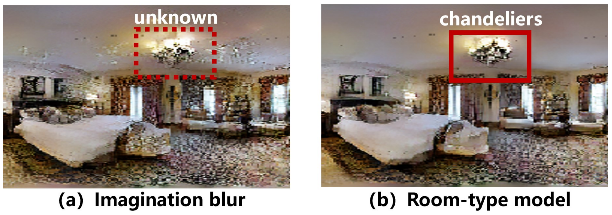


Figure 4: Before and after adding room-type model.

High-fidelity Image Generation

1) Inpaint Model. The inpaint model is an encoder-decoder-based network designed for depth and semantic image generation. For each navigable position $p_t^{g,N}$, the process begins by converting historical data H_t into corresponding structural maps x_t^{N+1} through a point cloud projector. Subsequently, the map is fed into an encoder-decoder based on RedNet and ResNet (He et al. 2016a) to generate one-hot semantic code e_t^{N+1} and d_t^{N+1}, obtained by:

$$X_t = \text{Encoder-Decoder}(x_t), \tag{5}$$

$$[d_{t+1}, e_{t+1}] = \sigma C_S(X) + (1-\sigma) C_F(X_t), \tag{6}$$

where σ is denoted as weight parameter $C_G(X_t)$. C_G, C_S, C_F stand for Gate-, Shortcut- and Final-Convolution layer.

As the imagination step upper bound M increases, the generated semantic images become increasingly ambiguous, with overlapping likelihoods for objects in the same pixel (Figure 4) (Kaneko and Harada 2021). The likelihood of different objects in the same pixel would be similar. For this reason, we incorporate the room type model. Given the high correlation between room information and objects (e.g., refrigerator-kitchen, bed-bedroom), SALI can ascertain the presence of specific objects by determining the room type based on commonsense knowledge. SALI uses a predefined object weight dictionary w for each room, refining one-hot semantic code e_{t+1} as $e_{t+1} \cdot (1+w)$. The updated code is then transformed to generate s_{t+1}.

2) Spade Model. The spade model is an image-to-image translation GAN (Wang et al. 2018a; Goodfellow et al. 2020) that enables high-resolution RGB generation capability (Park et al. 2019). The input to the spade model includes the embeddings of the RGB image r_t^N and the guidance image g_t^N, and the output is the RGB image r_t^{N+1}.

For the following navigable position prediction $p_t^{g,N+1}$, the generated RGB-D images r_t^{N+1} and d_t^{N+1} are fed into a BERT-based (Kenton and Toutanova 2019) waypoint model.

Cross-correction Global memory and local imagination mechanisms are designed to complement each other. Imagination enhances the image features of navigable nodes and facilitates map expansion. Conversely, the global map provides historical context and trajectory options, reducing information loss for imagination processes.

3.3 Training and Inference

Multimodal Transformer Pre-training Previous studies have shown the benefits of pre-training VLN models with auxiliary tasks. We pre-train our model using expert behavior and imitation learning. Pre-training tasks include masked language modeling (MLM) (Kenton and Toutanova 2019), masked region classification (MRC) (Lu et al. 2019), single-step action prediction (SAP) (Chen et al. 2021), and object grounding (OG) (Lin, Li, and Yu 2021).

Imagination Model Pre-training

1) Inpaint Model. With random noise added to the input, we trained the inpaint model by minimizing a combination of cross-entropy loss and mean absolute error (MAE):

$$\mathcal{L}_{inpaint} = -\lambda \sum\nolimits_i \hat{s}_i \log(s_t) + (1-\lambda)\|d_t - \hat{d}_t\|_1, \quad (7)$$

where λ is weight parameter. s_t, d_t are generated semantic and depth images, while $\hat{s}_t, \hat{d}_t$ are ground-truth images.

