Article

Not peer-reviewed version

From Instruction Following to Cognitive Navigation: A Survey on the Evolution of Vision-and-Language Navigation

Kailin Lyu [†,‡], Kangyi Wu [†], Pengna Li [†], Wenxuan Song, Di Wu, Jianwei He, Junting Chen, Ning Yang, Zebin Han, Kaiwen Luo, Liang Lin, Long Xiao, Xi Lin, Weigang Xue, Yongen Zhao, Kaiwen Xue, Zhiqiang Yuan, Yang Liu, WeiZhong Huang, Liwei Yang, Jinwei He, Nanxing Hu, Jinjun Wang, Ye Deng, Shaoqing Xu, Lin Zhao, Shengqian Qin, Fan Xu, Qingyi Si, Keji He, Jiaqi Peng, Hao Wu, Hao Chen, Xiaoyu Ma, Zhenghong Zhou, Zeqin Liao, Qiankun Li, Kun Wang, Ce Hao [*], Lianyu Hu [*], Dongrui Liu, Chuang Zhu, Yonggang Qi, Xingjun Ma, Hao Chen, Qing Guo, Weinan Zhang, Qi Li, Shanghang Zhang, Zhigang Zeng, Chunhua Shen, Yu-Gang Jiang, Tieniu Tan, Yang Liu

Posted Date: 30 June 2026

doi: 10.20944/preprints202606.2231.v1

Keywords: vision-and-language navigation; embodied intelligence; paradigm evolution; foundation models; open-world generalization

 doi:10.20944/preprints202606.2231.v1

Article

From Instruction Following to Cognitive Navigation: A Survey on the Evolution of Vision-and-Language Navigation

Kailin Lyu [1,2,†,‡], Kangyi Wu [3,†], Pengna Li [3,†], Wenxuan Song [4], Di Wu [1], Jianwei He [1], Junting Chen [5], Ning Yang [1,6], Zebin Han [7], Kaiwen Luo [8], Liang Lin [9], Long Xiao [1], Xi Lin [10], Weigang Xue [2], Yongen Zhao [2], Kaiwen Xue [11], Zhiqiang Yuan [12], Yang Liu [12], WeiZhong Huang [12], Liwei Yang [12], Jinwei He [12], Nanxing Hu [12], Jinjun Wang [3,13], Ye Deng [14], Shaoqing Xu [15], Lin Zhao [16], Shengqian Qin [17], Fan Xu [18], Qingyi Si [19], Keji He [20], Jiaqi Peng [21], Hao Wu [21], Hao Chen [11], Xiaoyu Ma [8], Zhenghong Zhou [8], Zeqin Liao [8], Qiankun Li [8], Kun Wang [8], Ce Hao [2,5,*], Lianyu Hu [8,*], Dongrui Liu [22], Chuang Zhu [11], Yonggang Qi [11], Xingjun Ma [23], Hao Chen [24], Qing Guo [25], Weinan Zhang [17,22], Qi Li [1], Shanghang Zhang [11], Zhigang Zeng [26], Chunhua Shen [24], Yu-Gang Jiang [23], Tieniu Tan [1] and Yang Liu [8]

[1] CASIA
[2] ZGCA
[3] XJTU
[4] HKUST(GZ)
[5] NUS
[6] NJU
[7] SEU
[8] NTU
[9] UCAS-IIE
[10] Johns Hopkins University
[11] BUPT
[12] Tencent
[13] Xpeng Inc
[14] SWUFE
[15] University of Macau
[16] BIT
[17] SJTU
[18] USTC
[19] JD Explore Academy
[20] SDU
[21] THU
[22] Shanghai AI Lab
[23] FDU
[24] ZJU
[25] NKU
[26] HUST
* Correspondence: cehao@berkeley.edu (C.H.); lianyu.hu@ntu.edu.sg (l.H.)
† These authors contributed equally to this work.
‡ Project leader.

Abstract

Vision-and-Language Navigation (VLN) requires embodied agents to ground natural language instructions in visual perception and make navigation decisions in complex 3D environments, making it a central problem in embodied artificial intelligence. Since the introduction of the Room-to-Room (R2R) benchmark, VLN has made substantial progress. In recent years, as research settings have gradually expanded from closed and single indoor benchmark scenarios to open-world environments, the field has undergone a profound paradigm shift from passive instruction following on fixed benchmarks to autonomous cognitive navigation in open-world settings. However, existing surveys mainly organize prior work according to technical taxonomies, lacking a systematic characterization of this paradigm evolution. To address this gap, this survey proposes an evolution-centered unified analytical framework that reviews contemporary VLN research across four progressive layers: perception, cognition,

 doi:10.20944/preprints202606.2231.v1

learning, and generalization. It reveals the intrinsic connections and evolutionary logic among different technical lines, identifies key open challenges at each dimension, and outlines future research directions. This survey aims to provide VLN researchers with a clear panoramic view of capability evolution, while offering the broader embodied intelligence community a systematic roadmap from closed-benchmark evaluation toward trustworthy open-world deployment.

Keywords: vision-and-language navigation; embodied intelligence; paradigm evolution; foundation models; open-world generalization

1. Introduction

Vision-and-Language Navigation (VLN) is a fundamental task in embodied artificial intelligence that requires agents to follow natural language instructions, perceive visual scenes, model spatial states, and execute sequential decisions to navigate toward target locations in previously unseen environments [1,2]. This task holds broad application prospects for multifunctional intelligent assistants in daily life, including domestic service robots, autonomous driving systems, and personal navigation assistants [3].

Since Anderson et al. [4] introduced the Room-to-Room (R2R) benchmark, VLN has attracted broad research attention and achieved significant progress. Early sequence-to-sequence methods have gradually been replaced by Transformer-based models [5–8], in recent years, the introduction of large-scale foundation models, including Large Language Models (LLMs) and Vision-Language Models (VLMs) [9–13], has further advanced VLN performance on standard benchmarks. However, closer examination reveals a persistent gap between strong benchmark performance and the capabilities required for real-world deployment [1,2,14]. When agents encounter ambiguous instructions, unseen environments, dynamic obstacles, or long-horizon navigation tasks, their performance remains clearly limited. To bridge this gap, recent frontier research has explored multiple directions, including architectural design [15,16], reasoning and planning [9,11,17], and cross-modal representation [18–20]. However, these advances often focus on specific models, local tasks, or individual capabilities, and the field still lacks a systematic characterization of the intrinsic connections among different technical lines and the logic of paradigm evolution. This fragmentation highlights the need for a structured survey to trace the development trajectory of VLN capabilities and integrate scattered advances into a systematic roadmap toward open-world navigation.

Although recent VLN research has been undergoing a paradigmatic shift from reactive instruction following to cognitive navigation, with profound implications for the field, existing VLN surveys have primarily categorized methods according to technical pipelines or focused on specific benchmarks and application scenarios [1,21–24]. In other words, while prior surveys provide valuable foundational summaries, they still lack a systematic examination of the paradigm-level transformations that are reshaping the field. Crucially, These transformations extend beyond methodological optimization and point to a renewed understanding of the nature of navigational intelligence, research trajectories, and future directions. Table 1 compares this survey with existing VLN surveys and highlights the key gaps that remain insufficiently covered in prior literature and motivate the contributions of this work.

 doi:10.20944/preprints202606.2231.v1

Table 1. Comparison with existing VLN surveys. ✓ indicates the dimension is covered; ✗ indicates not covered.

Dimension	Gu et al. [1]	Wu et al. [21]	Zhang et al. [22]	Khan et al. [23]	Pan et al. [24]	Ours
Temporal coverage	–2022	–2023	–2024	–2025	–2025	**2022–2026**
Core perspective	Task taxonomy	Task taxonomy	Foundation model tools	Task taxonomy	Foundation language models	**Paradigm evolution**
Object/landmark grounding	✓	✓	✓	✓	✓	✓
3D scene understanding	✗	✗	✗	✓	✓	✓
Streaming / video VLN	✗	✗	✗	✗	✗	✓
Audio-visual navigation	✗	✗	✗	✗	✗	✓
Memory & history modeling	✓	✓	✓	✓	✓	✓
World models	✗	✗	✗	✓	✓	✓
LLM/VLM-based reasoning & planning	✗	✗	✓	✓	✓	✓
Zero-shot & open-world generalization	✗	✓	✓	✓	✓	✓
Long-horizon navigation	✗	✓	✓	✓	✓	✓
Agentic navigation & self-correction	✗	✗	✓	✓	✓	✓
Continual / lifelong learning	✗	✗	✗	✗	✗	✓
Self-evolving navigation	✗	✗	✗	✗	✗	✓
Cross-platform navigation (UAV, outdoor)	✓	✓	✓	✓	✓	✓
City-scale outdoor VLN	✗	✓	✓	✓	✓	✓
Trustworthy & safety-aware VLN	✗	✓	✓	✓	✗	✓
Social-aware & human-in-the-loop VLN	✗	✗	✗	✗	✗	✓

To characterize the evolution of VLN, we review its ongoing shift from instruction following to cognitive navigation, as shown in Figure 1. This transition can be understood by analogy to human cognitive development:

1. **Perception Evolution.** Analogous to humans moving from coarse visual impressions to precise spatial cognition, VLN perception has evolved from panoramic vision-language alignment to contextualized spatial understanding, enabling semantic entity grounding, 3D spatial construction, and streaming multi-source perception.
2. **Cognition Evolution.** Similar to humans using mental maps to imagine routes before acting, VLN cognition has shifted from reactive decisions based on immediate observations to world-model-driven predictive planning, enabling agents to transition from observation-driven reaction to model-based deliberation.
3. **Learning Evolution.** Analogous to humans progressing from imitation to intrinsic learning, VLN learning has evolved from supervised imitation to reward-driven optimization, enabling agents to learn from experience and self-correct via expert trajectories and foundation-model-guided rewards.
4. **Generalization Evolution.** As humans transfer knowledge to novel settings and adapt lifelong, VLN generalization has evolved from closed-benchmark evaluation toward reliable open-world operation, spanning environment, horizon, lifelong, scene, and safety dimensions.

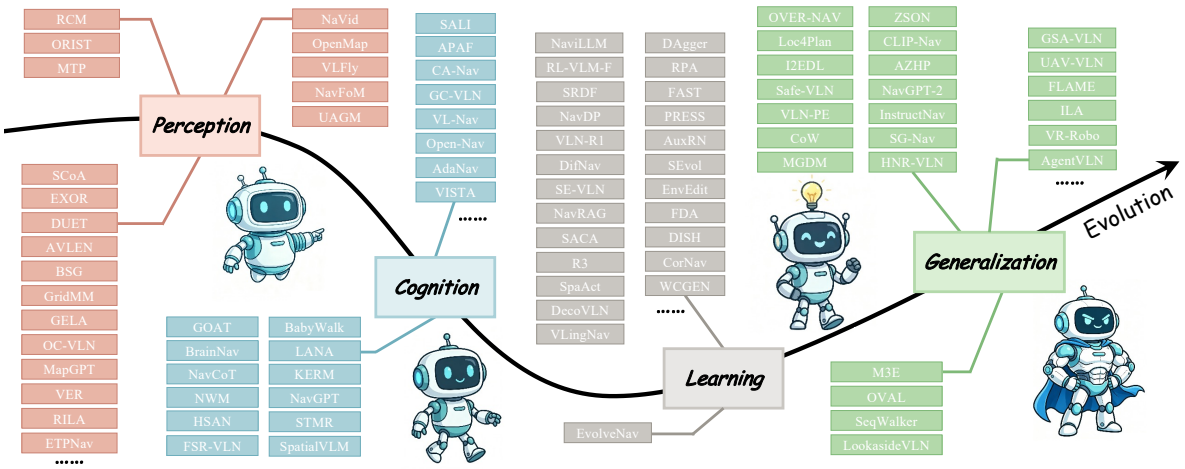

Figure 1. Evolutionary roadmap of Vision and Language Navigation from instruction following to cognitive navigation. Representative VLN studies are organized along two evolutionary trajectories. The main trajectory characterizes the progressive development across four core capability dimensions, including Perception, Cognition, Learning, and Generalization, while the branch trajectories within each dimension capture the methodological and paradigmatic evolution of specific research lines.

 doi:10.20944/preprints202606.2231.v1

This perspective suggests that VLN should be understood as a progressively organized system in which perception, reasoning, learning, and adaptation evolve in a hierarchical and mutually reinforcing manner. Unlike prior surveys that primarily categorize the field by technical components [1,21–24], this survey frames recent VLN research as part of an ongoing paradigm evolution. Such an evolution-centered perspective enables a more coherent characterization of the field's developmental trajectory and provides principled guidance for future research toward open-world cognitive navigation.

2. Preliminaries: The VLN Landscape

2.1. Task Formulation

VLN can be formalized as follows: an agent is situated in a 3D environment $\mathcal{E}$, which can be represented either as a connectivity graph of navigable viewpoints or as a continuous 3D space. At each time step t, the agent receives a visual observation o_t (typically an RGB or RGB-D panoramic image) and has access to a natural language instruction $\mathcal{I} = \{w_1, \ldots, w_L\}$. The agent must select an action a_t from its action space $\mathcal{A}$ to reach the target location specified by $\mathcal{I}$, with the episode terminating when the agent issues a STOP action or reaches the maximum number of steps.

Early VLN research predominantly adopted the discrete setting [5,6,25], where the environment is abstracted as an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and the agent selects the next viewpoint from neighboring nodes at each step. While such graph-based abstraction simplifies low-level motion modeling, its reliance on pre-defined topological structures introduces a substantial gap from the continuous physical space encountered by real-world robots. To bridge this gap, VLN-CE and subsequent works extend VLN to continuous environments [2,26,27], requiring agents to execute low-level control actions (e.g., Forward, Turn-Left, Turn-Right). This setting more closely approximates real-world deployment conditions but introduces new challenges including obstacle avoidance, motion planning, and fine-grained perception.

2.2. Representative Benchmarks

VLN has developed a rich and evolving benchmark ecosystem, in which benchmark design has progressively expanded the scope of navigation intelligence. The early stage was established by R2R [4], which formalized VLN as instruction following in unseen indoor environments built upon discrete viewpoint graphs. Shortly afterward, R4R [28] extended this setting to longer and more instruction-faithful trajectories, while Touchdown and StreetLearn moved VLN from indoor environments to outdoor street-view navigation [29,30]. Subsequently, the benchmark landscape diversified substantially: VLN-CE reformulated navigation in continuous environments with low-level control [2]; RxR introduced multilingual and densely aligned instructions for fine-grained language grounding [31]; REVERIE and ObjectNav emphasized object-centric goal grounding [32,33]; and CVDN, TEACh, and DialFRED incorporated dialogue, interaction, and task-oriented embodied execution [34–36]. More recent benchmarks further extend VLN toward long-horizon and persistent navigation [14,37–39], human-aware, safety-aware, continual, and feasibility-aware evaluation [40–44], as well as city-scale and UAV-based aerial navigation in simulated, real, and hybrid environments [45–48]. This evolution shows that VLN benchmarks have moved from short-range indoor instruction following on discrete graphs toward increasingly realistic evaluations of spatial grounding, continuous control, long-term planning, embodied interaction, open-world generalization, and trustworthy deployment. Table 2 summarizes representative VLN benchmarks in chronological order, with annotations of their environment types, task domains, and core characteristics.

 doi:10.20944/preprints202606.2231.v1

Table 2. Comprehensive summary of representative VLN benchmarks. "Sim." denotes simulator-based environments and "Real" denotes real-world environments.

Benchmark	Year	Environment	Domain	Highlight
R2R [4]	2018	Sim. (Matterport3D)	Indoor	Foundational VLN benchmark with step-by-step instructions
R4R [28]	2019	Sim. (Matterport3D)	Indoor	Long-path extension by concatenating R2R trajectories
Touchdown [29]	2019	Real (Street View)	Outdoor	First outdoor street-view VLN benchmark
StreetLearn [30]	2019	Real (Street View)	Outdoor	Large-scale street-level navigation
HANNA [25]	2019	Sim. (Matterport3D)	Indoor	Help-seeking navigation with subgoal requests
Just Ask [49]	2019	Sim. (Matterport3D)	Indoor	Active question-asking for ambiguity resolution
ALFRED [50]	2020	Sim. (AI2-THOR)	Indoor	Household task combining navigation and manipulation
REVERIE [32]	2020	Sim. (Matterport3D)	Indoor	Remote object grounding with high-level instructions
RxR [31]	2020	Sim. (Matterport3D)	Indoor	Multilingual extension with denser instructions
VLN-CE / R2R-CE [2]	2020	Sim. (Habitat)	Indoor	First continuous-environment VLN with low-level control
CVDN [34]	2020	Sim. (Matterport3D)	Indoor	Cooperative vision-and-dialog navigation
ObjectNav [33]	2020	Sim. (Habitat)	Indoor	Object-goal navigation in unseen environments
RoboSlang [51]	2020	Real	Indoor	Real-robot dialog-based VLN
Retouchdown [52]	2020	Real (Street View)	Outdoor	Refined Touchdown with cleaner annotations
SOON [53]	2021	Sim. (Matterport3D)	Indoor	Scenario-oriented object navigation with hierarchical reasoning
RxR-CE [2]	2021	Sim. (Habitat)	Indoor	Continuous-environment counterpart of RxR
Talk2Nav [54]	2021	Real (Street View)	Outdoor	Long-range outdoor navigation with attention dialog
TEACh [35]	2022	Sim. (AI2-THOR)	Indoor	Task-oriented embodied agent with chat dialogue
DialFRED [36]	2022	Sim. (AI2-THOR)	Indoor	Dialog-augmented household task execution
HM3D-AutoVLN [55]	2022	Sim. (HM3D)	Indoor	Auto-generated instructions on large-scale HM3D
IVLN [14]	2023	Sim. (Habitat)	Indoor	Iterative VLN with cross-episode persistent memory
AerialVLN [45]	2023	Sim. (UE4)	Outdoor	First city-scale UAV VLN benchmark
Safe-VLN [40]	2023	Sim. (Habitat)	Indoor	Collision-aware safe VLN-CE
HA-VLN [41]	2024	Sim. (Matterport3D)	Indoor	Human-aware VLN with dynamic human activities
R2R-IE-CE [56]	2024	Sim. (Habitat)	Indoor	Instruction error detection and localization
VLNCL [42]	2024	Sim. (Matterport3D)	Indoor	First continual learning benchmark for VLN
CVLN [43]	2024	Sim. (Habitat)	Indoor	Cross-domain continual VLN
NaviLLM-Bench [57]	2024	Sim. (Mixed)	Indoor	Unified evaluation across multiple VLN tasks
VLN-Video [58]	2024	Real (Driving Video)	Outdoor	Driving-video-based outdoor VLN
CityNav [46]	2024	Real (Aerial)	Outdoor	City-scale aerial navigation dataset
Open X-E [59]	2024	Real	Indoor/Outdoor	Cross-embodiment large-scale dataset
LHPR-VLN [37]	2025	Sim. (Habitat)	Indoor	First long-horizon VLN benchmark, ∼150-step trajectories
MG-VLN [38]	2025	Sim. (Habitat)	Indoor	Multi-goal sequential navigation
GSA-VLN [60]	2025	Sim. (Habitat)	Indoor	Generalized scene adaptation with memory bank
HA-VLN 2.0 [61]	2025	Sim. (Matterport3D)	Indoor	Multi-human social-norm-aware VLN
VLN-PE [62]	2025	Sim.+Real	Indoor	Physical-level platform across multi-embodiments
VR-Robo [63]	2025	Real-Sim-Real	Indoor	High-fidelity digital twins for sim-to-real transfer
OpenFly [47]	2025	Sim./Real	Outdoor	100K aerial trajectories, keyframe-aware UAV VLN
UAV-VLN [64]	2025	Sim. (UE4)	Outdoor	End-to-end velocity-yaw regression for UAV
StreamVLN [65]	2025	Sim. (Habitat)	Indoor	Streaming video-based VLN with online dialogue
CoNavBench [39]	2026	Sim. (Habitat)	Indoor	Multi-agent collaborative long-horizon VLN
VLNVerse [66]	2026	Sim. (Physics)	Indoor	Physics-aware large-scale VLN benchmark
VLN-NF [44]	2026	Sim. (Habitat)	Indoor	Feasibility-aware VLN with false-premise instructions
CoT-VLNBench [67]	2026	Sim. (Mixed)	Indoor	Visual chain-of-thought reasoning benchmark
AirNav [48]	2026	Real (Aerial)	Outdoor	Large-scale UAV VLN dataset for MLLM evaluation

2.3. Standard Metrics

VLN evaluation metrics quantify the effectiveness, efficiency, and trajectory quality of agent behavior, serving as the cornerstone for performance measurement and cross-work comparability. We introduce the commonly adopted metrics in VLN as follows:

- **Success Rate (SR).** The percentage of episodes in which the agent stops within a threshold distance (typically 3 meters) of the goal.
- **Oracle Success Rate (OSR).** SR computed using the closest point along the agent's trajectory to the goal, indicating whether the agent ever passes near the target.
- **Path Length (PL).** The total distance traveled by the agent during task completion, where shorter paths indicate higher navigation efficiency.
- **Success weighted by Path Length (SPL).** SR normalized by the ratio of shortest-path length to actual path length, penalizing unnecessarily long trajectories.
- **Navigation Error (NE).** The average distance between the agent's final position and the goal.
- **Normalized Dynamic Time Warping (nDTW).** A measure of the fidelity of the agent's trajectory to the reference path.
- **Trajectory Length (TL).** The total distance traveled by the agent during the navigation episode.

 doi:10.20944/preprints202606.2231.v1

3. Perception Evolution: From Visual Grounding to Situated Spatial Understanding

Vision-Language Navigation begins with perception. Early VLN agents primarily relied on 2D visual observations and cross-modal attention to associate natural language instructions with visual cues in the current scene. This setting was sufficient for benchmark-oriented navigation, where agents selected actions from discrete viewpoints based on panoramic images and textual instructions. However, real-world navigation requires more than recognizing objects in a single view. An embodied agent must identify language-referable entities, understand navigable space, maintain spatial structure across time, perceive three-dimensional geometry, process continuous sensory streams, and integrate multiple sources of environmental context.

