This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ProFocus: Proactive Perception and Focused Reasoning in Vision-and-Language Navigation

Wei Xue[1,2] Mingcheng Li[1,2] Xuecheng Wu[3] Jingqun Tang[4]
Dingkang Yang[1,2,†,§] Lihua Zhang[1,2,†]
[†]corresponding authors [§]project lead
[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]Fysics AI [3]Xi'an Jiaotong University [4]ByteDance
wxue24@m.fudan.edu.cn, {dkyang20, lihuazhang}@fudan.edu.cn

Abstract

*Vision-and-Language Navigation (VLN) requires agents to accurately perceive complex visual environments and reason over navigation instructions and histories. However, existing methods passively process redundant visual inputs and treat all historical contexts indiscriminately, resulting in inefficient perception and unfocused reasoning. To address these challenges, we propose **ProFocus**, a training-free progressive framework that unifies Proactive Perception and Focused Reasoning through collaboration between large language models (LLMs) and vision-language models (VLMs). For proactive perception, ProFocus transforms panoramic observations into structured ego-centric semantic maps, enabling the orchestration agent to identify missing visual information needed for reliable decision-making, and to generate targeted visual queries with corresponding focus regions that guide the perception agent to acquire the required observations. For focused reasoning, we propose Branch-Diverse Monte Carlo Tree Search (BD-MCTS) to identify top-k high-value waypoints from extensive historical candidates. The decision agent focuses reasoning on the historical contexts associated with these waypoints, rather than considering all historical waypoints equally. Extensive experiments validate the effectiveness of ProFocus, achieving state-of-the-art performance among zero-shot methods on R2R and REVERIE benchmarks.*

1. Introduction

Vision-and-Language Navigation (VLN) [2] enables agents to explore physical environments by following natural language instructions. Traditional training-based approaches [2, 8, 10], while achieving progress, remain data-intensive and environment-dependent [14]. The rapid development of large language models (LLMs) [24] and vision-language models (VLMs) [27] has enabled foundation model-based VLN [6, 53] through post-training adaptation [15, 49] or zero-shot prompting [5, 48, 55]. Despite these advances, current foundation model-based VLN methods [6, 48, 55] encounter two key limitations. (1) **Passive visual perception.** For VLM-driven research [5, 14, 52, 54], agents process panoramic or multi-view visual observations [29, 49, 50] whose redundancy inflates visual-token counts and causes attention diffusion across irrelevant features [1], obscuring instruction-critical fine-grained cues. (2) **Unfocused reasoning**. Both paradigms receive extensive historical contexts containing past observations and waypoints without prioritization. This hinders precise reasoning on high-value waypoints [5, 6, 54, 55]. Long trajectory histories dilute attention, preventing models from isolating key cues for accurate reasoning [20].

These *limitations* highlight two critical requirements: proactive acquisition of task-relevant visual observations to reduce perception redundancy, and focused reasoning over high-value waypoints within extensive historical contexts.

To address these challenges, we propose ProFocus, a training-free framework that unifies proactive perception and focused reasoning for efficient vision-and-language navigation. This is achieved through two core mechanisms, as illustrated in Figure 1. First, ProFocus introduces a reasoning-guided active perception mechanism to proactively acquire informative visual observations and assess the semantic values of navigable waypoints. It transforms panoramic observations into structured ego-centric semantic maps encoding object bounding boxes, depth, and directional relationships. Based on these maps, the orchestration agent (*i.e.*, LLM) performs spatial reasoning to infer missing visual cues and generate targeted visual queries with corresponding focused regions. The perception agent (*i.e.*, VLM) conducts fine-grained sensing exclusively within these regions. The orchestration agent itera-

tively evaluates information sufficiency and generates additional queries until adequate evidence is gathered, forming a closed perception-reasoning loop. Once sufficient information is acquired, the orchestration agent evaluates the semantic values of newly navigable waypoints to support subsequent focused reasoning. Furthermore, we design Branch-Diverse Monte Carlo Tree Search (BD-MCTS) to address spatial navigation uncertainty in unknown environments. Unlike standard MCTS that selects a single optimal action [4], BD-MCTS identifies top-k high-value candidates from extensive historical waypoints through path-aggregated value estimation, dynamic backpropagation, and branch-diversity constraints. The decision agent (*i.e.*, LLM) then performs focused reasoning over these filtered candidates, achieving efficient global navigation.

The main contributions are summarized as follows:

- A training-free progressive framework, ProFocus, that unifies proactive perception and focused reasoning for efficient vision-and-language navigation.
- A reasoning-guided proactive perception mechanism that establishes a closed perception-reasoning loop for iterative acquisition of instruction-relevant observations, rather than passively processing entire panoramic inputs.
- Branch-Diverse MCTS (BD-MCTS) that identifies top-k high-value waypoints from extensive historical candidates, guiding the decision agent to focus on the contexts associated with these waypoints rather than treating all historical contexts indiscriminately.

2. Related Work

2.1. Vision-and-Language Navigation

Benefiting from deep learning technologies [16, 19, 34, 35, 37–47], early methods enhanced VLN via learning [2, 7, 8, 10, 12, 56] and data augmentation [9, 31]. Traditional training-based methods, such as Speaker-Follower [9], EnvDrop [31], and PREVALENT [10], employed Seq2Seq [30] or Transformer-based architectures [32]. However, scarcity of instructions and limited scene diversity hindered robust navigation policies, resulting in poor generalization.

Foundation models have introduced two paradigms for VLN [6, 27, 29, 53]. The first centers on post-training adaptation [15, 21, 49, 52, 54], where models are fine-tuned on domain-specific datasets. For instance, NaviLLM [54] unifies diverse navigation tasks through schema-based instructions and multi-view fusion. These methods achieve strong in-domain results but require significant computation and struggle with generalization. The second paradigm explores training-free methods [5, 6, 22, 26, 55] exploiting zero-shot reasoning without task-specific tuning. NavGPT [55] converts panoramic scenes into textual descriptions for GPT-4-based action prediction, while AO-Planner [6] employs visual affordance prompts using SAM for waypoint planning.

2.2. Perception and Contextual Reasoning

Despite the remarkable progress brought by foundation model-based VLN paradigms [29, 50], these existing approaches [49, 55] still encounter two major challenges in perception and contextual reasoning. **(1) Perception.** Most existing methods [29, 49, 50] rely heavily on extensive visual inputs, such as panoramic observations [49, 55], multi-view images [29, 50, 54], or continuous video streams [51]. For example, NaviLLM [54] processes multi-view inputs via Vision Transformers and fuses them for holistic scene understanding. While such designs capture rich environmental information, they uniformly process dense visual tokens, resulting in visual token inflation. This passive perception paradigm suggests the need for proactive acquisition that selectively gathers instruction-relevant observations. **(2) Contextual Reasoning.** Existing methods often exhibit unfocused reasoning by treating historical contexts indiscriminately [6, 49, 55]. They typically store all past information in the form of historical frames [49, 54], textual descriptions [6, 55], or graph representations [5, 52]. Although approaches such as NavFoM's [50] token-based sampling and OpenNav's [26] multi-granularity dynamic memory (MGDM) attempt to manage memory, they still rely on recency-based heuristics or fixed policies, failing to dynamically assess the semantic importance of waypoints. For instance, NavFoM's [50] BATS allocates tokens based on temporal proximity rather than semantic relevance, hindering retrieval of critical waypoints during backtracking. These limitations motivate ProFocus, which introduces value-driven waypoint prioritization to enable focused reasoning over historical contexts.