For room-type model training, we get the one-hot codes of room-type for each viewpoint from the Matterport3D Simulator (Anderson et al. 2018b). Two ResNet-50 networks (He et al. 2016a), which have been previously trained on the ImageNet dataset (Russakovsky et al. 2015), are employed to encode the semantic image s and the depth image d, respectively. The room-type model takes in the encoded feature and outputs one-hot codes. We calculate cross-entropy loss between the predicted label l and true room type label $\hat{l}$:

$$\mathcal{L}_{room} = -\lambda \sum\nolimits_i \hat{l}_i \log(l_i). \quad (8)$$

2) Spade Model. The spade model is a GAN-based model. We train the generator with GAN hinge loss, feature matching loss, and perceptual loss (Koh et al. 2021). For PatchGAN-based discriminator (Phillip et al. 2017), we calculate the loss for ground-truth images $\hat{r}_t$ and generated images r_t, where ϕ^i and D^i denote the output of the i-th layer of the pre-trained VGG-19 model and discriminator:

$$\begin{aligned}\mathcal{L}_{spade}^{d} = &-\mathbb{E}[\min(0, -1 + D(\hat{r}_t))] \\ &-\mathbb{E}[\min(0, -1 + D(r_t))], \quad (9) \\ \mathcal{L}_{spade}^{g} = &-\lambda_G \mathbb{E}[D(r_t)] \\ &+\lambda_F \sum\nolimits_i^n \frac{\|\phi^i(\hat{r}_t) - \phi^i(r_t)\|_1}{n} \\ &+\lambda_P \sum\nolimits_i^n \frac{\|D^i(\hat{r}_t) - D^i(r_t)\|_1}{n}. \quad (10)\end{aligned}$$

We implement the waypoint prediction model following (Hong et al. 2022). For each viewpoint, we labeled its neighboring points' relative headings and distances into a 120×12 matrix representing 360 degrees and 3 meters (each element represents 3 degrees and 0.25 meters). A heat map, designated as $\hat{m} \in \mathbb{R}^{120\times12}$, is then generated as the ground-truth data through interpolation of the aforementioned matrix. For the training process, the above-mentioned two ResNet-50 networks are employed to encode the RGB image and depth image for embedding, e^r and e^d. Subsequently, e^r, e^d are merged by a non-linear layer and fed into the model, which generates a heatmap, $m \in \mathbb{R}^{120\times12}$. Subsequently, non-maximum suppression (NMS) (Hosang, Benenson, and Schiele 2017) is applied to obtain neighboring waypoints. The loss is calculated by:

$$\mathcal{L}_{waypint} = \frac{1}{120 \times 12} \sum\nolimits_{i=1}^{120} \sum\nolimits_{j=1}^{12} (m_{ij} - \hat{m}_{ij})^2. \quad (11)$$

Inference During inference, the agent constructs a global map of an imagination-based memory on the fly. It then reasons about the next action over the map, as explained in Section 3.1. As the memory expands, SALI will approach the correct point and find the shortest path to the target. The agent will stop if it selects the 'stop' node or reaches the maximum action steps.

4 Experiments

4.1 Task Setup

Datasets. We evaluate SALI on VLN benchmarks with both fine-grained instructions (R2R) (Anderson et al. 2018b) and coarse ones (REVERIE) (Qi et al. 2020). R2R provides step-by-step instructions with an average length of 32 words. REVERIE gives pre-defined object bounding boxes and instructions to guide agents in describing target objects' positions. Instructions consist of 21 words on average.

Evaluation Metrics. We adopt the evaluation metrics (Anderson et al. 2018a) used commonly by existing works: 1) Navigation Error (NE) calculates the distance between stop locations and target ones; 2) Trajectory Length (TL) represents the average path length; 3) Success Rate (SR) shows the ratio of stopping within 3m to the target; 4) Success rate weighted by Path Length (SPL) makes trade-off between SR and TL; 5) Oracle Success Rate (OSR) is the ratio of including a viewpoint along the path where the target position can be seen; 6) Remote Grounding Success (RGS) is the ratio of successfully executed instructions; 7) RGSPL is RGS penalized by path length.