Therefore, the evolution of perception in VLN can be understood as a transition from image-level grounding to embodied spatial understanding. Figure 2 illustrates this perceptual evolution by organizing representative studies along three major axes: semantic granularity, spatial structure, and input realism. This transition follows a progressive path: agents first learn to ground instructions in 2D visual features, then recognize objects and landmarks as semantic anchors, organize views into topological graphs, construct metric and map-based spatial representations, recover realistic 3D scene structure, process temporal observations from egocentric video, and finally integrate multi-source cues such as audio, dialogue, and human activities. Along this trajectory, perception evolves from a passive visual encoder into the sensory foundation for real-world embodied navigation. Building on this evolutionary perspective, Figure 3 further provides a structured taxonomy that maps these three axes to specific perceptual capabilities, representative research directions, and typical methods.

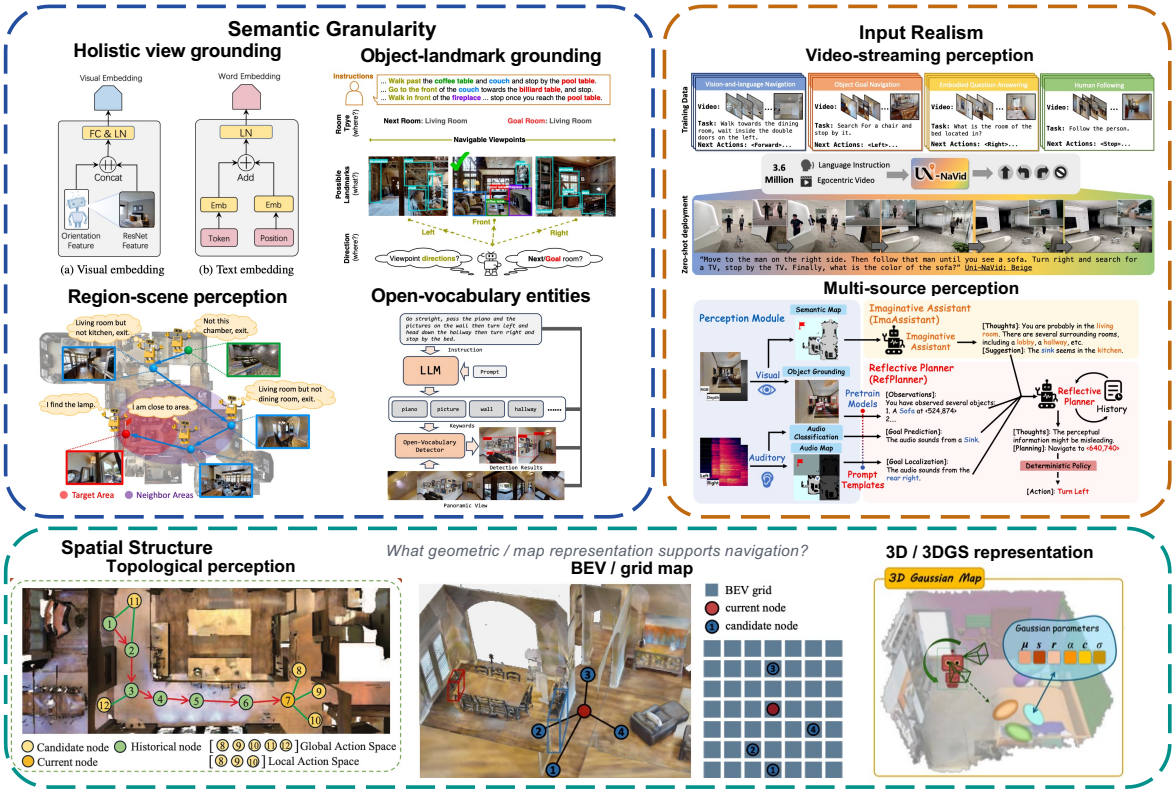

Figure 2. Perception evolution in vision-language navigation. The figure organizes representative VLN perception studies along three major axes: semantic granularity, input realism, and spatial structure. These axes characterize how agents evolve in recognizing language-referable visual content, processing more realistic sensory inputs, and constructing structured spatial representations for embodied navigation.

 doi:10.20944/preprints202606.2231.v1

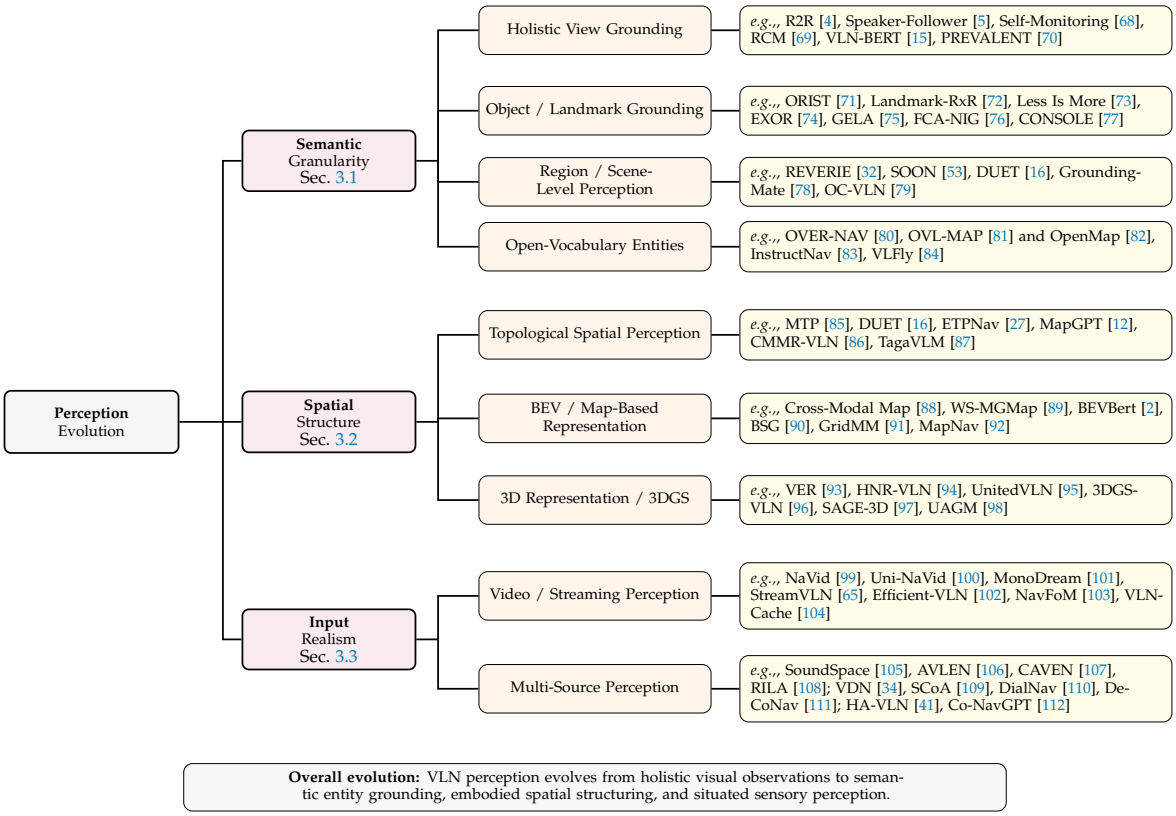

Figure 3. A structured taxonomy of perception evolution in VLN. VLN perception evolves along three complementary axes: semantic granularity, spatial structure, and input realism. These axes jointly move agents from image-level visual grounding toward situated embodied spatial understanding.

3.1. Semantic Granularity Evolution: From Holistic Views to Open-Vocabulary Semantic Anchors

Early VLN agents typically perceive the environment through holistic panoramic features and align them with natural language instructions using cross-modal attention or vision-language pre-training. While such image-level grounding is effective for benchmark-oriented instruction following, it is insufficient for real-world navigation, where instructions are rarely grounded in an entire scene. Instead, humans usually refer to objects, landmarks, room regions, and open-ended semantic concepts, such as "the sofa near the doorway," "the hallway on the left," or "the painting above the table." Therefore, VLN perception has gradually evolved from holistic image-level matching to language-referable semantic grounding. This evolution can be characterized along four stages: holistic image grounding, object/landmark grounding, region and scene understanding, and open-vocabulary entity perception. Together, these stages transform visual perception from recognizing a view to identifying the semantic anchors that make navigation instructions executable.

3.1.1. Holistic Image Perception

Early VLN perception was primarily built upon holistic visual-language grounding, where an agent encodes panoramic observations or candidate viewpoints as global visual features and aligns them with the language instructions for action prediction. The R2R benchmark [4] established this paradigm by introducing instruction-following navigation in Matterport3D panoramic environments, after which early agents such as Speaker-Follower [5], Self-Monitoring [68], and RCM [69] learned to match instructions with visual observations, navigation trajectories, or progress states through sequence modeling, cross-modal attention, and trajectory-level matching. These methods formed the basic perception interface of VLN: visual observations provide candidate views, language instructions specify navigation intent, and cross-modal grounding determines which direction should be selected at each step.

 doi:10.20944/preprints202606.2231.v1

With the rise of vision-language pre-training, this holistic perception paradigm was further strengthened. VLN-BERT [15] and PREVALENT [70] introduced pre-trained multimodal representations for instruction-conditioned navigation, while HAMT [8] incorporated long-horizon panoramic history into multimodal decision making. More recent alignment-oriented works, such as CSAP [113] and DELAN [114], further refine cross-modal perception by encouraging semantic alignment between language fragments and visual observations before or during multimodal fusion. These advances significantly improve the agent's ability to associate textual instructions with visual scenes, and they represent the transition from hand-designed visual-language fusion to pre-trained and alignment-aware multimodal perception.

However, despite these improvements, holistic image perception remains inherently coarse-grained. It treats the whole panorama as the basic perceptual unit, whereas human instructions usually rely on finer semantic anchors, such as objects, landmarks and regions. For example, "turn left at the sofa" requires the agent to ground the entity "sofa", not merely match the whole panorama. This motivates the shift from image-level perception to entity and scene-level perception, where agents identify the semantic anchors and contextual cues that make navigation instructions executable.

3.1.2. Entity and Scene-Level Perception

Rather than representing a scene only through global view features, VLN agents increasingly ground instructions in concrete semantic anchors, including objects, landmarks, room cues, and target-related regions. Early object-aware methods such as ORIST [71] introduce object- and room-level information into sequential decision making, while Landmark-RxR [72] provides fine-grained supervision between instruction spans and visual landmarks. Less Is More [73] highlights landmarks as compact navigation anchors, and EXOR [74] further models object-relation alignment.

Recent works move toward richer entity–context grounding. GELA [75] introduces GEL-R2R with entity phrase prediction, landmark bounding box prediction, and entity-landmark alignment objectives. FCA-NIG [76] reduces the cost of fine-grained supervision by automatically generating sub-instruction–sub-trajectory and entity–landmark annotations. CONSOLE [77] further extends landmark grounding to open-world sequential landmark discovery with large-model commonsense and CLIP-based discovery.

Beyond individual entities, goal-oriented VLN tasks require agents to understand where an entity is situated and what contextual cues make it identifiable. Benchmarks such as REVERIE [32] and SOON [53] move VLN from route following toward remote object localization and scenario-oriented object navigation, where the target may be described by its attributes, relations, room region, or nearby context. Methods such as DUET [16], Meta-explore [115], GroundingMate [78], OC-VLN [79], and image-goal auxiliary learning [116] further strengthen target grounding by connecting object-level perception with surrounding scene context and goal-oriented navigation. In this sense, VLN perception evolves from recognizing what entity is mentioned to understanding where it is situated and how it is embedded in the scene.

However, entity and scene-level grounding still often assumes that the relevant objects, landmarks, or room concepts are covered by the training vocabulary or benchmark annotations. Real-world navigation is more open-ended: users may refer to rare objects, long-tail landmarks, novel room categories, or free-form semantic concepts. This motivates the next stage of semantic granularity evolution: open-vocabulary entity perception.

3.1.3. Open-Vocabulary Entity Perception

While object, landmark, and scene-level perception helps agents identify language-referable anchors, many methods still assume that these entities are covered by benchmark training data. In real-world navigation, users may refer to long-tail objects, unseen landmarks, fine-grained attributes, or free-form semantic concepts. Thus, VLN perception must evolve from grounding known entities to discovering arbitrary entities described by natural language.

 doi:10.20944/preprints202606.2231.v1

This transition is enabled by open-vocabulary vision foundation models. CLIP [117] provides language-queryable visual representations, while Detic [118], SAM [119] and Grounding DINO [120] support open-set detection and segmentation. Building on these models, VLMaps [121], ConceptFusion [122], OpenScene [123], OVIR-3D [124], and HOV-SG [125] extend open-vocabulary semantics from 2D images to 3D maps, object instances, and hierarchical scene graphs. These works provide the perceptual basis for agents to locate and reason about novel entities in spatial environments.

Recent VLN methods further integrate open-vocabulary perception into navigation. CONSOLE [77] formulates VLN as open-world sequential landmark discovery using landmark commonsense and CLIP-based discovery. OVER-NAV [80] combines open-vocabulary detection with structured memory for iterative VLN. OVL-MAP [81] and OpenMap [82] build online or zero-shot visual-language maps for instruction grounding, while InstructNav [83] and VLFly [84] show how open-vocabulary goal understanding can support navigation in unexplored indoor or aerial environments.

However, open-vocabulary perception also brings new challenges: detected labels may be noisy or inconsistent across views, and recognizing an entity does not necessarily reveal its navigational role. Therefore, semantic openness must be complemented by spatial organization. This naturally leads to the next perception axis: how agents structure entities, landmarks, and regions into graphs, maps, and 3D spaces.

3.2. Spatial Structure Evolution: From Local Views to Embodied 3D Space

While semantic granularity evolution enables VLN agents to identify language-referable entities, navigation further requires understanding how these entities, viewpoints, and traversable regions are organized in space. Early VLN agents mainly reason over local panoramic observations or candidate views, which is insufficient for long-horizon navigation because the agent must remember where it has been, infer where it can go, and maintain spatial relations among landmarks and goals. Therefore, VLN perception has evolved from local view-based observation toward structured spatial representation. This evolution can be characterized along three stages: topological spatial perception, BEV and map-based spatial representation, and 3D/Gaussian-based spatial perception. Together, these stages transform VLN perception from recognizing what is visible to organizing where things are in an embodied environment.

3.2.1. Topological Spatial Perception

Topological methods represent viewpoints, explored locations, or predicted waypoints as nodes, and navigable transitions as edges. This allows the agent to reason about connectivity, visited history, and global route choices beyond the current panorama. Early works such as SSM [126] and CMTP [85] introduce structured scene memory or topological maps to support long-range reasoning and language-conditioned planning.

A representative milestone is DUET [16], which builds an online topological map and performs coarse-to-fine reasoning with a dual-scale graph transformer. Later works further extend topology-based perception to more realistic settings. ETPNav [27] performs online topological mapping in continuous environments by self-organizing predicted waypoints along the traversed path. MapGPT [12] converts an online topological map into language prompts, allowing GPT-4 [127] to reason over node information and topological relationships for zero-shot navigation. MAM [128] and COSMO [129] further explore how topological memory can be selectively stored or retrieved to reduce long-horizon memory cost. Recent works further integrate topology with foundation models: CMMR-VLN [86] equips LLM agents with continual multimodal memory retrieval over panoramas and landmarks, while TagaVLM [87] injects topological structures into VLMs for global action reasoning.

However, topological representations mainly answer the question of "where can I go?" They capture connectivity among viewpoints, but often provide limited metric geometry, spatial layout, obstacle structure, and fine-grained traversable regions. This motivates the transition from topology-

 doi:10.20944/preprints202606.2231.v1

centered spatial perception to BEV and map-based representations, which provide more explicit spatial layouts for navigation.

3.2.2. BEV and Map-Based Spatial Representation

BEV and map-based representations extend topology into metric and semantic spatial layouts. They encode navigable regions, object locations, directions, and history in top-down or map-like forms, enabling agents to reason about both their location and the structure of the environment. Early works such as Cross-modal Map [88] and WS-MGMap [89] show the importance of language-guided top-down semantic maps and multi-granularity spatial memory. BEVBert [130] further introduces map-based pre-training by combining local metric maps for aggregating incomplete observations with global topological maps for navigation dependency modeling.

More recent methods further strengthen map-based spatial perception. SUSA [131] enriches map-based VLN perception beyond RGB by combining view-level textual semantic panoramas with trajectory-level depth exploration maps, enabling hierarchical semantic-spatial grounding for local action selection and global route planning. BSG [90] constructs BEV scene graphs to encode indoor scene layouts and geometric cues, maintaining an online BEV-based global scene map during navigation. GridMM [91] builds a dynamically growing egocentric grid memory map, projecting historical observations into a unified top-down space and aggregating instruction-relevant visual clues in each grid region. MapNav [92] replaces raw historical frames with annotated semantic maps, using explicit textual labels on key regions as structured navigation cues for VLM-based agents. OVL-MAP [81] and OpenMap [82] further integrate open-vocabulary semantics into online or zero-shot visual-language maps, connecting the semantic openness discussed in the previous section with spatial organization.