3. Methodology

3.1. Overview

Problem formulation. We formalize VLN over a navigation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with waypoint set $\mathcal{V}$ and edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Given a natural language instruction $\mathcal{I} \in \mathcal{L}$, the agent navigates from an initial waypoint through $\mathcal{G}$ discovered online. At timestep t, the agent at waypoint v_t observes panoramic RGB image $\boldsymbol{\mathcal{O}}_t \in \mathbb{R}^{H \times W \times 3}$ and selects next waypoint v_{t+1} from navigable candidates $\mathcal{N}_t = \{v \in \mathcal{V} : (v_t, v) \in \mathcal{E}\}$. A memory bank $\mathcal{M} = \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_t\}$ stores multimodal contexts where each $\mathcal{S}_t$ records instruction, semantic map, and textual observations. A trajectory caption $\boldsymbol{\tau}_t$ constructed from $\mathcal{M}$ summarizes navigation history. Successful navigation requires perceiving $\boldsymbol{\mathcal{O}}_t$, reasoning over $\mathcal{I}$, $\mathcal{M}$ and $\boldsymbol{\tau}_t$, and selecting v_{t+1} from $\mathcal{N}_t$.

ProFocus framework. Rather than passively consuming predetermined visual inputs or uniformly treating all historical contexts, ProFocus addresses these challenges

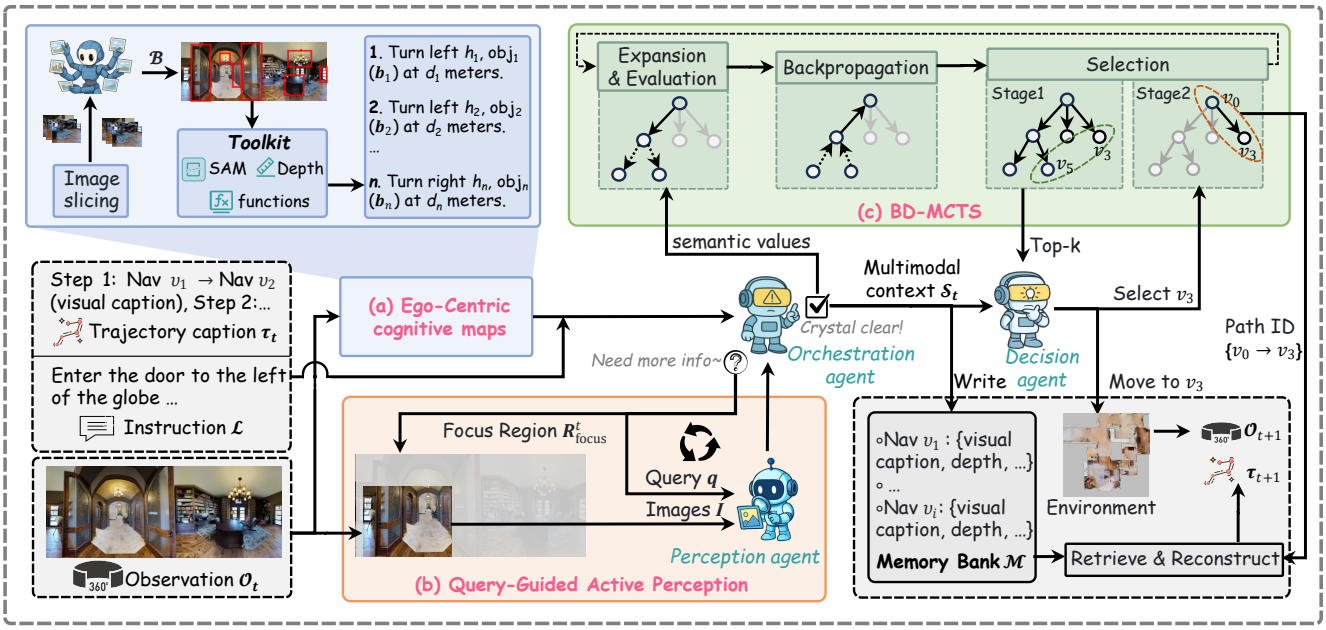


Figure 1. **Overview of the ProFocus framework.** ProFocus consists of two core mechanisms: *Proactive Perception*, which selectively acquires instruction-relevant visual evidence through a closed perception–reasoning loop, and *Focused Reasoning*, which leverages BD-MCTS to highlight top-k high-value waypoints for effective focused reasoning.

through a training-free framework unifying proactive perception and focused reasoning. The framework employs three specialized agents: orchestration agent $\mathcal{A}^{\boldsymbol{\theta}}_{\text{orch}}$ (LLM for spatial inference and semantic evaluation), perception agent $\mathcal{A}^{\boldsymbol{\phi}}_{\text{perc}}$ (VLM for fine-grained sensing), and decision agent $\mathcal{A}^{\boldsymbol{\psi}}_{\text{dec}}$ (LLM for reasoning over top-k candidates). As illustrated in Figure 1, ProFocus operates through two core mechanisms: Reasoning-Guided Proactive Perception (§3.2) transforms panoramic observations into structured semantic maps (Figure 1(a)) and establishes a closed perception-reasoning loop to iteratively acquire instruction-relevant observations (Figure 1(b)). Focused Reasoning via Branch-Diverse MCTS (§3.3) maintains a global search tree to identify top-k high-value waypoints from extensive historical contexts, enabling concentrated reasoning over strategically filtered candidates (Figure 1(c)).

3.2. Reasoning-Guided Proactive Perception

Current methods process token-heavy panoramic views obscuring fine-grained cues [49, 50] or rely on pre-captured descriptions omitting attributes [55]. Both exhibit passive perception, accepting predetermined inputs rather than actively querying instruction-relevant information. We establish a closed perception-reasoning loop enabling agents to determine *"what to perceive"* based on *"what is needed for decision-making."* This offers three advantages: (*i*) **reduced visual tokens** by focusing on instruction-relevant regions. (*ii*) **enhanced attribute recognition** through targeted queries capturing fine-grained details (*e.g.*, object color, texture, and spatial relations). (*iii*) **adaptive perception** adjusting to navigation requirements.