	Val Unseen-R2R				Test Unseen-R2R				Val Unseen-REVERIE			Test Unseen-REVERIE		
Methods	**NE↓**	OSR↑	**SR↑**	**SPL↑**	**NE↓**	OSR↑	**SR↑**	**SPL↑**	**SR↑**	**RGS↑**	**RGSPL↑**	**SR↑**	**RGS↑**	**RGSPL↑**
Seq2Seq (Anderson et al. 2018b)	7.81	28	21	-	7.85	27	20	-	56	36	26	55	32	22
RecBert (Hong et al. 2021)	3.93	-	63	57	4.09	70	63	57	30	18	15	29	16	13
HOP+ (Qiao et al. 2023)	3.49	-	67	61	3.71	-	66	60	36	22	19	34	20	17
DUET (Chen et al. 2022)	3.31	81	72	60	3.65	76	69	59	46	32	23	52	31	22
BEVBert (An et al. 2023)	2.81	84	75	64	3.13	81	73	62	51	34	24	52	32	22
Lily (Lin et al. 2023)	2.48	84	77	72	3.05	82	74	68	48	32	23	45	30	21
ScaleVLN (Wang et al. 2023b)	2.34	87	79	70	2.73	83	77	68	57	-	-	56	-	-
SALI (Ours)	**1.92**	**86**	**82**	**78**	**2.08**	**83**	**79**	**74**	**58**	**38**	**28**	**56**	**34**	**25**

Table 1: Comparison with SoTA methods on R2R and REVERIE benchmarks.

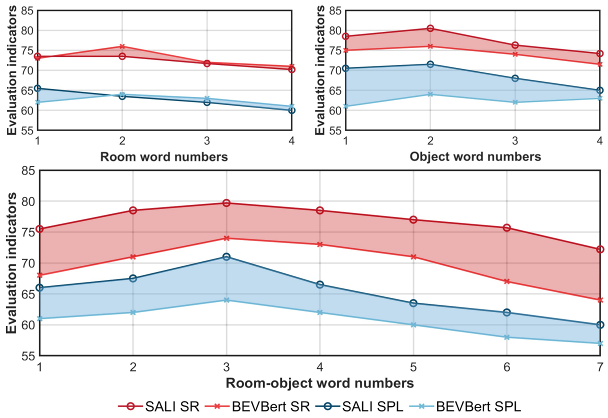


Figure 5: Comparison of SR and SPL between SALI and BEVBert models on different sub-datasets.

Implementation Details. We utilize ViT-B/16 (Dosovitskiy et al. 2020) and ResNet-50 (He et al. 2016b) to extract node features f_t, and adopt LXMERT (Tan and Bansal 2019) as our cross-modal transformer. During pre-training, we set the imagination history length K to 2. SALI is initially trained for 100k iterations with a batch size of 32 using a single Quadro RTX 8000 GPU, alongside training four imagination models. For fine-tuning, the trainable components include imagination node encoders, text encoders, and map encoders. Fine-tuning is conducted for 20k iterations with a batch size of 4 on four Quadro RTX 8000 GPUs. The best model checkpoint is selected based on SPL + SR.

4.2 Comparison with State-of-the-Art

R2R. Table 1 compares SALI's navigation performance with various state-of-the-art approaches on the R2R benchmark. SALI achieves superior performance across all metrics. Notably, the concurrent improvement in SR and SPL demonstrates that SALI effectively aligns scenes with instructions, resulting in more efficient navigation.

REVERIE. SALI also outperforms all previous models on the REVERIE benchmark, achieving the highest scores for RGS and RGSPL. This improvement is attributed to the imagination module, which enhances the agent's ability to recognize and associate objects within its environment.

4.3 Quantitative and Qualitative Analysis

SALI's episodic memory and simulation capabilities are evidenced by its ability to process complex environment-related commonsense, including associations between rooms, objects, and their spatial relationships.

Quantitative Study. We examine agents' navigation performance under instructions with complex object and room information. We divided the R2R validation unseen instruction set into:

- $\mathcal{S}_1$: Instructions with over two room-related terms and no object terms (e.g. "*Walk through **hallway** and towards **restroom**.*").
- $\mathcal{S}_2$: Instructions with over two object-related terms and no room terms (e.g. "*Turn left at the **oven** and past the **fridge**.*").
- $\mathcal{S}_3$: Complex instructions with over four combined room and object terms (e.g. "*Past the **door** next to the **TV** in the **living room**, then walk into the **kitchen**.*")

Performance on these instruction sets is illustrated in Figure 5. Compared to the prior best space-aware model (BEVBert, (An et al. 2023)), SALI achieves the largest performance improvement on $\mathcal{S}_3$ particularly in SR and SPL. SALI's ability to imagine spatial relationships allows it to effectively interpret intricate object-room associations, demonstrating improved navigation under complex instructions. This showcases SALI's commonsense reasoning capabilities, achieved through its imagination-memory mechanism, analogous to human episodic simulation memory.