Despite these advances, BEV and map-based representations remain abstractions of the physical environment. They are effective for encoding layout, traversability, and semantic regions, but they often lose fine-grained 3D geometry, occlusion structure, and physical executability. As VLN moves closer to real-world deployment, agents need richer spatial substrates that preserve the geometry, semantics, and visual realism of 3D environments.

3.2.3. 3D Spatial Representation

The next stage moves from 2D or 2.5D map abstractions to richer 3D spatial representations. Recent works explore different forms of 3D spatial perception for VLN. VER [93] voxelizes the physical world into structured 3D cells and aggregates multi-view 2D features into a unified 3D space. iPPD [132] uses global 3D semantic maps to generate and score path proposals. HNR-VLN [94] introduces neural radiance representations for continuous VLN, enabling lookahead exploration by predicting future environmental features.

More recently, 3D Gaussian Splatting has become an emerging spatial substrate for VLN. United-VLN [95] uses generalizable 3DGS-based pre-training to render high-quality panoramic observations and semantic features for future environment exploration. 3DGS-VLN [96] builds a unified Gaussian map with open-set semantics for geometry- and semantics-aware action prediction. SAGE-3D [97] further enhances 3DGS with semantic grounding and physics-aware execution interfaces, while UAGM [98] models geometric, semantic, and appearance uncertainties for more reliable spatial grounding. Complementary to explicit 3D maps, JanusVLN [133] explores dual implicit memory by decoupling spatial-geometric and visual-semantic representations, suggesting a more compact way to preserve 3D-aware navigation history.

These 3D representations move VLN perception from abstract spatial memory toward realistic embodied spatial perception. They better preserve geometry, semantics, appearance, and physical constraints, which are important for sim-to-real transfer and real-world deployment.

3.3. Input Realism Evolution: From Static Observations to Situated Sensory Streams

Classical VLN benchmarks often provide discrete panoramas, pre-defined candidate viewpoints, or precomputed spatial structures, while embodied agents perceive the world through continuous

 doi:10.20944/preprints202606.2231.v1

first-person streams, noisy sensors, human interaction, and dynamic social contexts. Therefore, VLN perception further evolves along the axis of input realism: from static image or panorama inputs to video-based streaming perception, and from vision-only sensing to multi-source situated perception. This evolution brings VLN closer to the sensory conditions of physical robots operating in real environments.

3.3.1. Video Streaming Perception

Video-based perception shifts VLN from isolated visual observations to continuous egocentric input. Video-based agents process first-person RGB streams and use temporal context to capture motion continuity, historical observations, and visual changes. This setting better matches real robot deployment, where observations arrive sequentially and navigation decisions must be made online.

A representative work is NaVid [99], which formulates VLN as video-based VLM planning. It directly takes monocular RGB video streams and language instructions as input and outputs the next navigation action. Uni-NaVid [100] further extends NaVid into multiple embodied navigation tasks under a shared input-output format. VLN-R1 [134] moves toward end-to-end egocentric video-to-action navigation by reinforcement fine-tuning LVLMs on first-person video streams and continuous action prediction. MonoDream [101] improves monocular VLN by learning navigation representations that predict panoramic and depth-aware latent cues from monocular inputs.

Recent streaming VLN also faces a growing long-context bottleneck. Beyond simply using more frames, recent works begin to select, compress, and reuse context. StreamVLN [65] introduces slow-fast context modeling for streaming VLN, using a fast dialogue context for responsive action generation and a slow memory context for compressing long-term visual history. Efficient-VLN [102] reduces the training and inference overhead of RGB-only VLN by designing efficient memory mechanisms for long visual contexts. STEP-Nav [135] further improves the efficiency of LLM-based streaming VLN by pruning spatially irrelevant image tokens and temporally redundant frames, while using distortion-aware fine-tuning to preserve navigation performance under compressed visual inputs. JanusVLN [133] decouples spatial-geometric and visual-semantic memory into compact implicit representations. NavFoM [103] adopts a forgetting-curve-inspired sampling strategy. VLN-Cache [104] reuses cached tokens across viewpoint transitions. These works suggest that input realism must be accompanied by context efficiency for real-time embodied navigation.

Although video streaming perception makes VLN inputs closer to real embodied sensing, these inputs remain primarily visual. Real environments also involve sound, dialogue, human activities, and multi-agent interactions. Recent works further extend VLN perception toward multi-source sensing.

3.3.2. Multi-Source Perception

In the real world, navigation decisions may depend on audio cues, human dialogue, social signals, and information from other agents. These cues can reveal events outside the field of view, resolve ambiguous instructions, indicate human intentions, or support collaboration. Thus, VLN perception evolves from seeing the environment to sensing and interpreting a situated environment.

Audio-visual language navigation is an important step in this direction. SoundSpaces [105] provides a foundation for audio-visual navigation by simulating spatialized sound in 3D environments, where agents use both vision and audio to locate sound-emitting targets. AVLEN [106] extends this to audio-visual-language embodied navigation, where the agent uses visual observations, audio events, and natural language. AVLMaps [136] stores audio, visual, and language cues in a unified 3D spatial map, enabling zero-shot multimodal navigation. CAVEN [107] further introduces a conversational audio-visual navigation framework in noisy environments, allowing the agent to interact with a human/oracle when audio cues are uncertain. RILA [108] moves this setting toward zero-shot semantic audio-visual navigation. ENMuS [137] proposes a noisy multi-source audio-visual navigation benchmark, BeDAViN. More recent works, such as NaVLA2 [138], further explore vision-language-audio-action modeling for multimodal instruction navigation.

 doi:10.20944/preprints202606.2231.v1

Dialogue-aware perception extends VLN from single-shot instruction following to interactive situated navigation. Vision-and-Dialog Navigation (VDN) [34] first uses dialog history to infer navigation goals, while SCoA [109] enables agents to actively decide when to communicate. AVDN [139] extends this setting to outdoor aerial scenes, and D-CVLN [43] studies how dialog-grounded experience can be retained across changing tasks. Recent works strengthen dialog-aware navigation with finer grounding and longer interaction. Fine-FG-AVDN [140] aligns dialog expressions with visual and navigation states, DialNav [110] studies multi-turn navigation with a remote guide, and DeCoNav [111] uses dialog for long-horizon collaborative navigation. Together, these works treat dialog as a situated perception channel for resolving ambiguity, updating goals, and coordinating navigation.

Human-aware and collaborative perception further pushes VLN towards real social environments. HA-VLN [41] introduces dynamic human activities into VLN, while HA-VLN 2.0 [61] adds multi-human interactions, social-awareness constraints, real-world validation, and an open leaderboard. Co-NavGPT [112] extends this direction to multi-robot collaborative visual semantic navigation. These works show that real-world VLN perception must account for people, communication, and other agents.

In summary, input realism evolves VLN perception from static benchmark observations to continuous and situated sensory streams. Video-based methods address temporal and online perception, while multi-source methods incorporate audio, dialogue, humans, and collaborative signals. Together with semantic granularity and spatial structure, this completes the perception-level evolution of VLN. Nevertheless, richer perception cannot guarantee intelligent navigation: agents must interpret instructions, reason spatial relations and actions. This naturally leads from perception to cognition.

3.4. Open Challenges in Perception

Despite the rapid evolution, several challenges remain open. First, semantic openness does not yet guarantee reliable navigational grounding: open-vocabulary detectors and language-queryable maps can recognize long-tail entities, but their labels may be noisy, view-dependent, or weakly connected to the actions required by an instruction. Second, richer spatial representations, including topological graphs, BEV maps, and 3D/Gaussian scene fields, improve structural awareness but introduce new trade-offs among online construction cost, memory consumption, update frequency, and physical executability. Third, many perceptual representations are still developed under relatively static assumptions, whereas real-world navigation requires agents to handle moving humans, changing objects, noisy sensors, and multi-source cues under real-time constraints. Finally, current evaluation is still dominated by policy-level navigation metrics, which indicate whether an agent reaches the goal but rarely diagnose whether failures originate from semantic grounding, spatial mapping, temporal perception, or downstream planning. Future VLN perception should therefore move toward open-vocabulary, uncertainty-aware, temporally updatable, computationally efficient, and independently evaluable representations.

4. Cognition Evolution: From Instruction Interpretation to Predictive World Modeling

Perception provides the sensory and spatial substrate for VLN, but perception alone does not determine navigation behavior. After identifying semantic anchors, organizing spatial structures, and receiving realistic sensory streams, an agent must still interpret what the language instruction requires, infer how goals and landmarks relate in space, decide which action sequence should be executed, and predict the future states resulting from its actions. This marks the transition from perception to cognition. Figure 4 illustrates this cognition evolution by organizing representative VLN studies along four major axes: instruction abstraction, spatial reasoning, deliberative planning, and world modeling.

 doi:10.20944/preprints202606.2231.v1

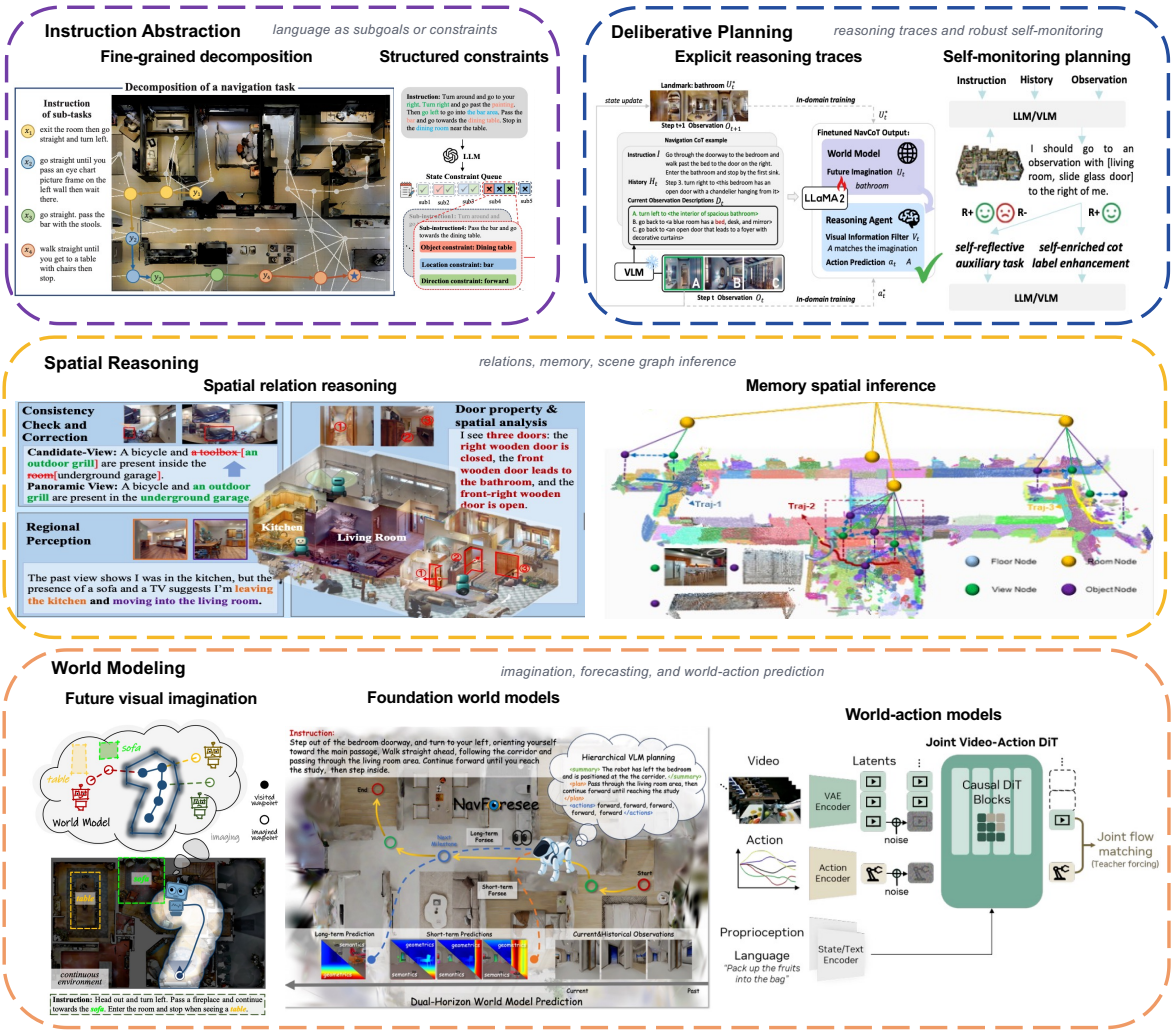

Figure 4. Cognition evolution in vision-language navigation. The figure organizes representative VLN cognition studies along four major axes: instruction abstraction, spatial reasoning, deliberative planning, and world modeling. These axes characterize how agents transform natural-language instructions into executable task structures, reason over navigation decisions, infer spatial relations, and model future states to support more reliable embodied navigation.

Following this evolutionary perspective, this section organizes VLN cognition along these four axes. First, instruction abstraction transforms raw natural language into executable task structures. Second, spatial reasoning infers relations among grounded entities, regions, and paths. Third, deliberative planning moves navigation from implicit action prediction to explicit reasoning and self-monitoring. Finally, world modeling enables agents to imagine future states and evaluate possible actions before execution. Building on this axis-level organization, Figure 5 further provides a structured taxonomy that maps these axes to concrete cognitive capabilities, representative research directions, and typical methods. Together, these axes describe how VLN agents evolve from instruction followers into cognitive navigators.

 doi:10.20944/preprints202606.2231.v1

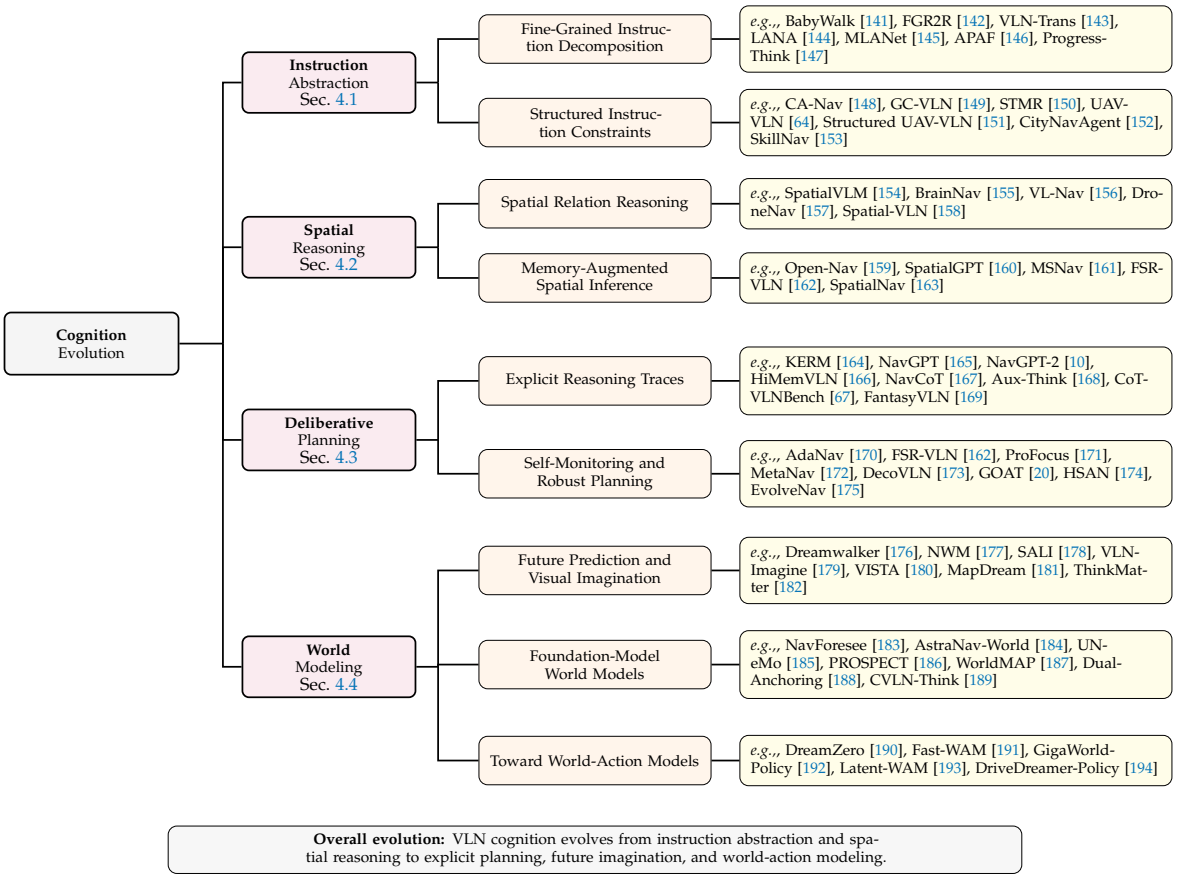

Figure 5. A structured taxonomy of cognition evolution in VLN. Cognition evolves along four complementary axes: instruction abstraction, spatial reasoning, deliberative planning, and world modeling. These axes move agents from instruction followers toward cognitive navigators that can structure tasks, infer spatial relations, reason over decisions, and imagine future states.

4.1. Instruction Abstraction Evolution: From Raw Instructions to Executable Task Structures

Navigation instructions are typically long, compositional, and temporally ordered. Early VLN agents often represent the instruction as a sentence-level embedding, whereas real-world navigation requires a more structured interpretation that identifies the currently relevant instruction segment, the corresponding subgoal, and the landmark-action associations that guide each navigation step. Thus, the first stage of VLN cognition is to abstract natural language into executable task structures.

4.1.1. Fine-Grained Instruction Decomposition

Fine-grained instruction decomposition aims to convert long and compositional navigation instructions into smaller executable units. Early works show that long-horizon navigation benefits from step-level language abstraction. BabyWalk [141] decomposes long instructions into short "BabySteps" and trains agents to complete them progressively. FGR2R [142] further constructs the dataset with sub-instruction and sub-path annotations, enabling agents to attend to the currently relevant instruction segment during navigation.

Recent works extend this idea from manually defined sub-instructions to adaptive and agent-aware instruction transformation. VLN-Trans [143] introduces a translator module that converts original instructions into easier-to-follow sub-instruction representations according to recognizable and distinctive landmarks. LANA [144] and Less Is More [73] also highlight that navigation language is structured around route descriptions, landmarks, and grounded instruction units. MLANet [145] further segments raw instructions into sub-instructions for continuous VLN and uses multi-level attention to select the active instruction segment. APAF [146] focuses on long-instruction VLN by aligning action-aware visual representations with action-oriented language instructions. Progress-

 doi:10.20944/preprints202606.2231.v1

Think [147] further models semantic progress reasoning by aligning visual history with instruction prefixes, helping agents estimate how far they have advanced within a multi-step instruction.

Overall, this line moves VLN from sentence-level instruction encoding to step-level task abstraction. However, decomposed instruction units alone do not fully specify how subgoals should be ordered, constrained, or grounded as an executable plan. This motivates structured instruction reasoning, where instructions are further represented as explicit constraints or task graphs.

4.1.2. Structured Instruction Constraints

Structured instruction constraints convert decomposed language units into explicit planning structures, such as constraints, graphs, semantic subgoals, or reusable skills. This makes instruction understanding more executable and interpretable. For example, CA-Nav [148] formulates zero-shot VLN-CE as constraint-aware sub-instruction completion, using a sub-instruction manager to track completion conditions and a value mapper to produce online navigation plans. GC-VLN [149] further represents instructions as graph constraints, decomposing them into waypoint nodes, object nodes, and spatial-relation edges for constraint solving and backtracking.