Ego-centric Semantic Map. As illustrated in Figure 1(a), given a panoramic observation $\mathcal{O}_t \in \mathbb{R}^{H\times W\times 3}$ at timestep t, we divide it into K directional views $\{\mathcal{I}_k\}_{k=1}^K$, each with overlapping fields of view. A vision-language model $F^{\text{scan}}_{\text{vlm}}: \mathbb{R}^{H\times W\times 3} \rightarrow \mathcal{P}(\mathbb{R}^4 \times \mathcal{V})$ is then used to process these views in parallel to detect all objects present in the scene:

$$\{(\boldsymbol{b}_i, obj_i)\}_{i=1}^{N_t} = F^{\text{scan}}_{\text{vlm}}(\{\mathcal{I}_k\}_{k=1}^K), \tag{1}$$

where $\boldsymbol{b}_i = (x_1, y_1, x_2, y_2) \in \mathbb{R}^4$ represents the bounding box of i, and $obj_i \in \mathcal{V}$ denotes its category. To obtain depth for objects, we use monocular depth estimation [3], $d_i \leftarrow f_{\text{depth}}(\mathcal{O}_t, \boldsymbol{m}_i)$, where $\boldsymbol{m}_i$ is the object mask obtained using SAM2 [28]. The heading angle of objects is computed as:

$$h_i = \pi \cdot \left(\frac{x_1 + x_2 - F}{F}\right) \in [-\pi, \pi], \tag{2}$$

where F is the field of view width. Finally, the ego-centric semantic map is constructed as:

$$\mathcal{C}_t = \{(h_i, obj_i(\boldsymbol{b}_i), d_i)\}_{i=1}^{N_t}, \tag{3}$$

which is formatted as natural language text, such as *"turn left/right h_i degrees, obj_i (bounding box $\boldsymbol{b}_i$) at d_i meters."* This structured map allows LLMs to reason about object locations (*e.g.*, "objects to the left"), enabling spatial understanding and focused region generation.

Reasoning-Driven Perception Loop. As shown in Figure 1(b), given the semantic map $\mathcal{C}_t$, trajectory history $\boldsymbol{\tau}_t$,

and instruction $\mathcal{I}$, the orchestration agent $\mathcal{A}^{\boldsymbol{\theta}}_{\text{orch}}$ initiates an active perception loop by generating a visual query $\boldsymbol{q} \in \mathcal{L}$ paired with a focused region $\boldsymbol{R}^t_{\text{focus}} \subseteq [0, H] \times [0, W]$:

$$(\boldsymbol{q}, \boldsymbol{R}^t_{\text{focus}}) = \mathcal{A}^{\boldsymbol{\theta}}_{\text{orch}}(\mathcal{C}_t, \boldsymbol{\tau}_t, \mathcal{I}, \mathcal{H}_{\text{query}}), \tag{4}$$

where $\mathcal{H}_{\text{query}} = \{(\boldsymbol{q}_j, \boldsymbol{a}^t_j)\}_{j=1}^{i-1}$ maintains the history of query-response pairs from previous iterations, preventing redundant queries. The perception agent $\mathcal{A}^{\boldsymbol{\phi}}_{\text{perc}}$ performs fine-grained analysis within the focused region $\boldsymbol{R}^t_{\text{focus}}$:

$$\boldsymbol{a}^t_i = \mathcal{A}^{\boldsymbol{\phi}}_{\text{perc}}(\boldsymbol{\mathcal{O}}_t|_{\boldsymbol{R}^t_{\text{focus}}}, \boldsymbol{q}), \tag{5}$$

where $\boldsymbol{\mathcal{O}}_t|_{\boldsymbol{R}^t_{\text{focus}}}$ denotes the cropped region of the panoramic observation $\boldsymbol{\mathcal{O}}_t$, and $\boldsymbol{a}^t_i \in \mathcal{L}$ represents the textual description of visual attributes extracted from the i-th query, including object color, texture, and spatial relationships. After each query and perception step, the orchestration agent evaluates whether the collected information is sufficient:

$$s_t = \mathcal{A}^{\boldsymbol{\theta}}_{\text{orch}}(\mathcal{C}_t, \mathcal{H}_{\text{query}}, \mathcal{I}), \quad s_t \in \{\text{sufficient}, \text{insufficient}\}. \tag{6}$$

If $s_t = \text{insufficient}$, the orchestration agent generates queries and focused regions for the perception agent, continuing the closed loop. This iterative cycle accumulates descriptions $\{\boldsymbol{a}^t_i\}_{i=1}^{n_t}$ from n_t iterations until the information is deemed sufficient. Once the perception loop terminates, the orchestration agent evaluates the semantic values of newly discovered waypoints $\mathcal{A}_{\text{new}}$ by leveraging all accumulated visual information:

$$V_{\text{sem}}(u) = \mathcal{A}^{\boldsymbol{\theta}}_{\text{orch}}(\mathcal{I}, \boldsymbol{\tau}_t, \{\boldsymbol{a}^t_i\}_{i=1}^{n_t}), \quad V_{\text{sem}}(u) \in [0, 1], \tag{7}$$

where $\{\boldsymbol{a}^t_i\}_{i=1}^{n_t}$ aggregates descriptions from perception loop, and $V_{\text{sem}}(u)$ quantifies how well waypoint u aligns with the navigation instruction $\mathcal{I}$ given trajectory history $\boldsymbol{\tau}_t$ and accumulated observations. Higher values indicate stronger semantic relevance to navigation goals, providing the foundation for value backpropagation in BD-MCTS (§3.3). Finally, the accumulated textual information is integrated with navigation instruction $\mathcal{I}$, trajectory history $\boldsymbol{\tau}_t$, and semantic map $\mathcal{C}_t$ to form the multimodal context $\mathcal{S}_t$:

$$\mathcal{S}_t = \{\mathcal{I}, \boldsymbol{\tau}_t, \mathcal{C}_t, \{\boldsymbol{a}^t_i\}_{i=1}^{n_t}\}, \tag{8}$$

where $\mathcal{C}_t$ encodes the semantic map with object locations. This unified textual representation $\mathcal{S}_t$ is stored in the memory bank $\mathcal{M}$ and enables the decision agent to perform global reasoning over complete navigation context.

3.3. Focused Reasoning via Branch-Diverse MCTS

As navigation progresses, the memory bank $\mathcal{M} = \{\mathcal{S}_0, \mathcal{S}_1, \ldots, \mathcal{S}_t\}$ accumulates extensive multimodal contexts. The decision agent $\mathcal{A}^{\boldsymbol{\psi}}_{\text{dec}}$ must reason over all accumulated contexts without prioritization, diffusing attention uniformly and leading to inefficient decision-making. Humans do not uniformly evaluate all past states when planning. Instead they *prune* low-value branches and use *prioritized memory access* to selectively replay task-relevant trajectories given current goals [13, 23]. Inspired by this, we propose Branch-Diverse MCTS (BD-MCTS), as illustrated in Figure 1(c). BD-MCTS filters a compact top-k set of high-value waypoints and iteratively refines their values with new observations. This enables the decision agent to focus reasoning on these candidates rather than diffusing attention across all historical contexts.