Qualitative Analysis. Figure 6 visualizes the imagination mechanism's process of generating spatial and image information. On the left, the imagined images for goal positions at varying distances are compared with corresponding real images. The peak signal-to-noise ratio (PSNR) (Korhonen and You 2012) of the generated image and the Pearson correlation coefficient (Cohen et al. 2009) with the real image were calculated. The shaded curve region represents the mean and variance of pixel errors of the two images, which can be approximated by visualizing the Pearson correlation coefficient and PSNR. For spatial waypoint prediction, the predicted heat map m_t after the non-maximum suppression (NMS) operation (top right) is visualized to show neighboring navigable positions.

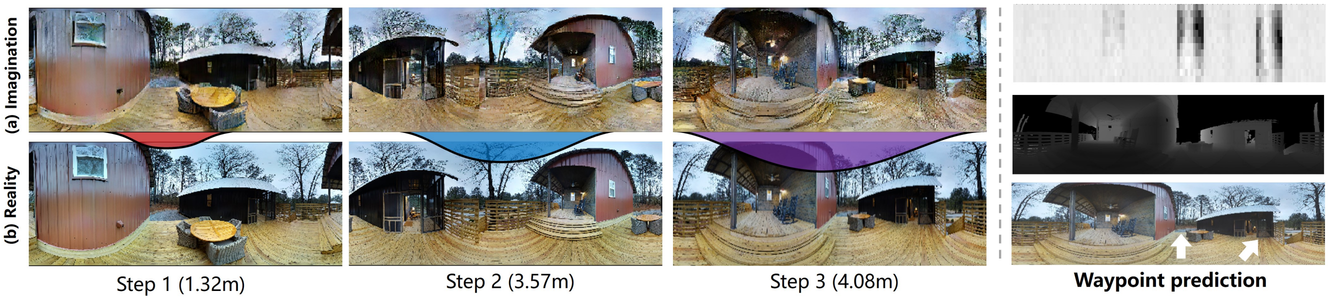


Figure 6: Imagined picture display and waypoint prediction schematic. The imagination error is visualized by the curve.

4.4 Ablation Study

Extensive experiments on the R2R val-unseen split were conducted to evaluate the proposed features.

Imagination vs. Reality. We first investigated the role of the hybrid mechanism between imagination and memory in improving navigation performance, as shown in Table 2. It indicates that the memory-based imaginary mechanism gets the SoTA performance, which is 12% and 9% higher than "reality only" in `SR` and `SPL`. At the same time, we find that navigation using only imagination is the least effective. We attribute this to the inability of the agent to effectively use and correlate the imagination results with its experience and trajectory. This suggests it's necessary to establish long-term memory mechanisms supporting imaginative abilities.

#	Memory Type	TL	NE↓	OSR↑	**SR↑**	**SPL↑**
1	**Reality**	12.59	3.39	78	70	61
2	**Imagination**	15.71	5.44	67	58	50
3	**Reality + Imagination**	10.34	1.92	86	82	70

Table 2: Ablation study of imagination-based memory.

Temporal and Spatial Imaginary. We investigated the impact of the spatial-temporal memory range on navigation performance, where the navigation output time step M and the upper node limit $\bar{N}$ define the memory range (Table 3), fixed input length $K = 2$. Results show that performance improves as the memory range expands. However, excessively large ranges ($M = 2, \bar{N} = 8$) lead to a decline in `SPL`, as agents at different positions tend to imagine the same destination, interfering with optimal decision-making. While navigation performance must be balanced with training time per epoch, inference time remains stable and does not significantly affect the results.