Recent aerial and long-horizon VLN works extend this idea to larger environments. STMR [150] converts instruction-related landmarks into semantic-topo-metric prompts for LLM-based aerial navigation. UAV-VLN [64] uses LLM-based goal parsing and visual grounding for aerial trajectory planning. Structured UAV-VLN [151] parses UAV instructions and aligns them with scene information, while CityNavAgent [152] decomposes urban aerial navigation into hierarchical semantic subgoals with global memory.

Another recent direction represents navigation requirements as skills. SkillNav [153] decomposes VLN tasks into interpretable atomic skills and selects them with a VLM-based router according to subgoals, observations, and history. Together, these works move instruction understanding from fine-grained decomposition to structured task constraints.

4.2. Spatial Reasoning Evolution: From Grounded Anchors to Relational Spatial Inference

Spatial reasoning operates on the semantic anchors and spatial structures built by perception. Its goal is not only to recognize landmarks, but also to infer their relative positions, distances, and relations to the agent's current state and future path. In this sense, VLN cognition moves from identifying what is mentioned to understanding how mentioned entities are spatially related.

4.2.1. Spatial Relation Reasoning

A first line of work strengthens the agent's ability to reason about spatial relations. SpatialVLM [154] shows that general VLMs still struggle with 3D spatial reasoning, such as distance, size, and relative position estimation, and improves this ability through large-scale spatial reasoning data. BrainNav [155] further introduces a bio-inspired spatial cognitive framework with dual maps and dual orientations to reduce spatial hallucination in real-world VLN. VL-Nav [156] integrates pixel-wise vision-language features with spatial reasoning for real-time robot navigation in indoor and outdoor environments.

Recent VLN works make spatial reasoning more task-specific. DroneNav [157] introduces unified text-visual representation and structured spatial reasoning for UAV-VLN. Spatial-VLN [158] explicitly addresses spatial perception bottlenecks in zero-shot VLN, including door interaction, multi-room navigation, and ambiguous instruction execution. Complementing these task-specific methods, NavSpace [195] introduces a spatial-intelligence benchmark for instruction navigation, evaluating agents' ability to translate vertical perception, metric movement, viewpoint shifting, spatial relations, environmental states, and space-structure understanding into executable navigation actions. These works move spatial cognition from simple landmark recognition toward explicit reasoning over directions, regions, transitions, and spatial affordances.

 doi:10.20944/preprints202606.2231.v1

4.2.2. Memory-Augmented Spatial Inference

Beyond local spatial relations, long-horizon navigation requires spatial inference over memory. Open-Nav [159] uses spatial-temporal chain-of-thought reasoning to connect instruction comprehension, progress estimation, and decision-making in continuous environments. SpatialGPT [160] further performs spatial CoT over structured spatial memory, combining local and global spatial context for zero-shot navigation. MSNav [161] integrates dynamic map memory, LLM-based spatial reasoning, and decision planning, using selective memory pruning to reduce overload while improving object-relation inference.

Recent works further organize spatial memory into richer structures. FSR-VLN [162] builds a hierarchical multimodal scene graph to support coarse-to-fine spatial retrieval and fast-to-slow reasoning. SpatialNav [163] leverages spatial scene graphs for zero-shot VLN, combining global spatial knowledge with agent-centric spatial representations and remote object localization. Together, these methods move spatial reasoning from local relation understanding toward memory-augmented inference over explored regions, landmarks, and paths. However, spatial reasoning mainly explains where things are and how they relate; navigation also requires deciding how to act under uncertainty, which motivates deliberative planning.

4.3. Deliberative Planning Evolution: From Implicit Policies to Explicit Reasoning

Early VLN agents often hide decision making inside an implicit navigation policy. Although such policies can work well on standard benchmarks, their reasoning process is difficult to inspect, diagnose, or correct. Recent works address this limitation by making navigation decisions more explicit, decomposable, and self-monitoring.

4.3.1. Explicit Reasoning Traces

Recent work has shifted VLN planning from implicit action policies to explicit reasoning processes. Agents are encouraged to generate interpretable reasoning traces for decision making. KERM [164] introduces external commonsense knowledge for object and relation reasoning, showing the value of knowledge-enhanced navigation decisions. LLM-based agents such as NavGPT [165] and NavGPT-2 [10] demonstrate that large models can serve as explicit navigation reasoners when provided with visual or textual scene descriptions.

Recent works render reasoning traces trainable and navigation-specific. Open-Nav [159] and HiMemVLN [166] explore zero-shot VLN in continuous environments using open-source LLMs combined with spatial-temporal chain-of-thought reasoning and a hierarchical memory system. Nav-CoT [167] introduces navigational chain-of-thought training to improve the accuracy and interpretability of LLM-based VLN. Aux-Think [168] studies auxiliary reasoning strategies for data-efficient VLN, while CoT-VLNBench [196] provides a large-scale benchmark with fine-grained CoT reasoning traces for quadruped robot navigation. FantasyVLN [169] further moves from explicit textual CoT to unified multimodal CoT, using compact latent reasoning to reduce token overhead while preserving reasoning-aware representations.

These works move VLN planning from black-box action prediction to explicit reasoning traces. However, explicit reasoning alone does not guarantee robust behavior: agents may reason unnecessarily, focus on irrelevant observations, or accumulate errors over long trajectories. This motivates adaptive and self-monitoring planning.

4.3.2. Self-Monitoring and Robust Planning

Recent works further make reasoning selective, corrective, and robust. AdaNav [170] dynamically triggers reasoning based on uncertainty, enabling difficulty-aware reasoning instead of fixed-step deliberation. FSR-VLN [162] introduces fast-and-slow reasoning, which adopts efficient matching for simple cases and deeper reasoning for difficult decisions. ProFocus [171] combines proactive perception with focused reasoning to reduce redundant visual processing and irrelevant historical context. MetaNav [172] introduces metacognitive reasoning to monitor exploration progress, detect

 doi:10.20944/preprints202606.2231.v1

wandering, and generate corrective rules. AwareVLN [197] advances by introducing sparse self-aware reasoning into RGB-only VLN-CE, enabling the agent to selectively assess scene context, instruction progress, navigation deviations, and next-step plans at key decision points for more explainable and robust action generation. DecoVLN [173] further decouples observation, reasoning, and correction, making error diagnosis and corrective navigation more explicit.

Robust planning also requires reducing biased or spurious decision logic. GOAT [20] introduces causal learning into VLN by modeling observable and unobservable confounders across vision, language, and history. HSAN [174] further combines semantic augmentation, optimal transport, and graph-driven reasoning for structured decision making. EvolveNav [175] moves toward self-improving embodied reasoning by combining formalized CoT supervised fine-tuning with self-reflective post-training, while DeepVLN [198] explores deep reasoning and collaborative mechanisms based on LLMs.

Together, these methods evolve VLN planning from explicit reasoning to adaptive, corrective, and self-improving decision making. However, they still mainly operate over current observations and memories. To reason beyond observed states, agents need internal models to predict future observations and action consequences, which motivates world model-based cognition.

4.4. World Model Evolution: From Reasoning over Observations to Imagining Future States

In this part, we summarize the evolution of world models in VLN. Unlike maps or memories that mainly describe what has been observed, world models aim to predict how the environment may unfold and how actions may change future observations. We discuss this evolution from three aspects: future prediction and visual imagination, foundation-model-driven world modeling, and the emerging world-action model paradigm for future VLN agents.

4.4.1. Future Prediction and Visual Imagination

Early world-model-based VLN focuses on future prediction and imagination. Dreamwalker [176] builds an internal abstract world model for continuous VLN, allowing the agent to simulate and evaluate candidate plans before taking real actions. Beyond VLN, NWM [177] further shows that action-conditioned video generation can serve as a general predictive model for visual navigation, indicating the potential of scaling world models with large egocentric video data. SALI [178] then introduces episodic simulation and episodic memory into VLN, using imagination-based memory to improve navigation in unseen environments. VLN-Imagine [179] provides an empirical study of visual imagination as additional landmark cues, showing that imagined subgoal visuals can improve navigation performance. VISTA [180] further proposes an imagine-and-align strategy, generating future visual imagination with diffusion priors and aligning it with current observations for action selection. MapDream [181] extends imagination from visual views to task-driven map learning, generating navigation-relevant map representations for instruction-conditioned planning. ThinkMatter [182] further explores panoramic-aware instructional semantics for monocular VLN, helping agents compensate for limited egocentric observations with richer panoramic context.

These methods shift VLN cognition from observe-and-act to imagine-and-act. However, generated futures can be noisy, ungrounded, or expensive to use directly. Recent work therefore explores how foundation models, latent predictive representations, and adaptive world models can make prediction more structured, efficient, and useful for navigation.

4.4.2. Foundation-Model and Self-Evolving World Models

Recent world-model methods integrate VLMs, generative models, predictive representations, and adaptive memory into navigation. NavForesee [183] unifies hierarchical language planning and dual-horizon navigation prediction, allowing a VLM to decompose instructions, track progress, and predict both short-term dynamics and long-term milestones. AstraNav-World [184] further tightens the coupling between future visual prediction and action planning, using bidirectional consistency between imagined scenes and planned trajectories to reduce error accumulation. UNeMo [185] intro-

 doi:10.20944/preprints202606.2231.v1

duces a multimodal world model that jointly reasons over visual observations, language instructions, and navigation actions, and uses predicted post-action visual states to refine navigation decisions. PROSPECT [186] moves this idea toward streaming VLN by learning latent predictive representations of both semantic and 3D spatial features, improving long-horizon robustness without adding inference-time prediction overhead. WorldMAP [187] converts world-model-generated future videos into semantic-spatial memory and planning-derived trajectory supervision through a teacher-student framework. Dual-Anchoring [188] complements these forward-prediction methods from the perspective of state consistency, using a landmark-centric world model to retrospectively verify past observations and mitigate memory drift in long-horizon VLN. CVLN-Think [189] further introduces counterfactual style adaptation for continuous VLN, highlighting the importance of causal reasoning when agents face distribution shifts and spurious visual-language correlations.

These works make world models more actionable: future prediction is not only used as imagined evidence, but also transformed into memory, supervision, latent representation, policy feedback, and state verification.

4.4.3. Toward World-Action Models for VLN

Although explicit world-action models have not yet become a mainstream solution in VLN, their rapid development in robotic manipulation and autonomous driving suggests a promising future direction. Different from conventional world models that mainly predict future observations, world-action models jointly model future world states and executable actions, thereby coupling imagination and control within the same predictive process. DreamZero [190] shows that a video-diffusion-based world-action model can serve as a zero-shot robot policy by jointly predicting future visual states and actions. Subsequent studies further examine the robustness and efficiency of this paradigm: Zhang et al. [199] compare world-action models with VLA policies under visual and language perturbations, while Fast-WAM [191] and GigaWorld-Policy [192] show that the benefits of world-action modeling can be retained with more efficient inference by reducing or decoupling test-time future imagination. In autonomous driving, Latent-WAM [193] and DriveDreamer-Policy [194] further extend this idea to latent-space or geometry-grounded world-action modeling, integrating future scene evolution with trajectory planning.

For VLN, this paradigm may offer a useful perspective because navigation requires both future scene prediction and executable action planning. A world-action formulation could jointly consider instruction-conditioned future views, action candidates, and progress states, helping agents assess whether a planned path is visually plausible and instruction-consistent before execution.

5. Learning Evolution: From Expert Imitation to Self-Improving Navigation

5.1. Open Challenges in Cognition

Although VLN cognition has evolved from instruction abstraction to spatial reasoning, deliberative planning, and world modeling, robust and faithful embodied cognition remains far from solved. First, decomposed instructions and structured constraints are still vulnerable to ambiguous language, implicit commonsense, and changing observations; an incorrect subgoal or constraint may mislead the entire navigation plan. Second, spatial reasoning remains a bottleneck for VLM-based agents, especially for quantitative distance estimation, relative orientation, occlusion reasoning, and inference over unseen spaces, which cannot be fully captured by textual scene summaries alone. Third, explicit reasoning traces improve interpretability but are not necessarily faithful to the executed policy: agents may generate plausible explanations, over-deliberate at simple steps, or propagate hallucinated landmarks through long-horizon planning. Fourth, world-model-based cognition introduces the promising ability to imagine future states, yet predicted futures can be noisy, computationally expensive, or weakly grounded in executable actions. Future VLN cognition should therefore move toward closed-loop, uncertainty-aware reasoning that can verify its own assumptions, align reasoning traces with actual decisions, and couple future prediction with physically feasible action planning.

 doi:10.20944/preprints202606.2231.v1

Perception and cognition provide the agent with the ability to observe, interpret, reason, and plan. However, a deployable VLN agent must also improve its behavior through data, feedback, interaction, and failure. Early VLN agents mainly learned by imitating expert demonstrations, which provides a strong and stable training signal but limits the agent to the distribution of annotated trajectories. As VLN moves toward real-world deployment, learning gradually evolves from expert imitation to reward-driven optimization, foundation-model-guided feedback, self-correction, and data-centric experience scaling. Figure 6 illustrates this learning evolution by organizing representative studies along four major axes: supervised alignment, reward-driven policy learning, self-improving navigation, and data-centric learning. This chapter organizes VLN learning along four evolution axes: demonstration-supervised learning, reward-driven policy learning, self-improving navigation, and data-centric learning. To further systematize this organization, Figure 7 presents a structured taxonomy that maps these axes to concrete learning capabilities, representative research directions, and typical methods.

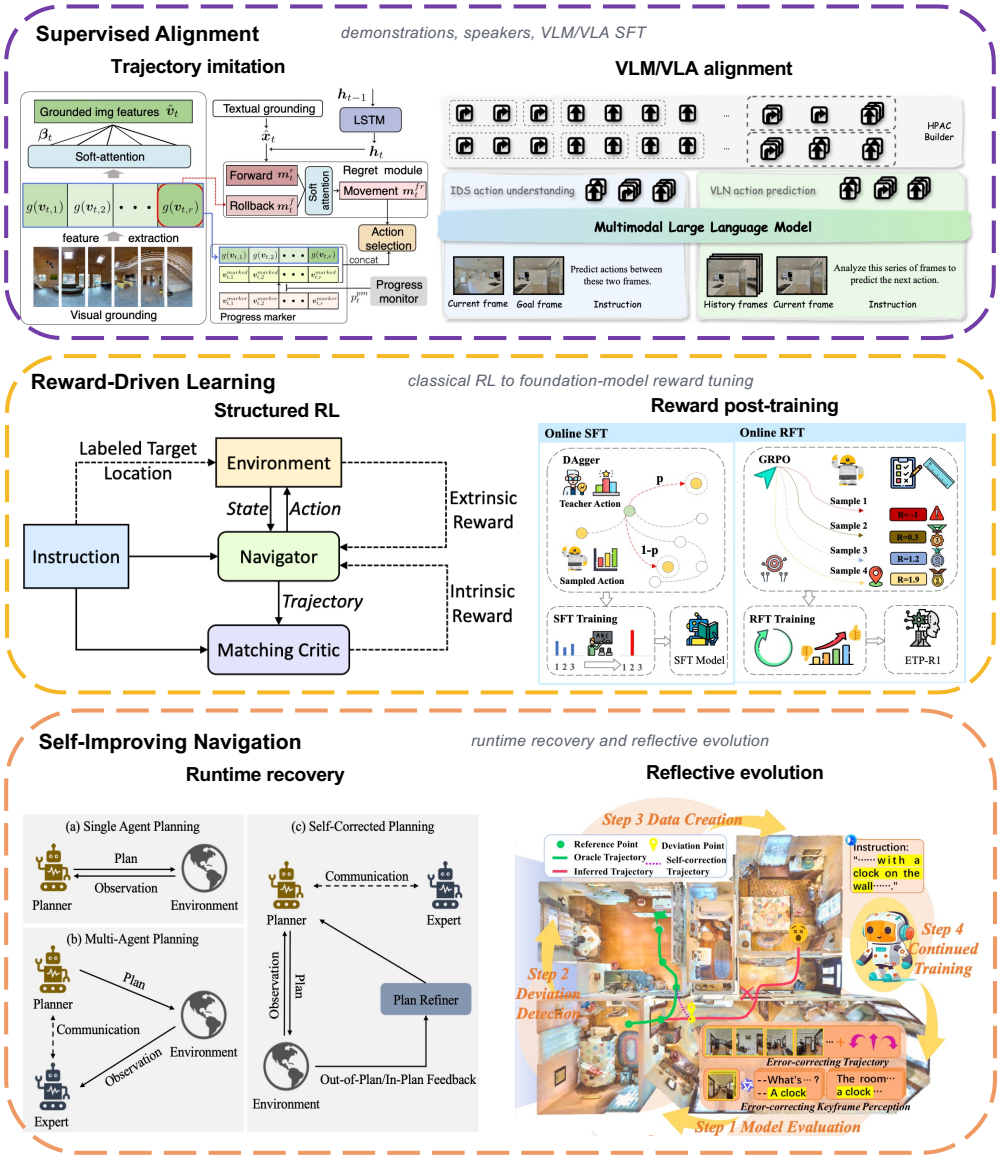

Figure 6. Learning evolution in VLN. The figure organizes representative VLN learning studies along four major axes: supervised alignment, reward-driven policy learning, self-improving navigation, and data-centric learning. These axes characterize how agents acquire navigation ability from expert demonstrations, interaction-based rewards, error-driven correction, and scalable navigation experience for more adaptive embodied navigation.

 doi:10.20944/preprints202606.2231.v1

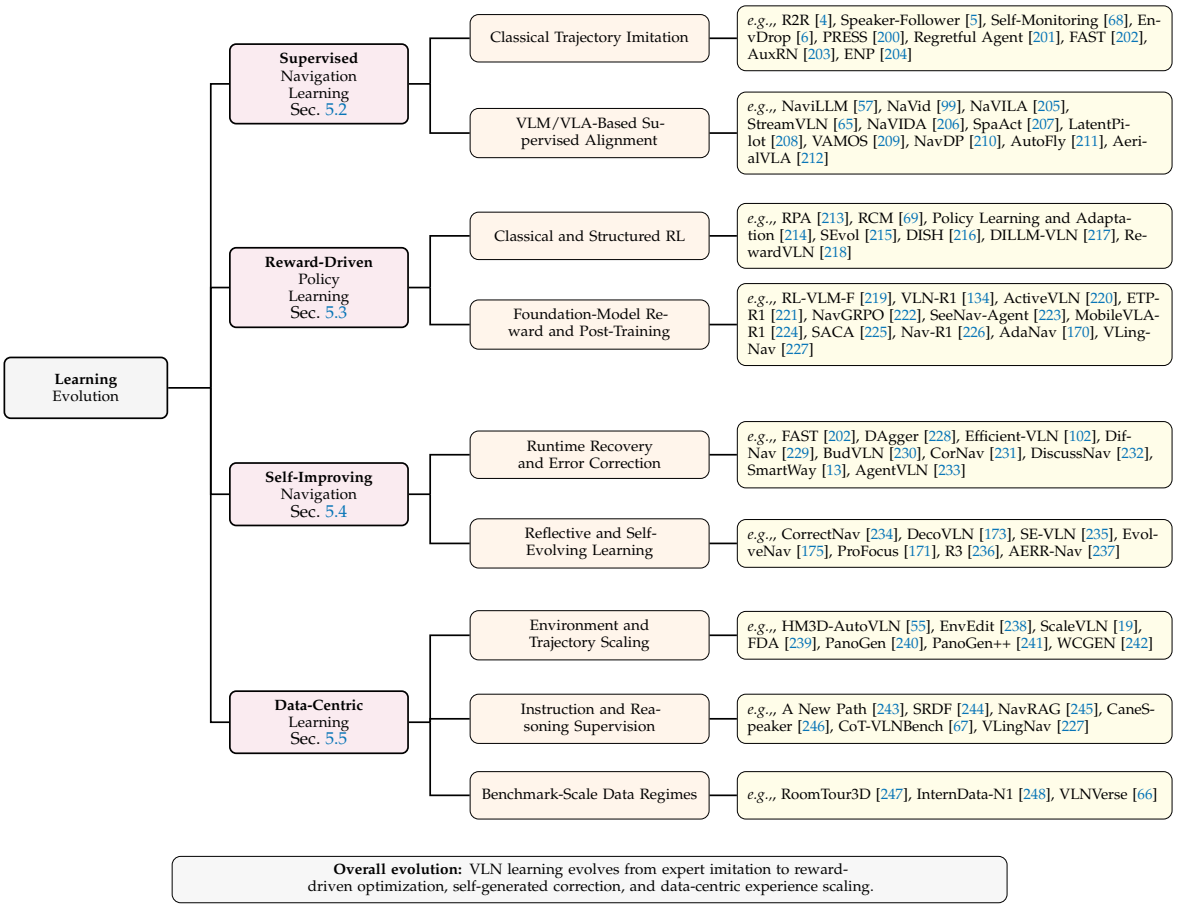

Figure 7. A structured taxonomy of learning evolution in VLN. Learning evolves along four complementary axes: supervised navigation learning, reward-driven policy learning, self-improving navigation, and data-centric learning. These axes move agents from expert imitation toward foundation-model alignment, interaction-based optimization, error-driven self-improvement, and scalable navigation experience.