Tree-Graph Adaptation. MCTS assumes acyclic state spaces, but navigation graphs contain cycles. We maintain search tree $\mathcal{T} = \langle \boldsymbol{V}_{\mathcal{T}}, \boldsymbol{E}_{\mathcal{T}}, Q, N \rangle$ where $\boldsymbol{V}_{\mathcal{T}}$ are discovered waypoints, $\boldsymbol{E}_{\mathcal{T}}$ are explored edges, Q maps waypoints to estimated quality, and N tracks visit counts. When expanding at v_t, new waypoints are added as children while cross-branch waypoints are skipped to preserve acyclicity.

Phase I: Expansion with Semantic Evaluation. Classical MCTS expands new nodes and estimates their values through random rollouts. We replace rollouts with semantic value estimation. At current waypoint v_t, we discover new navigable candidates $\mathcal{A}_{\text{new}}$ and expand the tree:

$$\boldsymbol{V}_{\mathcal{T}} \leftarrow \boldsymbol{V}_{\mathcal{T}} \cup \mathcal{A}_{\text{new}}, \quad \boldsymbol{E}_{\mathcal{T}} \leftarrow \boldsymbol{E}_{\mathcal{T}} \cup \{(v_t, u) : u \in \mathcal{A}_{\text{new}}\}. \tag{9}$$

Each expanded waypoint $u \in \mathcal{A}_{\text{new}}$ is initialized with the semantic value $V_{\text{sem}}(u)$ computed via Eq. 7, which integrates instruction $\mathcal{I}$, trajectory history $\boldsymbol{\tau}_t$, and accumulated visual observations $\{\boldsymbol{a}^t_i\}_{i=1}^{n_t}$ from proactive perception. This provides value estimates grounded in multimodal context:

$$Q(u) \leftarrow V_{\text{sem}}(u), \quad N(u) \leftarrow 0, \quad \forall u \in \mathcal{A}_{\text{new}}. \tag{10}$$

This unified phase enables semantically grounded initialization while avoiding computational cost of rollouts.

Phase II: Backpropagation with Dynamic Refinement. Classical MCTS backpropagates rollout results to update ancestor statistics. We backpropagate semantic values to dynamic refinement. We compute reward R_t as the average semantic value of newly expanded waypoints when $|\mathcal{A}_{\text{new}}| > 0$, otherwise using current quality $Q(v_t)$. This reward propagates along root-to-current path $\boldsymbol{\pi}(r, v_t)$:

$$N(v) \leftarrow N(v) + 1; \quad Q(v) \leftarrow Q(v) + \frac{R_t - Q(v)}{N(v)}. \tag{11}$$

This asymmetric backpropagation updates only the selected path, creating priority shifts. Low reward triggers backtracking while high reward reinforces forward exploration.

Phase III: Top-k Selection with Branch Diversity. After iterative expansion and backpropagation cycles, the search tree $\mathcal{T}$ accumulates value estimates across all explored waypoints. We distill the extensive tree into a compact set of top-k high-value candidates to guide the decision agent's focused reasoning. For each leaf node $v \in \mathcal{L}$, we evaluate its comprehensive quality by aggregating backpropagated values along its root-to-leaf path $\boldsymbol{\pi}(v) = [v_0 = r, v_1, \ldots, v_L = v]$. A leaf's trustworthiness depends not only on its own semantic value but also on the reliability of the trajectory leading to it. We compute path-aggregated

value by weighting each ancestor's quality $Q(v_i)$ by its normalized visit count, giving higher weight to frequently visited nodes representing verified navigation segments while allowing newly discovered leaves to contribute through their initial semantic values. This balances exploitation of reliable paths with exploration of novel discoveries. However, ranking by aggregated value may yield spatially distant waypoints that are impractical to navigate. To incorporate physical reachability constraints, we apply a distance penalty based on shortest path length $d_{\mathcal{G}}(v_t, v)$ in graph $\mathcal{G}$, normalized by maximum distance. The final scoring function combines path value with distance penalty:

$$\text{Score}(v) = V_{\text{path}}(v) - \lambda \cdot \frac{d_{\mathcal{G}}(v_t, v)}{\max_{u \in \mathcal{L}} d_{\mathcal{G}}(v_t, u)}, \tag{12}$$

where $V_{\text{path}}(v)$ denotes the visit-count weighted path-aggregated value, and λ controls penalty strength. Finally, to ensure strategic diversity spanning different exploration directions, we select top-k leaves from $\mathcal{L}$ ranked by $\text{Score}(v)$, with the constraint that each parent node contributes at most 2 children to the candidate set $\mathcal{C}_k$.

The decision agent $\mathcal{A}^{\psi}_{\text{dec}}$ performs fine-grained reasoning over the filtered candidate set $\mathcal{P}_{new}$:

$$v_{t+1} = \mathcal{A}^{\psi}_{\text{dec}}\left(\mathcal{P}_{new}, \mathcal{S}_t, \mathcal{M}\right), \tag{13}$$

where detailed historical context for each candidate is retrieved from memory bank $\mathcal{M}$ by indexing waypoint identifiers along path $\boldsymbol{\pi}(r, v)$, and current multimodal context $\mathcal{S}_t$ provides instruction, trajectory, and observations. This two-stage mechanism balances systematic global exploration with focused decision-making. BD-MCTS performs coarse-grained filtering from extensive history to top-k candidates, followed by LLM performing nuanced contextual reasoning. This avoids both premature commitment and attention diffusion. After selecting v_{t+1}, the agent transitions to the new waypoint, updates memory $\mathcal{M} \leftarrow \mathcal{M} \cup \{\mathcal{S}_{t+1}\}$. It then reconstructs the trajectory caption $\boldsymbol{\tau}_{t+1} \leftarrow f_{\text{rec}}(\{v_0, \ldots, v_{t+1}\}, \mathcal{M})$, where f_{rec} is the retrieve-and-reconstruct function. The agent obtains the observation $\boldsymbol{\mathcal{O}}_{t+1}$ and repeats this cycle until reaching the target.

4. Experiments

4.1. Experimental Setup

Datasets. Experiments are conducted on two widely-used VLN benchmarks: R2R [2] and REVERIE [25]. R2R requires agents to follow detailed step-by-step instructions to reach target locations, while REVERIE combines navigation with object grounding, requiring agents to locate specific objects based on high-level descriptions. Following recent foundation model-based VLN works [5, 55], we focus solely on navigation for REVERIE. We evaluate whether the agent successfully reaches the target location, without assessing object grounding.

Evaluation Metrics. Standard evaluation metrics [2] are adopted: (1) **Success Rate (SR)**: percentage of episodes stopping within 3m of target. (2) **Oracle Success Rate (OSR)**: SR with oracle stop policy. (3) **Navigation Error (NE)**: average distance to target in meters. (4) **Success weighted by Path Length (SPL)**: SR penalized by path efficiency. Following the REVERIE benchmark convention [5], we report OSR, SR, and SPL but omit NE, as the task prioritizes object-centric navigation success over precise distance measurement.