#	M	$\bar{N}$	OSR↑	**SR↑**	**SPL↑**	**Training Time**	**Inference Time**
1	0	0	78	70	61	0.54h	0.15h
2	1	4	82	76	67	0.74h	0.18h
3	2	4	86	82	71	1.32h	0.25h
4	2	8	84	82	68	2.51h	0.30h

Table 3: The effect of imagination range.

Auxiliary Model Effectiveness. The room type model and the waypoint model in the imagination model are auxiliary models that are not necessary for image generation. Waypoint predictions can be made by setting random directions instead. We examined the effect of the two models, as shown in Table 4. We found the agent containing the auxiliary model is optimal in the `SPL` metrics, reflecting that these two models are essential for decision-making.

#	Model Type	TL	NE↓	OSR↑	**SR↑**	**SPL↑**
1	**None**	12.59	3.39	78	70	61
2	**Room**	12.32	2.88	82	75	64
3	**Waypoint**	11.85	2.54	84	77	66
4	**Room + Waypoint**	10.34	1.92	86	80	71

Table 4: Ablation study of imagination auxiliary models.

Dynamic Decision-making. We evaluate the impact of combining imagined and real node embeddings on navigation performance using static and dynamic weighting strategies. Results show that dynamic weighting is more effective, as the weight of the imagination module decreases over time, reflecting reduced reliance on imagination in later navigation stages, analogous to human memory mechanisms.

#	Decisioning Weight	TL	NE↓	OSR↑	**SR↑**	**SPL↑**
1	**Dynamic**	10.34	1.92	86	80	71
2	**Fixed** ($\gamma_t = 0.5$)	12.33	2.25	82	76	67

Table 5: Ablation study of Decision Weight.

5 Conclusion

We propose SALI, an agent equipped with an imagination-integrated memory mechanism, inspired by human episodic memory and episodic simulation. Extensive experimental results demonstrate the effectiveness of the imagination module, enabling SALI to achieve SoTA performance in unseen environments. By leveraging its hybrid memory, SALI enhances robustness when navigating in complex environments. The future work will focus on optimizing the computational efficiency of the imagination process while preserving fine-grained results.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62303307, in part by National Key R&D Program of China under Grant No.2023YFB4705700, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102 and in part by the Fundamental Research Funds for the Central Universities.

References

An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; and Shao, J. 2023. Bevbert: Multimodal map pre-training for language-guided navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2737–2748.

An, D.; Wang, H.; Wang, W.; Wang, Z.; Huang, Y.; He, K.; and Wang, L. 2024. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Anderson, P.; Chang, A.; Chaplot, D. S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; et al. 2018a. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; and Van Den Hengel, A. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3674–3683.

Chaplot, D. S.; Salakhutdinov, R.; Gupta, A.; and Gupta, S. 2020. Neural topological slam for visual navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12875–12884.

Chen, S.; Guhur, P.-L.; Schmid, C.; and Laptev, I. 2021. History aware multimodal transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34.

Chen, S.; Guhur, P.-L.; Tapaswi, M.; Schmid, C.; and Laptev, I. 2022. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16537–16547.

Cohen, I.; Huang, Y.; Chen, J.; Benesty, J.; Benesty, J.; Chen, J.; Huang, Y.; and Cohen, I. 2009. Pearson correlation coefficient. *Noise Reduction in Speech Processing*, 1–4.

Cui, H.; Radosavljevic, V.; Chou, F.-C.; Lin, T.-H.; Nguyen, T.; Huang, T.-K.; Schneider, J.; and Djuric, N. 2019. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, 2090–2096. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Georgakis, G.; Schmeckpeper, K.; Wanchoo, K.; Dan, S.; Miltsakaki, E.; Roth, D.; and Daniilidis, K. 2022. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15460–15470.

Gomez, A.; Rousset, S.; and Baciu, M. 2009. Egocentric-updating during navigation facilitates episodic memory retrieval. *Acta Psychologica*, 132(3): 221–227.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016a. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016b. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hong, Y.; Wang, Z.; Wu, Q.; and Gould, S. 2022. Bridging the gap between learning in discrete and continuous environments for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15439–15449.

Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; and Gould, S. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, 1643–1653.