5.2. Supervised Navigation Learning: From Expert Imitation to Foundation-Model Alignment

Supervised navigation learning forms the foundation of VLN training. Given expert trajectories paired with natural language instructions, early agents learn to imitate demonstrated actions under teacher-forcing, student-forcing, or behavior-cloning objectives. With the rise of foundation models, this paradigm has further evolved from training task-specific policies to aligning large VLMs with navigation actions through supervised fine-tuning. Thus, supervised learning in VLN is not only for classical trajectory imitation, but also as the main mechanism for adapting foundation models to embodied navigation.

5.2.1. Classical Trajectory Imitation

Early VLN methods are largely built on demonstration-supervised learning, where agents learn action policies from instruction-trajectory pairs under teacher-forcing, student-forcing, or behavior-cloning objectives. The R2R benchmark [4] established this standard setting in discrete panoramic environments. Speaker-Follower [5] strengthens this paradigm with speaker-generated instructions and student-forcing training, while Self-Monitoring [68] augments imitation learning with progress estimation and visual-textual co-grounding. EnvDrop [6] improves generalization through environmental dropout and mixed imitation-reinforcement learning.

A central issue in trajectory imitation is the mismatch between expert states during training and agent-generated states during inference. Several classical methods address this exposure bias from different angles. PRESS [200] reduces the gap between expert actions and sampled test-time actions through stochastic action sampling. Regretful Agent [201] uses progress estimation as a heuristic for deciding whether to move forward or roll back, while FAST [202] introduces frontier-aware search with

 doi:10.20944/preprints202606.2231.v1

backtracking to recover from poor rollout decisions. CMG-AAL [249] further alternates teacher-forcing and student-forcing to balance expert supervision and agent rollout behavior.

Classical imitation learning is also strengthened by auxiliary and multitask supervision. AuxRN [203] introduces self-supervised auxiliary reasoning tasks, including previous-action explanation, progress estimation, next-orientation prediction, and trajectory-consistency evaluation. EAML [250] jointly learns VLN and navigation from dialog history with environment-agnostic representations, improving generalization across unseen environments. CITL [251] further improves supervised navigation backbones through coarse- and fine-grained contrastive instruction-trajectory learning. ENP [204] further formulates VLN policy learning through an energy-based objective over state-action pairs, offering an alternative to standard cross-entropy imitation.

This stage teaches agents basic navigation skills from human demonstrations and expert routes. Later transformer and pre-training methods, such as VLN-BERT [15], PREVALENT [70], AirBERT [252], HOP [18], and BEVBert [130], further improve representation learning, but their downstream policies remain largely aligned with expert trajectories through supervised fine-tuning. With the rise of VLMs and VLAs, supervised learning evolves from training task-specific navigation policies to aligning large multimodal models with navigation histories, action sequences, and embodied decision formats, motivating the next stage of VLM/VLA-based alignment.

5.2.2. VLM/VLA-Based Supervised Alignment

Recent VLM/VLA-based navigation methods extend supervised learning from task-specific policy training to large-model action alignment. NaviLLM [57] casts diverse embodied navigation tasks into generation problems and trains a generalist navigation model with multi-source supervised data. NaVid [99] further shows that video-based VLMs can be adapted to navigation by predicting the next action from egocentric video streams and language instructions. StreamVLN [65] uses expert video clips and general multimodal data to train a streaming vision-language-action model. Recent works further align foundation models with transition dynamics. NaVIDA [206] augments VLN training with inverse-dynamics supervision, encouraging the model to infer actions from visual changes and learn multi-step action chunks. SpaAct [207] provides a complementary action-aware transition learning perspective by supervising backward action retrospection and forward future-state prediction. LatentPilot [208] uses future observations during training to learn action-conditioned visual dynamics, while combining latent "dream-ahead" reasoning with on-policy expert takeover for more robust navigation.

Recent VLA methods further align large multimodal models with embodied action spaces. NaVILA [205] aligns a VLA with legged robot navigation by combining high-level language-conditioned decision making with low-level locomotion skills. VAMOS [209] introduces a hierarchical VLA model that decouples semantic planning from embodiment grounding, enabling steerable navigation across different robot embodiments. This trend also extends to continuous action generation and diffusion-based navigation policies. NavDP [210] introduces a navigation diffusion policy for sim-to-real continuous navigation, showing the potential of diffusion models for smooth action generation. AutoFly [211] proposes an end-to-end UAV VLA method with progressive two-stage training, while AerialVLA [212] maps visual observations and language instructions directly to continuous aerial control signals.

These works show that supervised learning remains central in the foundation-model era. Its role has expanded from imitating expert actions to aligning large multimodal models with video streams, navigation histories, transition dynamics, and embodied control formats. However, supervised alignment still depends on curated demonstrations or collected trajectories, motivating reward-driven learning from interaction and task feedback.

5.3. Reward-Driven Policy Learning: From Passive Imitation to Trial-and-Error Optimization

Reward-driven learning allows agents to improve through interaction feedback rather than only copying expert trajectories. Compared with supervised imitation, reinforcement learning can

 doi:10.20944/preprints202606.2231.v1

encourage exploration, recovery, and trajectory-level optimization. In VLN, this line evolves from handcrafted rewards and model-based planning to structured RL, foundation-model feedback, and reinforcement fine-tuning of large navigation models.

5.3.1. Classical and Structured Reinforcement Learning

Early RL-based VLN methods mainly introduce task rewards, progress rewards, and instruction-trajectory matching signals into policy learning. RPA [213] bridges model-free and model-based reinforcement learning by predicting future states and rewards before action selection. RCM [69] and its extended policy learning formulation [214] combine imitation learning with reinforced cross-modal matching, using both extrinsic and intrinsic rewards to improve trajectory-instruction alignment. SEvol [215] further introduces reinforced structured state evolution, maintaining graph-based navigation states to support reward-driven decision making.

Later works improve RL efficiency by adding hierarchy, structure, or language decomposition. DISH [216] discovers intrinsic subgoals through hierarchical reinforcement learning, decomposing long-horizon navigation into manager-worker decisions. DILLM-VLN [217] uses an LLM to decompose complex instructions into simpler sub-instructions and trains RL agents to complete them sequentially. RewardVLN [218] further explores visual-instruction alignment rewards and planning-ahead reward estimation for hybrid RL navigation.

These methods allow agents to optimize beyond one-step expert imitation. However, handcrafted or task-specific rewards are often sparse, brittle, or weakly aligned with language grounding. Distance-based feedback can indicate whether the agent approaches the goal, but it does not fully evaluate whether each step follows the instruction semantically.

5.3.2. Foundation-Model-Guided Reward and Post-Training

Foundation models provide a more semantic source of feedback for reward-driven learning. RL-VLM-F [219] shows that VLMs can serve as zero-shot reward models by comparing visual observations with language goals, replacing manually designed rewards with vision-language feedback. Although not designed specifically for VLN, it motivates the use of foundation models as reward generators, critics, or semantic evaluators for embodied navigation.

Recent VLN works further move toward reinforcement fine-tuning and active exploration. VLN-R1 [134] applies a two-stage SFT–RFT pipeline to continuous VLN, first aligning LVLM action predictions with expert demonstrations and then optimizing them with GRPO [253] and time-decayed rewards. ActiveVLN [220] explicitly enables active exploration through multi-turn RL, using a small amount of imitation learning for initialization and then optimizing self-collected rollouts with GRPO. ETP-R1 [221] applies reinforcement fine-tuning to graph-based VLN-CE, combining large-scale topological pretraining with online GRPO optimization. NavGRPO [222] encourage active exploration through outcome-based rewards, while SeeNav-Agent [223] and MobileVLA-R1 [224] introduce step-level or reasoning-action rewards to reduce sparse-feedback issues. SACA [225] addresses sparse outcome rewards by introducing step-aware contrastive alignment, which extracts dense supervision from imperfect trajectories and identifies valid prefixes and divergence points. Nav-R1 [226] further extends GRPO-style post-training to embodied reasoning and navigation by combining cold-start CoT initialization with format, understanding, and navigation rewards.

Reward-driven learning is also being integrated with adaptive reasoning. AdaNav [170] learns an uncertainty-aware reasoning policy through a heuristics-to-RL training process, dynamically deciding when deeper reasoning is needed. VLingNav [227] incorporates an online expert-guided reinforcement learning stage to move beyond pure imitation learning. In broader VLA navigation, UrbanVLA [254] adopts a two-stage SFT–RFT pipeline for urban micromobility, while TIC-VLA [255] combines imitation learning with online reinforcement learning to handle delayed reasoning and real-time control in dynamic environments.

Overall, reward-driven learning shifts VLN from passive imitation toward interaction-based optimization. The recent trend is not merely to design better rewards, but to use foundation models,

 doi:10.20944/preprints202606.2231.v1

dense step-level feedback, active exploration, and reinforcement fine-tuning to improve navigation policies.

5.4. Self-Improving Navigation: From External Feedback to Internal Error Correction

Self-improving navigation shifts learning from external supervision toward internal error diagnosis and correction. Instead of only receiving rewards after acting, the agent learns to detect deviations, recover from wrong decisions, reuse failure cases, and refine its future behavior. This is particularly important for long-horizon VLN, where small early mistakes can accumulate into large trajectory errors.

5.4.1. Runtime Recovery and Error-Driven Correction

Early correction mechanisms mainly focus on recovering from poor rollout decisions during execution. FAST [202] introduces frontier-aware search with backtracking, allowing the agent to roll back from unreliable paths and resume exploration from more promising states. This represents an early form of runtime correction, where navigation errors are handled through search and trajectory-level recovery.

DAgger-based methods provide another important form of error correction by addressing covariate shift during learning. DAgger [228] queries an oracle for corrective actions under the agent's own state distribution and iteratively aggregates these samples into training. Building on this idea, Efficient-VLN [102] introduces a dynamic hybrid policy that progressively adjusts the use of oracle guidance during navigation, while DifNav [229] incorporates DAgger-based online policy training and expert trajectory augmentation to improve robustness in erroneous states. BudVLN [230] further improves DAgger-style correction by addressing instruction-state misalignment, using retrospective rectification to synthesize semantically consistent corrective trajectories.

Recent foundation-model-based agents further extend correction to zero-shot and continuous VLN. CorNav [231] uses environmental feedback and multiple domain experts to refine future plans and adjust actions in zero-shot VLN. DiscussNav [232] performs multi-expert discussion before each movement, covering instruction understanding, scene perception, completion estimation, and decision checking. SmartWay [13] improves zero-shot VLN-CE with enhanced waypoint prediction, history-aware reasoning, and adaptive backtracking. AgentVLN [233] further integrates context-aware self-correction and active exploration into an agentic VLN framework, enabling recovery from occlusions and long-horizon error accumulation.

5.4.2. Reflective and Self-Evolving Learning

A second line turns errors and experiences into reusable training or reasoning signals. Beyond oracle-assisted correction, CorrectNav [234] introduces a self-correction flywheel for VLA navigation: model errors are converted into action-correction trajectories and perception-correction keyframes, which are then used to iteratively retrain the model. DecoVLN [173] decouples navigation process into observation, reasoning, and correction, using adaptive memory refinement and state-action pair-level corrective finetuning to reduce compounding errors in long-horizon navigation.

Recent works further move toward reflective and self-evolving agents. SE-VLN [235] stores successful and failed cases as reusable experience, retrieves them for thought-based reasoning, and uses reflection to support continual evolution during testing. EvolveNav [175] improves LLM-based VLN through formalized CoT supervised fine-tuning and self-reflective post-training, encouraging the model to learn correct reasoning patterns by contrasting them with wrong ones. ProFocus [171] introduces a perception-reasoning loop that actively queries missing visual information and focuses reasoning on high-value historical waypoints. R3 [236] further uses a regulator to decide when to rely on fast expert navigation and when to trigger slower multimodal reasoning. Related embodied navigation work such as AERR-Nav [237] also highlights adaptive switching among exploration, recovery, and reminiscing states for robust navigation in unknown environments.

 doi:10.20944/preprints202606.2231.v1

Together, these works shift VLN learning from externally specified supervision to self-generated improvement. The agent no longer treats mistakes only as failures, but as signals for recovery, reflection, and future policy refinement. Nevertheless, self-improvement is bounded by the diversity and quality of available experiences, which highlights the importance of data-centric learning.

5.5. Data-Centric Learning: From Limited Trajectories to Scalable Navigation Experience

Data-centric learning addresses a different bottleneck in VLN: the limited scale, diversity, and quality of navigation experience. Its evolution can be understood along three progressive levels. First, embodied experience scaling expands the environments, trajectories, visual observations, and generated worlds from which agents learn. Second, supervision signal refinement improves the quality of language, alignment, and reasoning supervision associated with these experiences. Third, benchmark-scale data regimes move beyond isolated datasets toward realistic, embodiment-aware data ecosystems. Consequently, VLN learning shifts from optimizing fixed datasets to constructing scalable and informative navigation experiences.

5.5.1. Environment and Trajectory Scaling

Early data-centric methods expand the visual and spatial diversity of training environments. HM3D-AutoVLN [55] automatically constructs large-scale VLN data from unlabeled 3D buildings by generating navigation graphs, pseudo object labels, and language instructions. EnvEdit [238] edits existing environments to create new visual variations, improving generalization to unseen scenes. ScaleVLN [19] further scales synthetic instruction–trajectory generation across HM3D and Gibson environments, producing millions of augmented navigation samples. FDA [239] complements these methods by perturbing visual observations in the frequency domain to improve robustness.

Generative environment augmentation further increases visual diversity. PanoGen [240] uses text-conditioned diffusion models to generate panoramic environments for VLN, while PanoGen++ [241] improves this direction with domain-adapted text-guided panoramic generation through inpainting and outpainting. WCGEN [242] emphasizes world-consistent data generation, aiming to improve data diversity while preserving physical and spatial consistency. Together, these works move VLN from reusing existing environments to generating more diverse and visually grounded navigation experiences.

5.5.2. Instruction and Reasoning Supervision

Another line scales the language and supervision side of VLN. A New Path [243] shows that large-scale in-domain instruction augmentation with imitation learning can substantially improve instruction-following agents. SRDF [244] introduces a self-refining data flywheel that iteratively improves both data quality and model learning. NavRAG [245] further generates user-demand instructions through retrieval-augmented LLMs, using hierarchical scene descriptions and simulated user roles to produce more realistic navigation requests. CaneSpeaker [246] explores LLM-assisted instruction generation to make navigation instructions more human-like.

Recent works also provide reasoning-oriented benchmarks. CoT-VLNBench [67] constructs a benchmark for visual chain-of-thought reasoning in vision-language-navigation robots, enabling more fine-grained evaluation and supervision of navigation reasoning. VLingNav [227] provides large-scale adaptive chain-of-thought navigation data, Nav-AdaCoT-2.9M, supporting supervised learning of reasoning-aware navigation behaviors. These works indicate that data-centric learning is no longer limited to more instructions, but increasingly focuses on richer alignment and reasoning traces.

5.5.3. Benchmark-Scale Data Regimes

Beyond individual augmentation methods, recent work begins to construct benchmark-scale data regimes that support generalist and embodiment-aware navigation learning. Recent datasets further broaden the data regime from synthetic indoor trajectories to real-world videos, generalist navigation corpora, and more realistic embodied benchmarks. RoomTour3D [247] introduces geometry-aware

 doi:10.20944/preprints202606.2231.v1

video-instruction tuning from web-based room-tour videos, providing real-world trajectories, instructions, and action-enriched supervision for embodied navigation. InternData-N1 [248] scales navigation data across simulation platforms and robot embodiments, supporting more general foundation-model training. VLNVerse [66] provides a versatile embodied simulation and evaluation benchmark, expanding VLN toward more realistic environments, embodiments, and action spaces. These benchmark-scale data regimes are not only evaluation platforms; they also define the environments, annotation styles, action spaces, and embodiment conditions from which future data-centric learning can be built.

Overall, data-centric learning expands the experience base of VLN agents from limited expert trajectories to scalable environments, generated observations, fine-grained annotations, reasoning data, real-world videos, and embodied benchmark ecosystems.

5.6. Open Challenges in Learning

Despite the shift from expert imitation to reward-driven optimization, self-improving navigation, and data-centric learning, several learning challenges remain open. First, supervised alignment still inherits the coverage and bias of demonstrations; agents trained on expert trajectories may fail in off-path states, rare instructions, or embodiment-specific action spaces. Second, reward-driven post-training offers a way to optimize beyond demonstrations, but navigation rewards are often sparse, delayed, and only partially aligned with instruction fidelity, making stable and semantically meaningful reinforcement learning difficult. Third, self-correction and reflective learning depend on reliable failure detection and error attribution; otherwise, agents may convert their own mistakes into misleading training signals or overfit to self-generated experience. Fourth, data-centric scaling increases environment and instruction diversity, but synthetic trajectories, generated instructions, and benchmark-scale corpora still require careful filtering, grounding verification, and sim-to-real validation. Future VLN learning should therefore combine scalable data generation with principled data quality control, dense semantic feedback, off-policy recovery supervision, and evaluation protocols that measure not only final success but also robustness under distribution shift, recovery from mistakes, and transfer across environments and embodiments.