Implementation Details. Our training-free framework consists of three specialized agents: an orchestration agent, a perception agent, and a decision agent, powered by LLM, VLM, and LLM respectively. We select state-of-the-art large language models Qwen3-Max [36] and DeepSeek-V3 [17], and vision-language models Qwen3-VL-Max [36] and GLM-4.5V [11] for evaluation. We denote these models as Q3, DS3, Q3VL, and GLM respectively for brevity. We evaluate two model configurations for ProFocus: (1) DeepSeek-V3 + GLM-4.5V. (2) Qwen3-Max + Qwen3-VL-Max. To enable fair comparison across different foundation models, following [5], we re-implement NavGPT with LLM backbones Q3 and DS3, and MapGPT with VLM backbones Q3VL and GLM on a randomly sampled validation subset under identical experimental settings. These re-implemented results are marked as † in Tables 1 and 2.

4.2. Performance Analysis and Comparisons

We analyze ProFocus's performance across four key dimensions: navigation capability, exploration-success trade-offs, long-trajectory robustness, and cross-dataset generalization.

Superior Navigation Capability. ProFocus achieves state-of-the-art zero-shot performance on both SR and SPL metrics (Tables 1 and 2). On R2R, ProFocus (Q3+Q3VL) reaches 52.5% SR and 39.8% SPL, substantially exceeding re-implemented baselines: NavGPT (Q3) at 47.0% SR and 38.4% SPL, MapGPT (Q3VL) at 32.0% SR and 28.7% SPL. The DS3+GLM configuration achieves 50.0% SR and 41.2% SPL, outperforming NavGPT (DS3) at 36.0% SR and 28.1% SPL, MapGPT (GLM) at 41.4% SR and 30.8% SPL. Navigation error remains competitive at 4.92m, comparable to NavGPT (Q3) at 4.82m. On REVERIE, ProFocus (Q3+Q3VL) achieves 40.0% SR and 24.8% SPL versus NavGPT (Q3) at 18.0% SR and 13.7% SPL, MapGPT (Q3VL) at 22.5% SR and 18.0% SPL. ProFocus (DS3+GLM) reaches 36.9% SR and 25.9% SPL versus MapGPT (GLM) at 37.1% SR and 24.7% SPL. The consistent improvements across SR and SPL demonstrate ProFocus's effectiveness in accurate navigation.

Balanced Exploration and Success. The OSR-SR gap exhibits task-dependent characteristics. In most standard

Table 1. Results on R2R validation unseen set. † denotes our re-implementation with different foundation models. **Bold** best, underline second-best *within each method group*.

Methods	NE↓	OSR↑	SR↑	SPL↑
Training-based VLN specialists				
Seq2Seq [2]	7.81	28	21	–
Speaker-Follower [9]	6.62	**45**	36	–
EnvDrop [31]	**5.22**	–	**52**	**48**
Pretraining-based VLN specialists				
PREVALENT [10]	4.71	–	58	53
HAMT [7]	**2.29**	–	66	**61**
DUET [8]	3.31	**81**	**72**	60
NavCoT [15]	6.26	48	40	37
NaviLLM [54]	3.51	–	67	59
Zero-shot VLN with Foundation Models				
NavGPT (GPT-4) [55]	6.46	42.0	34.0	29.0
MapGPT (GPT-4) [5]	6.29	57.6	38.8	25.8
MapGPT (GPT-4V) [5]	5.63	57.6	43.7	34.8
DiscussNav (GPT-4) [22]	5.32	61.0	43.0	40.0
MSNav (GPT-4o) [18]	5.24	65.0	46.0	40.0
NavGPT† (DS3)	6.52	50.0	36.0	28.1
NavGPT† (Q3)	**4.82**	57.5	47.0	38.4
MapGPT† (GLM)	5.00	**70.7**	41.4	30.8
MapGPT† (Q3VL)	5.27	47.5	32.0	28.7
Ours† (DS3+GLM)	5.21	63.0	50.0	**41.2**
Ours† (Q3+Q3VL)	4.92	65.0	**52.5**	39.8

navigation scenarios, ProFocus demonstrates highly effective stopping mechanisms. On R2R, MapGPT (GLM) creates a 29.3% gap between OSR and SR. ProFocus (Q3+Q3VL) significantly narrows this difference to 12.5%. This shows our agent efficiently identifies goal arrival to prevent over-exploration. However, object-grounding tasks like REVERIE introduce inherent localization uncertainty. This specific requirement alters the OSR-SR dynamics. The gap may fluctuate relative to certain baselines in these specific cases. Despite these minor fluctuations, ProFocus maintains superior overall accuracy. On REVERIE, ProFocus (Q3+Q3VL) achieves a 40.0% SR. ProFocus (DS3+GLM) reaches a 36.9% SR. These results prove that ProFocus robustly converts exploration into high navigation success across diverse tasks.

Long-Trajectory Robustness. To evaluate robustness in challenging scenarios, we analyze the 30 longest navigation episodes from R2R validation unseen set (Figure 2). These long trajectories amplify decision complexity, requiring agents to maintain coherent reasoning over dozens of historical waypoints. ProFocus demonstrates advantages in these conditions. ProFocus (Q3+Q3VL) achieves 50.0% SR, exceeding MapGPT (GLM) at 40.0%. NavGPT variants show weaker performance at 33.3% and 22.2% SR. ProFocus (DS3+GLM) reaches 46.7% SR, consistently outperforming baselines. The performance advantage in long trajectories validates ProFocus's capability to effectively handle extensive historical contexts without performance degradation as trajectory complexity increases.

Table 2. Results on REVERIE validation unseen set. † denotes our re-implementation with different foundation models. **Bold** best, underline second-best *within each method group*.

Methods	OSR↑	SR↑	SPL↑
Training-based VLN specialists			
Seq2Seq [2]	8.1	4.2	2.8
FAST-MATTN [25]	28.2	14.4	7.2
NavQ [33]	**62.0**	**54.1**	**39.2**
Pretraining-based VLN specialists			
HAMT [7]	35.4	31.6	29.6
DUET [8]	50.0	**45.8**	35.3
NaviLLM [54]	**53.7**	44.6	**36.6**
Zero-shot VLN with Foundation Models			
NavGPT (GPT-4) [55]	28.3	19.2	14.6
MapGPT (GPT-4) [5]	42.6	28.4	14.5
MapGPT (GPT-4V) [5]	36.8	31.6	20.3
NavGPT† (DS3)	34.0	21.0	14.3
NavGPT† (Q3)	24.0	18.0	13.7
MapGPT† (GLM)	50.5	37.1	24.7
MapGPT† (Q3VL)	36.5	22.5	18.0
Ours† (DS3+GLM)	**57.1**	36.9	**25.9**
Ours† (Q3+Q3VL)	51.7	**40.0**	24.8

Cross-Dataset and Cross-Model Consistency. ProFocus maintains robust performance across diverse settings. Both configurations consistently outperform corresponding baselines on R2R (detailed step-by-step instructions) and REVERIE (object-centric navigation). ProFocus (Q3+Q3VL) achieves the highest SR on both datasets: 52.5% on R2R and 40.0% on REVERIE among re-implemented methods. ProFocus (DS3+GLM) reaches 50.0% SR on R2R and 36.9% SR on REVERIE, maintaining competitive performance. ProFocus demonstrates consistent advantages across different foundation model combinations, substantially exceeding corresponding baselines on both datasets. This stability demonstrates that ProFocus's performance gains are consistent across different foundation model capabilities and dataset characteristics.