Hosang, J.; Benenson, R.; and Schiele, B. 2017. Learning non-maximum suppression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4507–4515.

Irshad, M. Z.; Mithun, N. C.; Seymour, Z.; Chiu, H.-P.; Samarasekera, S.; and Kumar, R. 2022. Semantically-aware spatio-temporal reasoning agent for vision-and-language navigation in continuous environments. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 4065–4071. IEEE.

Kaneko, T.; and Harada, T. 2021. Blur, noise, and compression robust generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13579–13589.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, 2. Minneapolis, Minnesota.

Koh, J. Y.; Lee, H.; Yang, Y.; Baldridge, J.; and Anderson, P. 2021. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14738–14748.

Korhonen, J.; and You, J. 2012. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth International Workshop on Quality of Multimedia Experience*, 37–38. IEEE.

Kühn, S.; and Gallinat, J. 2014. Segregating cognitive functions within hippocampal formation: A quantitative meta-analysis on spatial navigation and episodic memory. *Human Brain Mapping*, 35(4): 1129–1142.

Li, M.; Wang, Z.; Tuytelaars, T.; and Moens, M.-F. 2023. Layout-aware dreamer for embodied visual referring expression grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1386–1395.

Lin, C.; Jiang, Y.; Cai, J.; Qu, L.; Haffari, G.; and Yuan, Z. 2022. Multimodal transformer with variable-length memory for vision-and-language navigation. In *European Conference on Computer Vision*, 380–397. Springer.

Lin, K.; Chen, P.; Huang, D.; Li, T. H.; Tan, M.; and Gan, C. 2023. Learning vision-and-language navigation from youtube videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8317–8326.

Lin, X.; Li, G.; and Yu, Y. 2021. Scene-intuitive agent for remote embodied visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7036–7045.

Liu, R.; Wang, X.; Wang, W.; and Yang, Y. 2023. Bird's-eye-view scene graph for vision-language Navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10968–10980.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 32.

Mallya, A.; Wang, T.-C.; Sapra, K.; and Liu, M.-Y. 2020. World-consistent video-to-video synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, 359–378. Springer.

Moudgil, A.; Majumdar, A.; Agrawal, H.; Lee, S.; and Batra, D. 2021. Soat: A scene-and object-aware transformer for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 34.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346.

Phillip, I.; Jun-Yan, Z.; Tinghui, Z.; Alexei, A.; et al. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 3.

Qi, Y.; Pan, Z.; Hong, Y.; Yang, M.-H.; Van Den Hengel, A.; and Wu, Q. 2021. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1655–1664.

Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W. Y.; Shen, C.; and Hengel, A. v. d. 2020. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9982–9991.

Qiao, Y.; Qi, Y.; Hong, Y.; Yu, Z.; Wang, P.; and Wu, Q. 2023. Hop+: History-enhanced and order-aware pre-training for vision-and-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7): 8524–8537.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211–252.

Shah, D.; Osiński, B.; Levine, S.; et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on Robot Learning*, 492–504. PMLR.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Tang, H.; Yan, R.; and Tan, K. C. 2017. Cognitive navigation by neuro-inspired localization, mapping, and episodic memory. *IEEE Transactions on Cognitive and Developmental Systems*, 10(3): 751–761.

Temeltas, H.; and Kayak, D. 2008. SLAM for robot navigation. *IEEE Aerospace and Electronic Systems Magazine*, 23(12): 16–19.

Wang, H.; Liang, W.; Van Gool, L.; and Wang, W. 2023a. Dreamwalker: Mental planning for continuous vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10873–10883.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018a. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8798–8807.

Wang, X.; Xiong, W.; Wang, H.; and Wang, W. Y. 2018b. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 37–53.

Wang, Z.; Li, J.; Hong, Y.; Wang, Y.; Wu, Q.; Bansal, M.; Gould, S.; Tan, H.; and Qiao, Y. 2023b. Scaling data generation in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12009–12020.

Wang, Z.; Li, X.; Yang, J.; Liu, Y.; and Jiang, S. 2023c. Gridmm: Grid memory map for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15625–15636.