6. Generalization Evolution: From Closed Benchmarks to Open-World Deployment

Perception, cognition, and learning endow VLN agents with environment observation, instruction comprehension, path planning, and behavioral optimization. However, these capabilities have largely been developed within closed benchmark ecosystems [2,4,31], where agents learn from limited environments and are evaluated on held-out buildings from the same dataset. As VLN advances toward real-world deployment, a fundamental question emerges: how can these capabilities transfer to open, dynamic, and continuously evolving environments? This marks a shift from capability construction to generalization evolution [9,80]. Figure 8 illustrates this generalization evolution by organizing representative studies along five major axes: environment generalization, horizon generalization, lifelong adaptation, scene generalization, and safety generalization. This chapter is organized along five dimensions: environment generalization, horizon generalization, lifelong adaptation, scene generalization, and safety generalization. To further systematize this organization, Figure 9 presents a structured taxonomy that maps these dimensions to representative generalization capabilities, research directions, and typical methods. Along this trajectory, generalization has evolved from a closed-dataset metric into a core capability for sustained open-world operation.

 doi:10.20944/preprints202606.2231.v1

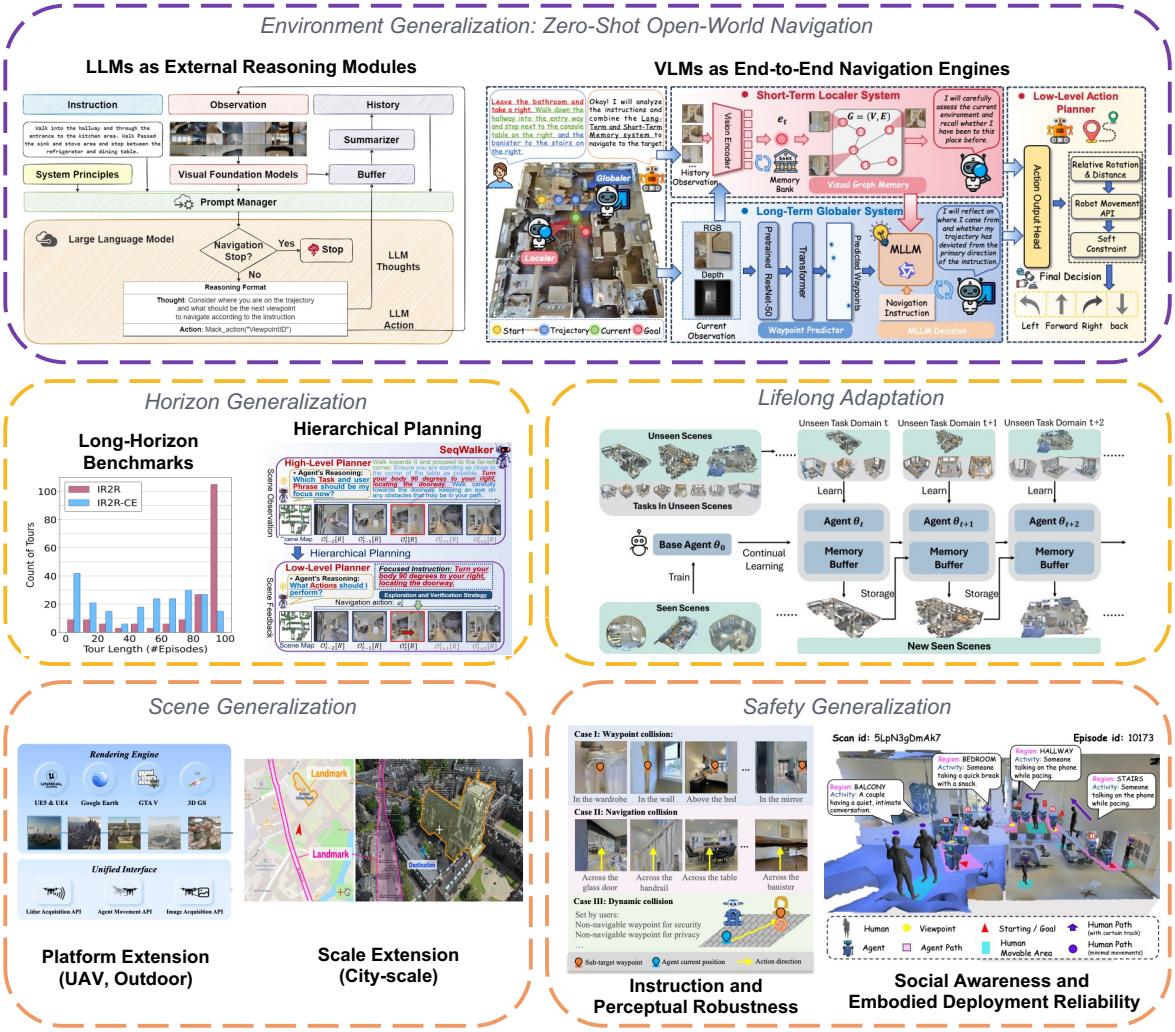

Figure 8. Generalization evolution in vision-language navigation. The figure organizes representative VLN generalization studies along five major axes: environment generalization, horizon generalization, lifelong adaptation, scene generalization, and safety generalization. These axes characterize how VLN agents evolve from closed-set benchmark evaluation toward zero-shot open-world navigation, long-horizon agentic decision making, continual self-evolution, cross-platform and city-scale deployment, and trustworthy real-world operation.

 doi:10.20944/preprints202606.2231.v1

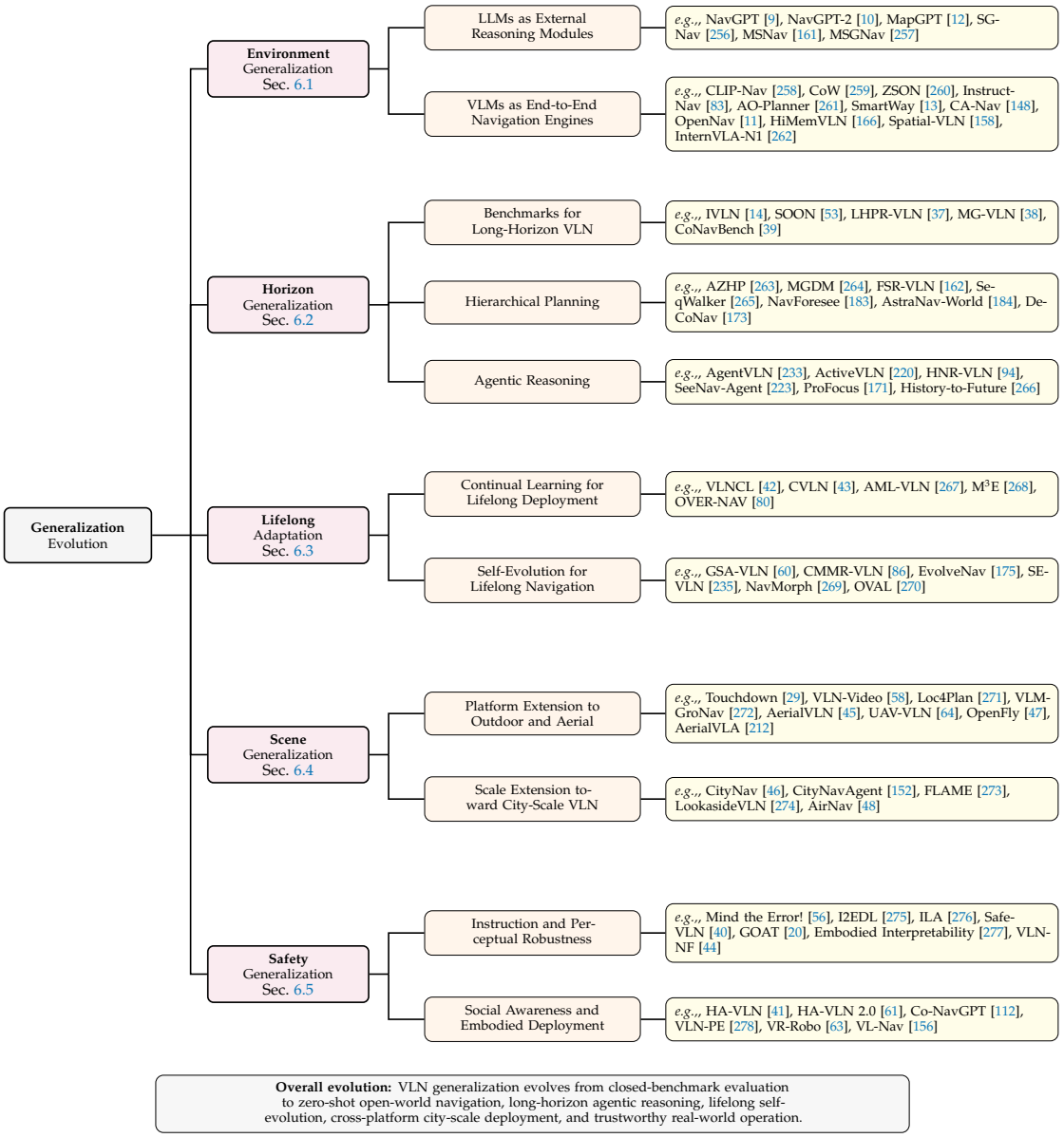

Figure 9. A structured taxonomy of generalization evolution in VLN along five dimensions: environment generalization, horizon generalization, lifelong adaptation, scene generalization, and safety generalization, collectively advancing VLN from benchmark-oriented evaluation toward sustained and reliable open-world operation.

6.1. Environment Generalization: From Closed-Set Evaluation to Zero-Shot Open-World Navigation

The generalization of conventional VLN methods is primarily constrained by the environmental distribution covered by their training data, leading to substantial performance degradation under distribution shift in open scenarios. Foundation models offer a promising alternative, as the world knowledge and spatial priors acquired through large scale multimodal pretraining enable zero-shot VLN without task specific training. Existing methods have evolved in two stages, from using LLMs as reasoning modules to employing VLMs as end to end navigation engines.

6.1.1. LLMs as External Reasoning Modules

Early zero-shot VLN methods employ LLMs as reasoning engines, translating visual observations into structured text and leveraging commonsense knowledge to guide navigation. NavGPT [9] is the first purely LLM-driven VLN agent, feeding serialized panoramic descriptions into GPT-4 for chain-of-thought action generation, but reveals failures in directional reasoning and spatial grounding.

 doi:10.20944/preprints202606.2231.v1

NavGPT-2 [10] addresses this by incorporating visual features, demonstrating that combining LLM reasoning with visual perception substantially improves zero-shot performance.

Subsequent works externalize spatial burdens into explicit structures. MapGPT [12] converts an online topological map into text prompts for global path planning. SG-Nav [256] extends this to hierarchical 3D scene graphs with chain-of-thought prompting over object-, group-, and room-level nodes. MSNav [161] further introduces hierarchical spatial memory spanning room-level relations and cross-room topology, enabling multi-scale reasoning. More recently, MSGNav [257] extends this paradigm by replacing traditional text-based scene graph relations with dynamically allocated multi-modal 3D scene graphs (M3DSG), preserving fine-grained visual information for zero-shot embodied navigation.

Together, LLM-based VLN has evolved from pure language reasoning, through visual feature injection, to external topological scaffolds and hierarchical memory. However, fine-grained structures in continuous observations are inevitably lost regardless of compression strategy, motivating the shift toward end-to-end VLM-based navigation. However, fine-grained structures in continuous observations are inevitably lost regardless of compression strategy, motivating the shift toward end-to-end VLM-based navigation.

6.1.2. VLMs as End-to-End Navigation Engines

LLM-based zero-shot VLN is inherently constrained by the modality conversion bottleneck, where fine-grained visual structures are irreversibly compressed into symbolic text. VLMs [279–281] offer a direct remedy by jointly processing visual and linguistic inputs. Early CLIP-based methods (e.g., CLIP-Nav [258], CoW [259], ZSON [260]) validated cross-modal matching but achieved limited zero-shot success due to underused reasoning capacity.

Subsequent VLMs with dialogue-based reasoning address this limitation. InstructNav [83] employs GPT-4V as a zero-shot task planner with spatial-action prompts, but formulates subtasks solely as semantic goals. AO-Planner [261] integrates high-level planning with affordance-based path generation in a zero-shot closed loop. SmartWay [13] strengthens constraint awareness via DINOv2 [282] with adaptive backtracking for error correction. Spatial-VLN [158] further addresses the spatial perception bottleneck of zero-shot VLN by augmenting VLM-based agents with explicit spatial perception modules, compensating for the spatial reasoning weakness of foundation models through depth-aware geometric grounding for more reliable zero-shot navigation. Beyond modular augmentation, InternVLA-N1 [262] frames VLN as a native foundation-model capability through a dual-system architecture that integrates a high-level VLM planner with a low-level diffusion policy. This design mitigates the slow inference, discrete token-action outputs, and weak geometric-spatial reasoning of prior VLM/LLM-based VLN systems, enabling real-time reactive planning, smoother continuous trajectory generation, and more reliable spatial grounding. By using learned latent plans instead of ambiguous pixel goals, it achieves improved performance with zero-shot cross-embodiment transfer. Despite these advances, per-step VLM invocation creates a deployment bottleneck in latency and API cost. CA-Nav [148] restructures the pipeline into one-time instruction decomposition followed by continuous constraint monitoring via BLIP-2 [283] and Grounding DINO [120]. OpenNav [11] demonstrates that open-source VLMs [281] with external modules enable real-robot deployment free from commercial API dependence. HiMemVLN [166] identifies the navigation amnesia problem and introduces hierarchical memory to close the gap with commercial counterparts.

Overall, VLM-based navigation has evolved from shallow perception bridging the modality gap, through integrated perception-reasoning engines, to deployment-aware designs jointly optimizing inference cost, memory persistence, and operational robustness.

6.2. Horizon Generalization: From Short-Horizon Instruction Following to Long-Horizon Agentic Navigation

Standard VLN benchmarks such as R2R [4] and RxR [31] typically involve only 5-8 decision steps, far below real-world complexity where tasks require hundreds of consecutive decisions. In contrast, tasks such as multi room service and building scale guidance require hundreds of consecutive

 doi:10.20944/preprints202606.2231.v1

decisions and multi floor traversal. These settings introduce three core challenges: error accumulation over extended trajectories, memory decay in long temporal contexts, and exploration-exploitation imbalance without global topological priors. To address these, long-horizon VLN research has evolved from benchmark redefinition to hierarchical planning and agentic autonomous navigation.

6.2.1. Benchmarks and Evaluation for Long-Horizon VLN

Long-horizon VLN requires evaluation protocols that capture challenges absent from standard benchmarks. IVLN [14] initiates this effort by extending episodic evaluation to persistent environments for cross-episode knowledge accumulation, while SOON [53] complements it with situation-oriented object navigation requiring multi-stage compositional reasoning. Building on these foundations, LHPR-VLN [37] establishes the first dedicated long-horizon benchmark with trajectories exceeding 50 steps and abstract goal interpretation, accompanied by the NavGen platform for scalable scenario generation. Extending beyond single-task evaluation, MG-VLN [38] introduces multi-goal sequential navigation that jointly assesses planning capability and execution persistence, and CoNavBench [39] further scales to multi-agent collaboration across over 4,000 episodes evaluating both individual competence and coordination efficiency.

Together, these benchmarks trace a progressive evolution from single-agent knowledge accumulation, through multi-goal planning, to multi-agent collaboration, collectively delineating the problem space for long-horizon methods.

6.2.2. Hierarchical Planning for Long-Horizon Navigation

The core complexity of long-horizon tasks stems from the granularity mismatch between abstract goals and atomic actions. Early work addresses this through spatial decomposition: AZHP [263] separates zone-level subgoal planning from intra-zone execution, and MGDM [264] advances this by explicitly decoupling LLM-based semantic planning from low-level action execution, confining error accumulation within subgoal boundaries. Subsequent methods further improve efficiency: FSR-VLN [162] draws on fast-slow dual-system theory to reduce inference latency through CLIP-based candidate filtering with uncertainty-triggered VLM verification, while SeqWalker [265] extends hierarchical planning to multi-goal settings with cross-stage state consistency.

More recent works push beyond reactive decomposition toward predictive reasoning. NavForesee [183] introduces a unified world model for dual-horizon prediction that simultaneously anticipates observations and subgoal completion probabilities, and AstraNav-World [184] further reduces error accumulation through bidirectional consistency between imagined scenes and planned trajectories. DeCoNav [173] complements this with dialogue-enhanced collaboration, demonstrating that solicited human feedback can exceed the performance ceiling of purely autonomous planning.

Together, hierarchical planning reflects a progressive shift from reactive decomposition, through predictive reasoning, to collaborative problem solving.

6.2.3. Agentic Reasoning for Long-Horizon Navigation

While hierarchical planning alleviates granularity mismatch, it assumes passive execution of predefined plans. In practice, long-horizon deployment inevitably encounters trajectory drift and unforeseen changes, requiring autonomous failure detection and recovery. AgentVLN [233] addresses this comprehensively by integrating task decomposition, persistent memory, adaptive exploration, and context-aware self-correction, while ActiveVLN [220] complements it through multi-turn reinforcement learning that enables exploration strategies to emerge from self-generated experiences.

Beyond reactive recovery, anticipatory capabilities further deepen agent autonomy. HNR-VLN [94] leverages neural radiance fields to synthesize observations from unvisited viewpoints before irreversible decisions, transforming navigation into a predictive process. SeeNav-Agent [223] reduces perceptual hallucinations via dual-view prompts with step-level reward optimization, and Pro-Focus [171] introduces a training-free, reasoning-driven perception loop for fine-grained recognition within key regions. History to Future [266] deepens agentic reasoning through an experience-and-

 doi:10.20944/preprints202606.2231.v1

thought paradigm, leveraging historical trajectories and predictive thoughts to guide future decisions in continuous environments.

Together, agentic reasoning reflects a progressive deepening from reactive execution toward proactive perception, self-correction, and continual optimization.

6.3. Lifelong Adaptation: From Episodic Isolation to Continual Learning and Self-Evolution

The preceding sections share a key assumption: agents are reset after each episode with no knowledge carried across tasks. Real-world deployment, however, demands continuous operation in dynamically changing environments, introducing two core challenges: catastrophic forgetting, where learning in new environments degrades prior capabilities, and knowledge accumulation, where agents must integrate cross-episode experiences into reusable priors. To address these, lifelong adaptation research has evolved from stateless episodic deployment toward autonomous agents integrating continual learning and self-evolution.

6.3.1. Continual Learning for Lifelong Deployment

Standard VLN evaluation deploys agents with frozen parameters, overlooking distribution drift in continual deployment. VLNCL [42] first formalizes this by modeling environments as sequential data streams with Dual-loop Scenario Replay balancing forgetting mitigation and knowledge accumulation. CVLN [43] extends this to cross-domain settings where visual and layout shifts impose stronger forgetting pressure, while AML-VLN [267] introduces TuKA, a parameter-efficient Tucker decomposition storing shared knowledge in core tensors with condition-specific adaptation in factor matrices.

However, replay-based methods face scalability bottlenecks as storage grows linearly with environments. M3E [268] addresses this through a replay-free mixture-of-experts framework encoding new knowledge via environment-aware hierarchical gating without disrupting existing parameters, and OVER-NAV [80] further improves reuse by organizing semantics from LLMs and open-vocabulary detectors [120,282] into a persistent omnigraph for structured cross-visit retrieval.

Together, VLN continual learning has evolved from replay-based anti-forgetting, through parameter-efficient decomposition, to replay-free expert architectures approaching lifelong scalability.

6.3.2. Self-Evolution for Lifelong Navigation

Continual learning provides forgetting-resistant support, but lifelong navigation further requires self-evolution: extracting reusable knowledge from the agent's own experience without external supervision. GSA-VLN [60] establishes this through passive experience retention via graph-based memory with environment-specific online training. CMMR-VLN [86] upgrades this into active refinement through structured reflection, enabling causal attribution on failures and historical experience retrieval.

Recent methods advance toward reasoning-level self-improvement. EvolveNav [175] integrates chain-of-thought training with self-reflective post-training for iterative reasoning correction, while SE-VLN [235] achieves training-free test-time self-evolution through hierarchical memory and retrieval-augmented reflection, providing first evidence of sustained test-time gains. NavMorph [269] complements this by adaptively modeling latent representations for foresight-based decision-making, and OVAL [270] extends self-evolution to object-level lifelong navigation with selective forgetting for memory compactness.

Together, self-evolving navigation reflects a progressive shift from passive adaptation, through active reflection, to training-free continuous evolution.

6.4. Scene Generalization: From Structured Indoor Environments to Cross-Platform and City-Scale Navigation

The preceding dimensions of generalization have been studied primarily within the structured indoor ecosystem of Matterport3D [284]. However, real-world deployment demands navigation in unstructured open-world settings, from last-mile delivery across mixed indoor-outdoor spaces to aerial inspection and outdoor search and rescue, introducing cross-platform and cross-scale challenges

 doi:10.20944/preprints202606.2231.v1

that structured indoor methods cannot address. Accordingly, recent research has advanced scene generalization along two directions: cross-platform embodiment extension and cross-scale environmental expansion.