4.3. Ablation Studies

We evaluate two ablation variants under two foundation model configurations (Tables 3 and 4): (*i*) ProFocus w/o BD-MCTS removes the focused reasoning mechanism, requiring the decision agent to reason over all historical waypoints without top-k filtering. (*ii*) ProFocus w/o PP removes the proactive perception mechanism, processing full

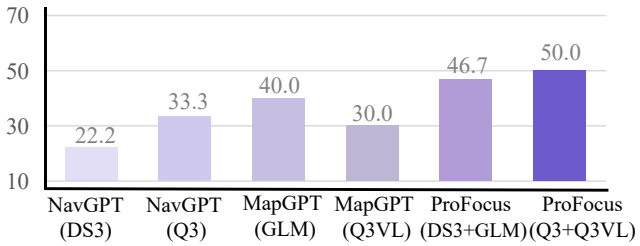


Figure 2. Performance comparison on the 30 longest navigation episodes from R2R validation unseen set.

Table 3. Ablation study on R2R validation unseen set. BD-MCTS: Branch-Diverse MCTS for focused reasoning, PP: Proactive Perception mechanism.

Methods	NE ↓	OSR ↑	SR ↑	SPL ↑
DS3+GLM				
ProFocus	5.21	63.0	50.0	41.2
ProFocus w/o BD-MCTS	5.75	54.2	46.9	40.0
ProFocus w/o PP	6.45	50.0	42.0	30.0
Q3+Q3VL				
ProFocus	4.92	65.0	52.0	39.8
ProFocus w/o BD-MCTS	6.19	53.66	50.0	38.4
ProFocus w/o PP	5.97	54.0	48.0	34.4

Table 4. Ablation study on REVERIE validation unseen set. BD-MCTS: Branch-Diverse MCTS for focused reasoning, PP: Proactive Perception mechanism.

Methods	OSR ↑	SR ↑	SPL ↑
DS3+GLM			
ProFocus	57.14	36.90	25.85
ProFocus w/o BD-MCTS	42.67	33.33	26.73
ProFocus w/o PP	36.00	32.50	23.05
Q3+Q3VL			
ProFocus	51.67	40.00	24.83
ProFocus w/o BD-MCTS	48.89	37.78	27.37
ProFocus w/o PP	34.00	30.00	18.51

panoramic views without targeted visual queries.

Impact of BD-MCTS on Exploration Capability. Removing BD-MCTS causes OSR degradation across both datasets. On R2R (Table 3), OSR drops from 63.0% to 54.2% for DS3+GLM and from 65.0% to 53.66% for Q3+Q3VL. On REVERIE (Table 4), the impact is even more pronounced: OSR decreases from 57.14% to 42.67% for DS3+GLM and from 51.67% to 48.89% for Q3+Q3VL. In contrast, SR degradation is modest. The disproportionate OSR versus SR degradation suggests that without BD-MCTS, agents explore less effectively. They fail to discover target-proximate waypoints even when navigation succeeds. Navigation error increases substantially on R2R (4.92m to 6.19m for Q3+Q3VL), validating the importance of systematic historical context filtering for spatial reasoning.

Impact of Proactive Perception on Path Efficiency. Removing proactive perception causes severe performance degradation across all metrics. SPL losses are particularly dramatic. On R2R (Table 3), SPL drops from 41.2% to 30.0% for DS3+GLM and from 39.8% to 34.4% for Q3+Q3VL. SR decreases from 50.0% to 42.0% for DS3+GLM and from 52.0% to 48.0% for Q3+Q3VL. On REVERIE (Table 4), OSR drops dramatically from 51.67% to 34.0% for Q3+Q3VL and from 57.14% to 36.0% for DS3+GLM. The substantial OSR and SPL degradation indicates that without targeted visual queries, agents struggle to acquire instruction-relevant spatial details. This leads to inefficient exploration and poor navigation decisions.

Task-Specific Sensitivity. The two mechanisms exhibit different sensitivities across R2R and REVERIE. BD-MCTS removal causes greater OSR loss on REVERIE than R2R. For DS3+GLM, REVERIE shows 14.47% decline versus 8.8% on R2R. This suggests REVERIE's object-centric navigation demands more effective historical context management. Conversely, proactive perception removal causes greater SPL loss on R2R than REVERIE. For DS3+GLM, R2R shows 11.2% absolute decline in SPL. This indicates R2R's detailed step-by-step instructions require more precise visual attribute acquisition for path efficiency. An interesting phenomenon emerges on REVERIE: removing BD-MCTS slightly increases SPL from 25.85% to 26.73% for DS3+GLM and from 24.83% to 27.37% for Q3+Q3VL, while reducing OSR and SR. This suggests that without top-k filtering, agents adopt more conservative, shorter paths at the cost of exploration breadth and success rate.

Cross-Model Consistency. Both mechanisms demonstrate consistent contributions across foundation model combinations. Removing BD-MCTS consistently reduces OSR across configurations and datasets. Removing proactive perception causes substantial losses in both OSR and SPL. On REVERIE, Q3+Q3VL shows greater sensitivity to proactive perception removal with OSR dropping 17.67% compared to 21.14% for DS3+GLM. Both configurations exhibit comparable degradation on R2R. This cross-model consistency validates that both mechanisms provide fundamental architectural improvements rather than compensating for specific foundation model weaknesses.

4.4. Qualitative Analysis

Figure 3 (left) illustrates the advantage of proactive perception in semantic value estimation. Without proactive perception (bottom, yellow), the agent processes the full panoramic view and identifies that waypoint 1 leads to a corridor. It assigns the highest value of 0.9 to waypoint 1 and only 0.2 to waypoint 2. However, this estimation is