6.4.1. Platform Extension from Indoor Ground Navigation to Outdoor Navigation

Prior to 2023, VLN research was almost exclusively confined to structured indoor environments. Ground-level extension began with Touchdown [29], constructing the first outdoor VLN benchmark on Google Street View. VLN-Video [58] addresses data scarcity by generating instructions from driving videos across multiple cities, while Loc4Plan [271] introduces a locate-before-plan paradigm grounding spatial position against street-view landmarks, demonstrating that outdoor VLN demands reasoning strategies absent from indoor methods. VLM-GroNav [272] further extends to unstructured terrains by integrating VLMs with proprioceptive sensing for terrain-aware navigation.

Parallel to ground extension, aerial platforms address an equally critical gap. AerialVLN [45] constructs the first city-scale UAV benchmark covering five cities with trajectories exceeding 200 meters, though its baselines rely on indoor-style waypoint graphs misaligned with flight dynamics. UAV-VLN [64] addresses this through end-to-end regression of velocity vectors and yaw rates from monocular inputs, OpenFly [47] contributes data-level advances with 100K trajectories and keyframe-aware modeling, while AerialVLA [212] introduces a VLA paradigm that tokenizes continuous control for autoregressive trajectory generation.

Together, platform extension has progressively expanded VLN from structured indoor environments, through outdoor streets and unstructured terrains, to city-scale aerial settings.

6.4.2. Scale Extension toward City-Scale VLN

Platform extension addresses in what environment navigation occurs, while scale extension further addresses over what spatial extent it operates. Indoor VLN typically spans tens of meters, whereas city-scale tasks involve kilometer-range planning and geographic commonsense reasoning far exceeding any single visual model. CityNav [46] establishes the data foundation with the first city-scale aerial navigation dataset comprising 32,637 trajectories across three cities. However, visual observations alone are insufficient at this scale, as landmark-associated geographic commonsense requires world knowledge from large-scale pretraining. CityNavAgent [152] accordingly introduces LLM-based geographic reasoning, decomposing complex city instructions into executable subgoal sequences mapped to visual recognition or local planning. FLAME [273] further demonstrates that primary gains at city scale arise from rearchitecting agents around multimodal LLM primitives with perceiver-style cross-attention over street-view sequences, rather than extrapolating indoor methods outward. LookasideVLN [274] reveals that aerial scale extension simultaneously demands representational reconstruction, encoding UAV orientation as a joint yaw-pitch-roll representation through three-dimensional orientation attention. More recently, AirNav [48] establishes a large-scale UAV VLN benchmark constructed from real urban aerial data with natural and diverse instructions, providing a unified evaluation foundation for city-scale aerial navigation in the era of multimodal large language models.

Together, scale extension reflects a progressive deepening from street-level perception toward city-level geographic reasoning, advancing VLN from local environmental understanding to global spatial cognition.

6.5. Safety Generalization: From Controlled Simulation to Trustworthy Real-World Deploymen

The preceding dimensions of generalization implicitly assume agents operate within safe, controlled simulations. Real-world deployment requires navigation agents not only to reach targets accurately but also to meet trustworthiness standards in instruction robustness, perceptual reliability, social norm compliance, and embodied transfer fidelity. Accordingly, recent research advances safety generalization along two complementary directions: instruction and perceptual robustness targeting input-side reliability, and social-aware embodied deployment targeting output-side reliability.

 doi:10.20944/preprints202606.2231.v1

6.5.1. Instruction and Perceptual Robustness

The preceding dimensions assume accurate instructions and ideal perceptual conditions, yet real-world deployment involves erroneous instructions and visual disturbances from lighting and occlusion. Robustness research first emerges at the instruction level, where Mind the Error! [56] systematically exposes agents' lack of critical evaluation for erroneous instructions, and I2EDL [275] advances this toward interactive error resolution through proactive user querying. Research then extends to the perceptual level, with ILA [276] demonstrating that minor illumination changes alone cause substantial navigation degradation through adversarial lighting manipulation. These findings motivate deeper improvements at the training paradigm level: Safe-VLN [40] incorporates collision avoidance as an explicit objective proving safety and performance are not mutually exclusive, while GOAT [20] more fundamentally eliminates spurious correlations through causal reasoning, establishing structural defenses against distribution shift. Building on this causal perspective, Embodied Interpretability [277] further investigates how VLA policies under distribution shift may rely on spurious visual correlations rather than task-relevant causes, formulating embodied interpretability as the bridge between causal understanding and generalization. VLN-NF [44] complements these with feasibility-aware navigation, exposing agents' brittleness to false-premise instructions where targets may be absent from specified rooms.

6.5.2. Social Awareness and Embodied Deployment Reliability

The second line of safety generalization addresses reliable transfer from static unpopulated environments toward dynamic human-populated scenes and real physical worlds. In social navigation, HA-VLN [41] first incorporates dynamic human activities requiring agents to respect personal space, HA-VLN 2.0 [61] demonstrates that explicit social modeling simultaneously improves success rate and collision avoidance, and Co-NavGPT [112] extends social constraints to multi-robot collaboration. Building upon this, research advances into embodied sim-to-real transfer: VLN-PE [278] reveals distinct failure modes across humanoid, quadruped, and wheeled embodiments through a physical-level platform, VR-Robo [63] improves transfer reliability via high-fidelity digital twins in a real-to-sim-to-real framework, 3D feature fields [278] bridge the modality gap by providing consistent representations across simulated and real environments, and VL-Nav [156] validates efficient architectures meeting real-time deployment requirements.

6.6. Open Challenges in Generalization

Despite the rapid evolution, several challenges remain open. First, foundation-model-based zero-shot navigation relies on iterative LLM/VLM prompting that incurs prohibitive inference cost, calling for efficient strategies that support real-time and long-horizon consistent tracking. More importantly, strong performance on existing benchmarks does not necessarily translate into robust real-world deployment, as agents may overfit to dataset-specific layouts, instruction styles, or action distributions. A promising direction is to move toward VLM-native navigation architectures and Reinforcement Learning with Verifiable Rewards (RLVR)-based post-training, enabling navigation capabilities to be learned through interaction and environmental feedback rather than relying solely on benchmark-specific supervision. Second, long-horizon and lifelong navigation are bottlenecked by catastrophic forgetting and memory-compute trade-offs, where deciding what to retain and forget in dynamic environments under limited compute remains a critical challenge for continual learning. Third, cross-embodiment transfer and dynamic environment adaptation remain largely unexplored, lacking unified embodiment-agnostic representations and rapid re-adaptation mechanisms to dynamic changes. Fourth, safety generalization still lacks a unified framework spanning instruction robustness, adversarial perturbations, privacy, and social compliance. Finally, success-rate-centric evaluation neither characterizes open-world distribution shifts nor diagnoses reliability, safety, or persistence. Future VLN generalization should therefore move toward unified, deployment-aware frameworks that

 doi:10.20944/preprints202606.2231.v1

jointly optimize spatial reasoning, efficient inference, lifelong adaptation, cross-embodiment transfer, and trustworthy operation, supported by multidimensional open-world evaluation protocols.

7. Conclusion

VLN is undergoing a paradigm shift from passive instruction following to autonomous cognitive navigation. To characterize this transition, this survey proposes a unified analytical framework organized around evolutionary logic. The framework systematically reviews recent advances in VLN across four progressive dimensions: perception, cognition, learning, and generalization, while revealing the intrinsic connections and evolutionary mechanisms among different technical paradigms. Specifically, at the perception level, VLN has evolved along three axes, namely semantic granularity, spatial structure, and input realism, from panoramic vision and language alignment toward contextualized spatial understanding. At the cognition level, VLN has evolved from reactive decision making to world model driven predictive planning, shifting navigation from immediate perception based action to prospective reasoning over internal representations. At the learning level, VLN has expanded from expert imitation to reward driven autonomous optimization and self correction, enabling agents to continuously improve through experience. At the generalization level, research has progressed across five dimensions, namely environment, horizon, lifelong adaptation, scenario diversity, and safety, moving the field from closed benchmark evaluation toward trustworthy deployment in open world settings. This evolutionary trajectory reshapes the meaning of navigation intelligence, which is no longer confined to success rate metrics on individual benchmarks but is instead reflected in an integrated capability that combines multimodal perception, autonomous learning, and open world generalization. Building on these multidimensional technical developments, this survey further identifies several fundamental open challenges and outlines future directions, with the aim of providing the VLN community with a clear roadmap for capability development and advancing VLN from an instruction follower toward a genuinely embodied cognitive navigator.

References

1. Gu, J.; Stefani, E.; Wu, Q.; Thomason, J.; Wang, X. Vision-and-language navigation: A survey of tasks, methods, and future directions. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2022, pp. 7606–7623.
2. Krantz, J.; Wijmans, E.; Majumdar, A.; Batra, D.; Lee, S. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In Proceedings of the European Conference on Computer Vision. Springer, 2020, pp. 104–120.
3. Plikynas, D.; Žvironas, A.; Budrionis, A.; Gudauskis, M. Indoor navigation systems for visually impaired persons: Mapping the features of existing technologies to user needs. *Sensors* **2020**, *20*, 636.
4. Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.; Gould, S.; Van Den Hengel, A. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3674–3683.
5. Fried, D.; Hu, R.; Cirik, V.; Rohrbach, A.; Andreas, J.; Morency, L.P.; Berg-Kirkpatrick, T.; Saenko, K.; Klein, D.; Darrell, T. Speaker-follower models for vision-and-language navigation. *Advances in neural information processing systems* **2018**, *31*.
6. Tan, H.; Yu, L.; Bansal, M. Learning to navigate unseen environments: Back translation with environmental dropout. In Proceedings of the Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2610–2621.
7. Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; Gould, S. A recurrent vision-and-language bert for navigation. *arXiv preprint arXiv:2011.13922* **2020**.
8. Chen, S.; Guhur, P.L.; Schmid, C.; Laptev, I. History aware multimodal transformer for vision-and-language navigation. *Advances in neural information processing systems* **2021**, *34*, 5834–5847.
9. Zhou, G.; Hong, Y.; Wu, Q. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 7641–7649.

 doi:10.20944/preprints202606.2231.v1

10. Zhou, G.; Hong, Y.; Wang, Z.; Wang, X.E.; Wu, Q. Navgpt-2: Unleashing navigational reasoning capability for large vision-language models. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 260–278.
11. Qiao, Y.; Lyu, W.; Wang, H.; Wang, Z.; Li, Z.; Zhang, Y.; Tan, M.; Wu, Q. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 6710–6717.
12. Chen, J.; Lin, B.; Xu, R.; Chai, Z.; Liang, X.; Wong, K.Y. Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation. In Proceedings of the Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024, pp. 9796–9810.
13. Shi, X.; Li, Z.; Lyu, W.; Xia, J.; Dayoub, F.; Qiao, Y.; Wu, Q. Smartway: Enhanced waypoint prediction and backtracking for zero-shot vision-and-language navigation. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025, pp. 16923–16930.
14. Krantz, J.; Banerjee, S.; Zhu, W.; Corso, J.; Anderson, P.; Lee, S.; Thomason, J. Iterative vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14921–14930.
15. Hong, Y.; Wu, Q.; Qi, Y.; Rodriguez-Opazo, C.; Gould, S. Vln bert: A recurrent vision-and-language bert for navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 1643–1653.
16. Chen, S.; Guhur, P.L.; Tapaswi, M.; Schmid, C.; Laptev, I. Think global, act local: Dual-scale graph transformer for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 16537–16547.
17. Li, P.; Wu, K.; Xu, S.; Li, F.; Zhao, L.; Chen, L.; Yang, Z.X.; Zheng, N. Think before Go: Hierarchical Reasoning for Image-goal Navigation. *arXiv preprint arXiv:2604.17407* **2026**.
18. Qiao, Y.; Qi, Y.; Hong, Y.; Yu, Z.; Wang, P.; Wu, Q. Hop: History-and-order aware pre-training for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15418–15427.
19. Wang, Z.; Li, J.; Hong, Y.; Wang, Y.; Wu, Q.; Bansal, M.; Gould, S.; Tan, H.; Qiao, Y. Scaling data generation in vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 12009–12020.
20. Wang, L.; He, Z.; Dang, R.; Shen, M.; Liu, C.; Chen, Q. Vision-and-language navigation via causal learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 13139–13150.
21. Wu, W.; Chang, T.; Li, X.; Yin, Q.; Hu, Y. Vision-language navigation: a survey and taxonomy. *Neural Computing and Applications* **2024**, *36*, 3291–3316.
22. Zhang, Y.; Ma, Z.; Li, J.; Qiao, Y.; Wang, Z.; Chai, J.; Wu, Q.; Bansal, M.; Kordjamshidi, P. Vision-and-language navigation today and tomorrow: A survey in the era of foundation models. *arXiv preprint arXiv:2407.07035* **2024**.
23. Khan, J.; Aafaq, N.; Ali, Q.; Mohsin, M. A comprehensive review of recent advancements in vision-and-language navigation. *Discover Computing* **2026**, *29*, 167.
24. Pan, H.; Huang, S.; Yang, J.; Mi, J.; Li, K.; You, X.; Liang, P.; Yang, J.; Liu, Y.; Zhang, J.; et al. Robot Navigation via Foundation Language Models: A Review. *ACM Computing Surveys* **2026**, *58*, 1–38.
25. Nguyen, K.; Dey, D.; Brockett, C.; Dolan, B. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12527–12537.
26. Krantz, J.; Gokaslan, A.; Batra, D.; Lee, S.; Maksymets, O. Waypoint models for instruction-guided navigation in continuous environments. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15162–15171.
27. An, D.; Wang, H.; Wang, W.; Wang, Z.; Huang, Y.; He, K.; Wang, L. Etpnav: Evolving topological planning for vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**.
28. Jain, V.; Magalhaes, G.; Ku, A.; Vaswani, A.; Ie, E.; Baldridge, J. Stay on the path: Instruction fidelity in vision-and-language navigation. In Proceedings of the Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1862–1872.

 doi:10.20944/preprints202606.2231.v1

29. Chen, H.; Suhr, A.; Misra, D.; Snavely, N.; Artzi, Y. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12538–12547.
30. Mirowski, P.; Banki-Horvath, A.; Anderson, K.; Teplyashin, D.; Hermann, K.M.; Malinowski, M.; Grimes, M.K.; Simonyan, K.; Kavukcuoglu, K.; Zisserman, A.; et al. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292* **2019**.
31. Ku, A.; Anderson, P.; Patel, R.; Ie, E.; Baldridge, J. Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In Proceedings of the Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4392–4412.
32. Qi, Y.; Wu, Q.; Anderson, P.; Wang, X.; Wang, W.Y.; Shen, C.; Hengel, A.v.d. Reverie: Remote embodied visual referring expression in real indoor environments. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 9982–9991.
33. Batra, D.; Gokaslan, A.; Kembhavi, A.; Maksymets, O.; Mottaghi, R.; Savva, M.; Toshev, A.; Wijmans, E. Objectnav revisited: On evaluation of embodied agents navigating to objects. *arXiv preprint arXiv:2006.13171* **2020**.
34. Thomason, J.; Murray, M.; Cakmak, M.; Zettlemoyer, L. Vision-and-dialog navigation. In Proceedings of the Conference on Robot Learning. PMLR, 2020, pp. 394–406.
35. Padmakumar, A.; Thomason, J.; Shrivastava, A.; Lange, P.; Narayan-Chen, A.; Gella, S.; Piramuthu, R.; Tur, G.; Hakkani-Tur, D. Teach: Task-driven embodied agents that chat. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 2017–2025.
36. Gao, X.; Gao, Q.; Gong, R.; Lin, K.; Thattai, G.; Sukhatme, G.S. Dialfred: Dialogue-enabled agents for embodied instruction following. *IEEE Robotics and Automation Letters* **2022**, *7*, 10049–10056.
37. Song, X.; Chen, W.; Liu, Y.; Chen, W.; Li, G.; Lin, L. Towards long-horizon vision-language navigation: Platform, benchmark and method. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 12078–12088.
38. Zhang, J.; Ma, K. MG-VLN: Benchmarking Multi-Goal and Long-Horizon Vision-Language Navigation with Language Enhanced Memory Map. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 7750–7757.
39. Wang, T.; Li, X.; Lu, F.; Gong, T.; Dong, J.; Xue, W.; Qu, S.; Bai, C.; Chen, G. CoNavBench: Collaborative Long-Horizon Vision-Language Navigation Benchmark. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
40. Yue, L.; Zhou, D.; Xie, L.; Zhang, F.; Yan, Y.; Yin, E. Safe-vln: Collision avoidance for vision-and-language navigation of autonomous robots operating in continuous environments. *IEEE Robotics and Automation Letters* **2024**, *9*, 4918–4925.
41. Li, H.; Li, M.; Cheng, Z.Q.; Dong, Y.; Zhou, Y.; He, J.Y.; Dai, Q.; Mitamura, T.; Hauptmann, A.G. Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions. *Advances in Neural Information Processing Systems* **2024**, *37*, 119411–119442.
42. Li, Z.; Lv, Y.; Tu, Z.; Shang, D.; Qiao, H. Vision-language navigation with continual learning. *arXiv preprint arXiv:2409.02561* **2024**.
43. Jeong, S.; Kang, G.C.; Choi, S.; Kim, J.; Zhang, B.T. Continual vision-and-language navigation. *arXiv preprint arXiv:2403.15049* **2024**.
44. Su, H.T.; Wang, T.J.; Yeh, J.F.; Sun, M.; Hsu, W.H. Vln-nf: Feasibility-aware vision-and-language navigation with false-premise instructions. *arXiv preprint arXiv:2604.10533* **2026**.
45. Liu, S.; Zhang, H.; Qi, Y.; Wang, P.; Zhang, Y.; Wu, Q. Aerialvln: Vision-and-language navigation for uavs. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15384–15394.
46. Lee, J.; Miyanishi, T.; Kurita, S.; Sakamoto, K.; Azuma, D.; Matsuo, Y.; Inoue, N. CityNav: Language-Goal Aerial Navigation Dataset Using Geographic Information **2024**.
47. Gao, Y.; Li, C.; You, Z.; Liu, J.; Li, Z.; Chen, P.; Chen, Q.; Tang, Z.; Wang, L.; Yang, P.; et al. OpenFly: A comprehensive platform for aerial vision-language navigation. *arXiv preprint arXiv:2502.18041* **2025**.
48. Cai, H.; Rao, Y.; Huang, L.; Zhong, Z.; Dong, J.; Tan, J.; Lu, W.; Zhong, R. AirNav: A Large-Scale Real-World UAV Vision-and-Language Navigation Dataset with Natural and Diverse Instructions. *arXiv preprint arXiv:2601.03707* **2026**.