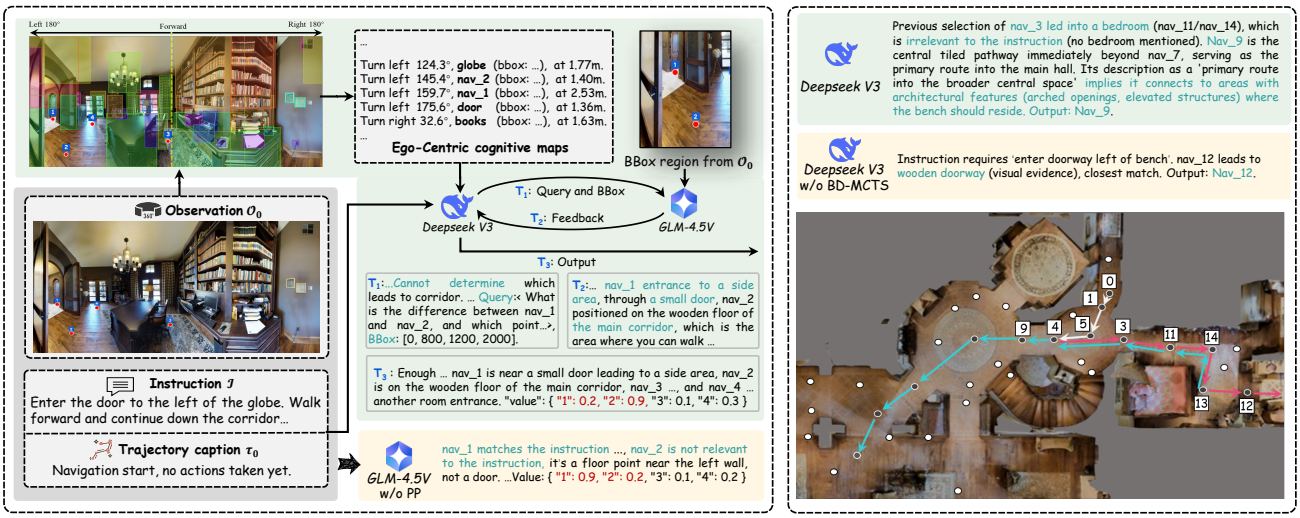


Figure 3. **Qualitative case study demonstrating proactive perception and focused reasoning.** *Left*: Proactive perception enables accurate semantic value estimation by acquiring fine-grained visual details. *Right*: BD-MCTS-guided top-k focused reasoning enables error correction through global historical awareness.

incorrect. The passive perception misses the fine-grained visual details that waypoint 1 actually connects to a small doorway leading to a different room, while waypoint 2 is the correct corridor path. With proactive perception (top, green), the orchestration agent identifies missing information and generates a targeted visual query with focused regions. The perception agent examines waypoints 1 and 2 in detail. It discovers that waypoint 1 leads to a small doorway (side room) while waypoint 2 provides better corridor access. Value estimates are corrected: waypoint 1 drops to 0.2 and waypoint 2 rises to 0.9. This prevents a critical navigation error. This demonstrates that reasoning-guided queries and focused perception acquire instruction-relevant attributes missed by passive processing, enabling accurate semantic value estimation for reliable waypoint selection.

Figure 3 (right) demonstrates the advantage of BD-MCTS-guided top-k focused reasoning in error correction. The navigation instruction requires sequential progression: approach the clock, turn right through a doorway, enter the doorway left of the bench, pass through the wooden entry, and stop at the table. The agent previously made an incorrect decision at nav_3, leading to a bedroom area (nav_11/nav_14) that deviates from the correct path. Nav_9 represents the central tiled pathway beyond nav_7, serving as the primary route into the main hall. With BD-MCTS (top, green), the decision agent receives top-k candidates including nav_9 and focuses reasoning on its associated historical context. By analyzing nav_9's visual information describing it as "the primary route into the broader central space implying connection to areas with architectural features," the agent recognizes the path deviation and corrects the error by backtracking from nav_13 to nav_9. Without BD-MCTS (bottom, yellow), the agent lacks top-k guidance and processes all historical candidates uniformly. At nav_12, it observes a "wooden doorway" and directly matches it with the instruction's later-stage landmark "entry way made out of wood," ignoring that the current location (bedroom area) is irrelevant to the instruction stage (main hall area). This superficial visual matching causes the agent to continue on the incorrect path without reflecting on historical decisions. The case validates that top-k focused reasoning enables global historical awareness and path correction, while undirected reasoning over all candidates leads to myopic reactive behavior that matches current observations to instruction fragments without considering path validity or navigation stage consistency.

5. Conclusion

This paper presents ProFocus, a training-free progressive framework for vision-and-language navigation. By unifying proactive perception and focused reasoning, ProFocus mitigates the issues of perception redundancy and unfocused reasoning in foundation model-based VLN. Experimental results on the R2R and REVERIE benchmarks demonstrate ProFocus's competitive zero-shot performance, with significant improvements in navigation success rates and path efficiency.

The training-free design facilitates easy deployment in practical applications. Future work will extend ProFocus to long-horizon, multi-goal tasks in complex environments, such as robotic assistance for individuals with disabilities, where focused reasoning and proactive perception hold significant potential for enhancing efficiency and adaptability.

References

[1] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29915–29926, 2025. 1

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2018. 1, 2, 5, 6

[3] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. 3

[4] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012. 2

[5] Jiaqi Chen, Bingqian Lin, Ran Xu, Zhenhua Chai, Xiaodan Liang, and Kwan-Yee K. Wong. MapGPT: Map-guided prompting with adaptive path planning for vision-and-language navigation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 1, 2, 5, 6

[6] Jiaqi Chen, Bingqian Lin, Xinmin Liu, Lin Ma, Xiaodan Liang, and Kwan-Yee K Wong. Affordances-oriented planning using foundation models for continuous vision-language navigation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 23568–23576, 2025. 1, 2

[7] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5834–5847, 2021. 2, 6

[8] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. 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 (CVPR)*, pages 16537–16547, 2022. 1, 2, 6

[9] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2, 6

[10] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13137–13146, 2020. 1, 2, 6

[11] Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025. 5

[12] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. VLN BERT: A recurrent vision-and-language BERT for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2021. 2

[13] Quentin J. M. Huys, Neir Eshel, Elizabeth O'Nions, Luke Sheridan, Peter Dayan, and Jonathan P. Roiser. Bonsai trees in your head: How the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLOS Computational Biology*, 8(3):e1002410, 2012. 4

[14] Bingqian Lin, Yunshuang Nie, Khun Loun Zai, Ziming Wei, Mingfei Han, Rongtao Xu, Minzhe Niu, Jianhua Han, Liang Lin, Cewu Lu, et al. Evolvenav: Self-improving embodied reasoning for llm-based vision-language navigation. *arXiv e-prints*, pages arXiv–2506, 2025. 1

[15] Bingqian Lin, Yunshuang Nie, Ziming Wei, Jiaqi Chen, Shikui Ma, Jianhua Han, Hang Xu, Xiaojun Chang, and Xiaodan Liang. NavCoT: Boosting LLM-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2025. 1, 2, 6

[16] Lin Lin, Jiefeng Long, Zhihe Wan, Yuchi Wang, Dingkang Yang, Shuang Yang, Yueyang Yao, Xu Chen, Zirui Guo, Shengqiang Li, et al. Sail-embedding technical report: Omni-modal embedding foundation model. *arXiv preprint arXiv:2510.12709*, 2025. 2

[17] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 5

[18] Chenghao Liu, Zhimu Zhou, Jiachen Zhang, Minghao Zhang, Songfang Huang, and Huiling Duan. MSNav: Zero-shot vision-and-language navigation with dynamic memory and LLM spatial reasoning. *arXiv preprint arXiv:2508.16654*, 2025. 6

[19] Keliang Liu, Dingkang Yang, Ziyun Qian, Weijie Yin, Yuchi Wang, Hongsheng Li, Jun Liu, Peng Zhai, Yang Liu, and Lihua Zhang. Reinforcement learning meets large language models: A survey of advancements and applications across the llm lifecycle. *arXiv preprint arXiv:2509.16679*, 2025. 2

[20] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024. 1

[21] Qingxiang Liu, Ting Huang, Zeyu Zhang, and Hao Tang. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*, 2025. 2

[22] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: Visual language navigation via

multi-expert discussions. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 17380–17387, 2024. 2, 6

[23] Marcelo G. Mattar and Nathaniel D. Daw. Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, 21(11):1609–1617, 2018. 4

[24] Wenying Niu, Hongru Ma, and Yue Zhai. Research on task planning methods for embodied intelligence based on large language models. In *Proceedings of the IEEE International Symposium on Intelligent Robotics and Systems (ISoIRS)*, pages 1–5, 2025. 1

[25] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9982–9991, 2020. 5, 6

[26] Yanyuan Qiao, Wenqi Lyu, Hui Wang, Zixu Wang, Zerui Li, Yuan Zhang, Mingkui Tan, and Qi Wu. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 6710–6717, 2025. 2

[27] Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. A survey of multilingual large language models. *Patterns*, 6 (1), 2025. 1, 2

[28] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[29] Xinshuai Song, Weixing Chen, Yang Liu, Weikai Chen, Guanbin Li, and Liang Lin. Towards long-horizon vision-language navigation: Platform, benchmark and method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12078–12088, 2025. 1, 2

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2

[31] Hao Tan, Licheng Yu, and Mohit Bansal. Learning to navigate unseen environments: Back translation with environmental dropout. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2610–2621, 2019. 2, 6

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2

[33] Peiran Xu, Xicheng Gong, and Yadong Mu. NavQ: Learning a Q-model for foresighted vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6327–6341, 2025. 6

[34] Wei Xue and Hong He. A progressive learning classifier based on dynamic differential weighted network for feature identification of brain network series. *Knowledge-Based Systems*, 274:110661, 2023. 2

[35] Wei Xue, Hong He, Yanbing Wang, and Ying Zhao. Sagn: Sparse adaptive gated graph neural network with graph regularization for identifying dual-view brain networks. *IEEE Transactions on Neural Networks and Learning Systems*, 36 (5):8085–8099, 2024. 2

[36] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 5

[37] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, pages 1642–1651, 2022. 2

[38] Dingkang Yang, Shuai Huang, Shunli Wang, Yang Liu, Peng Zhai, Liuzhen Su, Mingcheng Li, and Lihua Zhang. Emotion recognition for multiple context awareness. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 144–162, 2022.

[39] Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, Xiao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, and Lihua Zhang. Context de-confounded emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19005–19015, 2023.

[40] Dingkang Yang, Mingcheng Li, Linhao Qu, Kun Yang, Peng Zhai, Song Wang, and Lihua Zhang. Asynchronous multimodal video sequence fusion via learning modality-exclusive and-agnostic representations. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[41] Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. Towards multimodal sentiment analysis debiasing via bias purification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[42] Dingkang Yang, Jinjie Wei, Dongling Xiao, Shunli Wang, Tong Wu, Gang Li, Mingcheng Li, Shuaibing Wang, Jiawei Chen, Yue Jiang, et al. Pediatricsgpt: Large language models as chinese medical assistants for pediatric applications. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:138632–138662, 2024.

[43] Dingkang Yang, Kun Yang, Haopeng Kuang, Zhaoyu Chen, Yuzheng Wang, and Lihua Zhang. Towards context-aware emotion recognition debiasing from a causal demystification perspective via de-confounded training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[44] Dingkang Yang, Kun Yang, Mingcheng Li, Shunli Wang, Shuaibing Wang, and Lihua Zhang. Robust emotion recognition in context debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12447–12457, 2024.

[45] Dingkang Yang, Mingcheng Li, Xuecheng Wu, Zhaoyu Chen, Kaixun Jiang, Keliang Liu, Peng Zhai, and Lihua Zhang. Improving multimodal sentiment analysis via modality optimization and dynamic primary modality selection. *arXiv preprint arXiv:2511.06328*, 2025.

[46] Dingkang Yang, Jinjie Wei, Mingcheng Li, Jiyao Liu, Lihao Liu, Ming Hu, Junjun He, Yakun Ju, Wei Zhou, Yang Liu, et al. Medaide: Information fusion and anatomy of medical intents via llm-based agent collaboration. *Information Fusion*, page 103743, 2025.

[47] Dingkang Yang, Dongling Xiao, Jinjie Wei, Mingcheng Li, Zhaoyu Chen, Ke Li, and Lihua Zhang. Improving factuality in large language models via decoding-time hallucinatory and truthful comparators. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 25606–25614, 2025. 2

[48] Hang Yin, Xiuwei Xu, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. UniGoal: Towards universal zero-shot goal-oriented navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19057–19066, 2025. 1

[49] Zhuoyuan Yu, Yuxing Long, Zihan Yang, Chengyan Zeng, Hongwei Fan, Jiyao Zhang, and Hao Dong. Correctnav: Self-correction flywheel empowers vision-language-action navigation model. In *Proceedings of the AAAI Conference on Artificial Intelligence(AAAI)*, pages 18737–18745, 2026. 1, 2, 3

[50] Jiazhao Zhang, Anqi Li, Yunpeng Qi, Minghan Li, Jiahang Liu, Shaoan Wang, Haoran Liu, Gengze Zhou, Yuze Wu, Xingxing Li, et al. Embodied navigation foundation model. *arXiv preprint arXiv:2509.12129*, 2025. 1, 2, 3

[51] Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan champion Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Proceedings of the Robotics: Science and Systems (RSS)*, 2025. 2

[52] Pingrui Zhang, Yifei Su, Pengyuan Wu, Dong An, Li Zhang, Zhigang Wang, Dong Wang, Yan Ding, Bin Zhao, and Xuelong Li. Cross from left to right brain: Adaptive text dreamer for vision-and-language navigation. *arXiv preprint arXiv:2505.20897*, 2025. 1, 2

[53] Yue Zhang, Ziqiao Ma, Jialu Li, Yanyuan Qiao, Zun Wang, Joyce Chai, Qi Wu, Mohit Bansal, and Parisa Kordjamshidi. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *Transactions on Machine Learning Research (TMLR)*, 2024. 1, 2

[54] Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13624–13634, 2024. 1, 2, 6

[55] Gengze Zhou, Yicong Hong, and Qi Wu. NavGPT: Explicit reasoning in vision-and-language navigation with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7641–7649, 2024. 1, 2, 3, 5, 6

[56] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10012–10022, 2020. 2