 doi:10.20944/preprints202606.2231.v1

49. Chi, T.C.; Shen, M.; Eric, M.; Kim, S.; Hakkani-Tur, D. Just ask: An interactive learning framework for vision and language navigation. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2020, Vol. 34, pp. 2459–2466.
50. Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; Fox, D. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10740–10749.
51. Banerjee, S.; Thomason, J.; Corso, J. The robotslang benchmark: Dialog-guided robot localization and navigation. In Proceedings of the Conference on Robot Learning. PMLR, 2021, pp. 1384–1393.
52. Mehta, H.; Artzi, Y.; Baldridge, J.; Ie, E.; Mirowski, P. Retouchdown: Releasing touchdown on StreetLearn as a public resource for language grounding tasks in street view. In Proceedings of the Proceedings of the third international workshop on spatial language understanding, 2020, pp. 56–62.
53. Zhu, F.; Liang, X.; Zhu, Y.; Yu, Q.; Chang, X.; Liang, X. Soon: Scenario oriented object navigation with graph-based exploration. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 12689–12699.
54. Vasudevan, A.B.; Dai, D.; Van Gool, L. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision* **2021**, *129*, 246–266.
55. Chen, S.; Guhur, P.L.; Tapaswi, M.; Schmid, C.; Laptev, I. Learning from unlabeled 3d environments for vision-and-language navigation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 638–655.
56. Taioli, F.; Rosa, S.; Castellini, A.; Natale, L.; Del Bue, A.; Farinelli, A.; Cristani, M.; Wang, Y. Mind the error! detection and localization of instruction errors in vision-and-language navigation. In Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2024, pp. 12993–13000.
57. Zheng, D.; Huang, S.; Zhao, L.; Zhong, Y.; Wang, L. Towards learning a generalist model for embodied navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13624–13634.
58. Li, J.; Padmakumar, A.; Sukhatme, G.; Bansal, M. Vln-video: Utilizing driving videos for outdoor vision-and-language navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 18517–18526.
59. O'Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 6892–6903.
60. Hong, H.; Qiao, Y.; Wang, S.; Liu, J.; Wu, Q. General scene adaptation for vision-and-language navigation. *arXiv preprint arXiv:2501.17403* **2025**.
61. Dong, Y.; Wu, F.; He, Q.; Cheng, Z.Q.; Li, H.; Li, M.; Cheng, Z.; Zhou, Y.; Sun, J.; Dai, Q.; et al. Ha-vln 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions. *arXiv preprint arXiv:2503.14229* **2025**.
62. Wang, L.; Xia, X.; Zhao, H.; Wang, H.; Wang, T.; Chen, Y.; Liu, C.; Chen, Q.; Pang, J. Rethinking the embodied gap in vision-and-language navigation: A holistic study of physical and visual disparities. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 9455–9465.
63. Zhu, S.; Mou, L.; Li, D.; Ye, B.; Huang, R.; Zhao, H. Vr-robo: A real-to-sim-to-real framework for visual robot navigation and locomotion. *IEEE Robotics and Automation Letters* **2025**.
64. Saxena, P.; Raghuvanshi, N.; Goveas, N. Uav-vln: End-to-end vision language guided navigation for uavs. In Proceedings of the 2025 European Conference on Mobile Robots (ECMR). IEEE, 2025, pp. 1–6.
65. Wei, M.; Wan, C.; Yu, X.; Wang, T.; Yang, Y.; Mao, X.; Zhu, C.; Cai, W.; Wang, H.; Chen, Y.; et al. Streamvln: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240* **2025**.
66. Lin, S.; Li, Z.; Zhao, X.; Zhou, G.; Wang, L.; Wei, R.; Tang, R.; Li, J.; Wang, H.; Pang, J.; et al. VLNVerse: A Benchmark for Vision-Language Navigation with Versatile, Embodied, Realistic Simulation and Evaluation. *arXiv preprint arXiv:2512.19021* **2025**.
67. Zhao, X.; Liu, C.; Ji, R.; Zhang, Z.; Zhu, M.; Song, L.; Ren, Z.; Qingliang, L.; Gao, Y.; Du, Z.; et al. CoT-VLNBench: A Benchmark for Visual Chain-of-Thought Reasoning in Vision-Language-Navigation Robots. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 36573–36581.

 doi:10.20944/preprints202606.2231.v1

68. Ma, C.Y.; Lu, J.; Wu, Z.; AlRegib, G.; Kira, Z.; Socher, R.; Xiong, C. Self-monitoring navigation agent via auxiliary progress estimation. *arXiv preprint arXiv:1901.03035* **2019**.
69. Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.F.; Wang, W.Y.; Zhang, L. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6629–6638.
70. Hao, W.; Li, C.; Li, X.; Carin, L.; Gao, J. Towards learning a generic agent for vision-and-language navigation via pre-training. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 13137–13146.
71. Qi, Y.; Pan, Z.; Hong, Y.; Yang, M.H.; Van Den Hengel, A.; Wu, Q. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1655–1664.
72. He, K.; Huang, Y.; Wu, Q.; Yang, J.; An, D.; Sima, S.; Wang, L. Landmark-rxr: Solving vision-and-language navigation with fine-grained alignment supervision. *Advances in Neural Information Processing Systems* **2021**, *34*, 652–663.
73. Wang, S.; Montgomery, C.; Orbay, J.; Birodkar, V.; Faust, A.; Gur, I.; Jaques, N.; Waters, A.; Baldridge, J.; Anderson, P. Less is more: Generating grounded navigation instructions from landmarks. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15428–15438.
74. Zhang, Y.; Kordjamshidi, P. Explicit object relation alignment for vision and language navigation. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, 2022, pp. 322–331.
75. Cui, Y.; Xie, L.; Zhang, Y.; Zhang, M.; Yan, Y.; Yin, E. Grounded entity-landmark adaptive pre-training for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 12043–12053.
76. Cui, Y.; Xie, L.; Zhao, Y.; Sun, J.; Yin, E. Generating Vision-Language Navigation Instructions Incorporated Fine-Grained Alignment Annotations. *Information Fusion* **2025**, p. 104107.
77. Lin, B.; Nie, Y.; Wei, Z.; Zhu, Y.; Xu, H.; Ma, S.; Liu, J.; Liang, X. Correctable landmark discovery via large models for vision-language navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, *46*, 8534–8548.
78. Liu, Q.; Zhang, S.; Qiao, Y.; Zhu, J.; Li, X.; Guo, L.; Wang, Q.; He, X.; Wu, Q.; Liu, J. GroundingMate: Aiding Object Grounding for Goal-Oriented Vision-and-Language Navigation. In Proceedings of the 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025, pp. 1775–1784.
79. Raychaudhuri, S.; Ta, D.; Ashton, K.; Chang, A.X.; Wang, J.; Bucher, B. Nl-slam for oc-vln: Natural language grounded slam for object-centric vln. *arXiv preprint arXiv:2411.07848* **2024**.
80. Zhao, G.; Li, G.; Chen, W.; Yu, Y. Over-nav: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16296–16306.
81. Wen, S.; Zhang, Z.; Sun, Y.; Wang, Z. Ovl-map: An online visual language map approach for vision-and-language navigation in continuous environments. *IEEE Robotics and Automation Letters* **2025**.
82. Li, D.; Yang, Z.; Qi, G.; Pang, S.; Shang, G.; Ma, Q.; Yang, Z. OpenMap: Instruction Grounding via Open-Vocabulary Visual-Language Mapping. In Proceedings of the Proceedings of the 33rd ACM International Conference on Multimedia, 2025, pp. 7444–7452.
83. Long, Y.; Cai, W.; Wang, H.; Zhan, G.; Dong, H. Instructnav: Zero-shot system for generic instruction navigation in unexplored environment. *arXiv preprint arXiv:2406.04882* **2024**.
84. Zhang, Y.; Yu, H.; Xiao, J.; Feroskhan, M. Grounded vision-language navigation for uavs with open-vocabulary goal understanding. *arXiv preprint arXiv:2506.10756* **2025**.
85. Chen, K.; Chen, J.K.; Chuang, J.; Vázquez, M.; Savarese, S. Topological planning with transformers for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11276–11286.
86. Li, H.; Dong, X.; Jiang, H.; Zhou, Y.; Ma, X. CMMR-VLN: Vision-and-Language Navigation via Continual Multimodal Memory Retrieval. *arXiv preprint arXiv:2603.07997* **2026**.
87. Liu, J.; Zhang, Z.; Li, X.; Wang, B.; Hu, Y.; Yin, B. TagaVLM: Topology-Aware Global Action Reasoning for Vision-Language Navigation. *arXiv preprint arXiv:2603.02972* **2026**.

 doi:10.20944/preprints202606.2231.v1

88. Georgakis, G.; Schmeckpeper, K.; Wanchoo, K.; Dan, S.; Miltsakaki, E.; Roth, D.; Daniilidis, K. Cross-modal map learning for vision and language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 15460–15470.
89. Chen, P.; Ji, D.; Lin, K.; Zeng, R.; Li, T.; Tan, M.; Gan, C. Weakly-supervised multi-granularity map learning for vision-and-language navigation. *Advances in Neural Information Processing Systems* **2022**, *35*, 38149–38161.
90. Liu, R.; Wang, X.; Wang, W.; Yang, Y. Bird's-eye-view scene graph for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 10968–10980.
91. Wang, Z.; Li, X.; Yang, J.; Liu, Y.; Jiang, S. Gridmm: Grid memory map for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International conference on computer vision, 2023, pp. 15625–15636.
92. Zhang, L.; Hao, X.; Xu, Q.; Zhang, Q.; Zhang, X.; Wang, P.; Zhang, J.; Wang, Z.; Zhang, S.; Xu, R. Mapnav: A novel memory representation via annotated semantic maps for vlm-based vision-and-language navigation. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 13032–13056.
93. Liu, R.; Wang, W.; Yang, Y. Volumetric environment representation for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 16317–16328.
94. Wang, Z.; Li, X.; Yang, J.; Liu, Y.; Hu, J.; Jiang, M.; Jiang, S. Lookahead exploration with neural radiance representation for continuous vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 13753–13762.
95. Dai, G.; Zhao, J.; Chen, Y.; Qin, Y.; Zhao, H.; Xie, G.; Yao, Y.; Shu, X.; Li, X. Unitedvln: Generalizable gaussian splatting for continuous vision-language navigation. *arXiv preprint arXiv:2411.16053* **2024**.
96. Gao, J.; Liu, R.; Wang, W. 3d gaussian map with open-set semantic grouping for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 9252–9262.
97. Miao, B.; Wei, R.; Ge, Z.; Gao, S.; Zhu, J.; Wang, R.; Tang, S.; Xiao, J.; Tang, R.; Li, J.; et al. Towards Physically Executable 3D Gaussian for Embodied Navigation. *arXiv preprint arXiv:2510.21307* **2025**.
98. Gao, J.; Liu, R.; Xu, Y.; Cao, T.; Zhang, Y.; Zhang, Z.; Peng, S.; Yang, Y.; Wang, W. Uncertainty-aware gaussian map for vision-language navigation. In Proceedings of the The Fourteenth International Conference on Learning Representations, 2026.
99. Zhang, J.; Wang, K.; Xu, R.; Zhou, G.; Hong, Y.; Fang, X.; Wu, Q.; Zhang, Z.; Wang, H. Navid: Video-based vlm plans the next step for vision-and-language navigation. *arXiv preprint arXiv:2402.15852* **2024**.
100. Zhang, J.; Wang, K.; Wang, S.; Li, M.; Liu, H.; Wei, S.; Wang, Z.; Zhang, Z.; Wang, H. Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. *arXiv preprint arXiv:2412.06224* **2024**.
101. Wang, S.; Wang, Y.; Fan, Z.; Wang, Y.; Chen, M.; Wang, K.; Su, Z.; Li, W.; Cai, X.; Jin, Y.; et al. Monodream: Monocular vision-language navigation with panoramic dreaming. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 10074–10082.
102. Zheng, D.; Huang, S.; Li, Y.; Wang, L. Efficient-VLN: A Training-Efficient Vision-Language Navigation Model. *arXiv preprint arXiv:2512.10310* **2025**.
103. Zhang, J.; Li, A.; Qi, Y.; Li, M.; Liu, J.; Wang, S.; Liu, H.; Zhou, G.; Wu, Y.; Li, X.; et al. Embodied navigation foundation model. *arXiv preprint arXiv:2509.12129* **2025**.
104. Zheng, Z.; Mao, Z.; Zhou, X.; Chen, J.; Li, M.; Sun, X.; Zou, H.; Zhang, Z.; Liu, X.; Cao, D.; et al. VLN-Cache: Enabling Token Caching for VLN Models with Visual/Semantic Dynamics Awareness. *arXiv preprint arXiv:2603.07080* **2026**.
105. Chen, C.; Jain, U.; Schissler, C.; Gari, S.V.A.; Al-Halah, Z.; Ithapu, V.K.; Robinson, P.; Grauman, K. Soundspaces: Audio-visual navigation in 3d environments. In Proceedings of the European conference on computer vision. Springer, 2020, pp. 17–36.
106. Paul, S.; Roy-Chowdhury, A.; Cherian, A. Avlen: Audio-visual-language embodied navigation in 3d environments. *Advances in Neural Information Processing Systems* **2022**, *35*, 6236–6249.
107. Liu, X.; Paul, S.; Chatterjee, M.; Cherian, A. Caven: An embodied conversational agent for efficient audio-visual navigation in noisy environments. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2024, Vol. 38, pp. 3765–3773.

 doi:10.20944/preprints202606.2231.v1

108. Yang, Z.; Liu, J.; Chen, P.; Cherian, A.; Marks, T.K.; Le Roux, J.; Gan, C. Rila: Reflective and imaginative language agent for zero-shot semantic audio-visual navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 16251–16261.
109. Zhu, Y.; Weng, Y.; Zhu, F.; Liang, X.; Ye, Q.; Lu, Y.; Jiao, J. Self-motivated communication agent for real-world vision-dialog navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1594–1603.
110. Han, L.; Min, H.; Hwangbo, G.; Choi, J.; Seo, P.H. DialNav: Multi-turn Dialog Navigation with a Remote Guide. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 8514–8523.
111. Zhou, S.; Wu, Y.; Wang, T.; Li, X.; Chen, G.; Liu, L.; Bai, C.; Li, X. DeCoNav: Dialog enhanced Long-Horizon Collaborative Vision-Language Navigation. *arXiv preprint arXiv:2604.12486* **2026**.
112. Yu, B.; Kasaei, H.; Cao, M. Co-navgpt: Multi-robot cooperative visual semantic navigation using large language models. *arXiv preprint arXiv:2310.07937* **2023**.
113. Wu, S.; Fu, X.; Wu, F.; Zha, Z.J. Cross-modal semantic alignment pre-training for vision-and-language navigation. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 4233–4241.
114. Du, M.; Wu, B.; Zhang, J.; Fan, Z.; Li, Z.; Luo, R.; Huang, X.J.; Wei, Z. Delan: Dual-level alignment for vision-and-language navigation by cross-modal contrastive learning. In Proceedings of the Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 4605–4616.
115. Hwang, M.; Jeong, J.; Kim, M.; Oh, Y.; Oh, S. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6683–6693.
116. Li, S.; Wang, Z.; Zhou, G.; Li, J.; Zeng, X.; Wang, L.; Qiao, Y.; Wu, Q.; Bansal, M.; Wang, Y. Learning Goal-Oriented Language-Guided Navigation with Self-Improving Demonstrations at Scale. *arXiv preprint arXiv:2509.24910* **2025**.
117. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
118. Zhou, X.; Girdhar, R.; Joulin, A.; Krähenbühl, P.; Misra, I. Detecting twenty-thousand classes using image-level supervision. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 350–368.
119. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 4015–4026.
120. Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In Proceedings of the European conference on computer vision. Springer, 2024, pp. 38–55.
121. Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Visual language maps for robot navigation. In Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023, pp. 10608–10615.
122. Jatavallabhula, K.M.; Kuwajerwala, A.; Gu, Q.; Omama, M.; Chen, T.; Maalouf, A.; Li, S.; Iyer, G.; Saryazdi, S.; Keetha, N.; et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241* **2023**.
123. Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; Funkhouser, T.; et al. Openscene: 3d scene understanding with open vocabularies. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 815–824.
124. Lu, S.; Chang, H.; Jing, E.P.; Boularias, A.; Bekris, K. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In Proceedings of the Conference on Robot Learning. PMLR, 2023, pp. 1610–1620.
125. Werby, A.; Huang, C.; Büchner, M.; Valada, A.; Burgard, W. Hierarchical open-vocabulary 3d scene graphs for language-grounded robot navigation. In Proceedings of the First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024, 2024.
126. Wang, H.; Wang, W.; Liang, W.; Xiong, C.; Shen, J. Structured scene memory for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2021, pp. 8455–8464.
127. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.

 doi:10.20944/preprints202606.2231.v1

128. He, K.; Jing, Y.; Huang, Y.; Lu, Z.; An, D.; Wang, L. Memory-adaptive vision-and-language navigation. *Pattern Recognition* **2024**, *153*, 110511.
129. Zhang, S.; Qiao, Y.; Wang, Q.; Yan, Z.; Wu, Q.; Wei, Z.; Liu, J. Cosmo: Combination of selective memorization for low-cost vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 5511–5522.
130. An, D.; Qi, Y.; Li, Y.; Huang, Y.; Wang, L.; Tan, T.; Shao, J. Bevbert: Multimodal map pre-training for language-guided navigation. *arXiv preprint arXiv:2212.04385* **2022**.
131. Zhang, X.; Xu, Y.; Li, J.; Liu, R.; Hu, Z. Agent journey beyond rgb: Hierarchical semantic-spatial representation enrichment for vision-and-language navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 18791–18799.
132. Wang, Z.; Li, M.; Wu, M.; Moens, M.F.; Tuytelaars, T. Instruction-guided path planning with 3D semantic maps for vision-language navigation. *Neurocomputing* **2025**, *625*, 129457.
133. Zeng, S.; Qi, D.; Chang, X.; Xiong, F.; Xie, S.; Wu, X.; Liang, S.; Xu, M.; Wei, X. Janusvln: Decoupling semantics and spatiality with dual implicit memory for vision-language navigation. *arXiv preprint arXiv:2509.22548* **2025**.
134. Qi, Z.; Zhang, Z.; Yu, Y.; Wang, J.; Zhao, H. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221* **2025**.
135. Lu, Y.; Sun, S.; Liu, N.; Jiang, B.; Zhang, Y.; Chen, J.; Du, C. STEP-Nav: Spatial-Temporal Efficient Visual Token Pruning for Vision-and-Language Navigation with Large Language Models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 24097–24105.
136. Huang, C.; Mees, O.; Zeng, A.; Burgard, W. Audio visual language maps for robot navigation. In Proceedings of the International Symposium on Experimental Robotics. Springer, 2023, pp. 105–117.
137. Shi, Z.; Zhang, L.; Li, L.; Shen, Y. Towards audio-visual navigation in noisy environments: a large-scale benchmark dataset and an architecture considering multiple sound-sources. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 14673–14680.
138. Fan, J.; Chen, P.; Li, C.; Du, Q.; Chen, J.; Tan, M. NaVLA^2: A Vision-Language-Audio-Action Model for Multimodal Instruction Navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 18234–18242.
139. Fan, Y.; Chen, W.; Jiang, T.; Zhou, C.; Zhang, Y.; Wang, X. Aerial vision-and-dialog navigation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, 2023, pp. 3043–3061.
140. Su, Y.; An, D.; Chen, K.; Yu, W.; Ning, B.; Ling, Y.; Huang, Y.; Wang, L. Learning fine-grained alignment for aerial vision-dialog navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 7060–7068.
141. Zhu, W.; Hu, H.; Chen, J.; Deng, Z.; Jain, V.; Ie, E.; Sha, F. Babywalk: Going farther in vision-and-language navigation by taking baby steps. In Proceedings of the Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 2539–2556.
142. Hong, Y.; Rodriguez, C.; Wu, Q.; Gould, S. Sub-instruction aware vision-and-language navigation. In Proceedings of the Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), 2020, pp. 3360–3376.
143. Zhang, Y.; Kordjamshidi, P. Vln-trans: Translator for the vision and language navigation agent. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 13219–13233.
144. Wang, X.; Wang, W.; Shao, J.; Yang, Y. Lana: A language-capable navigator for instruction following and generation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 19048–19058.
145. He, Z.; Wang, L.; Li, S.; Yan, Q.; Liu, C.; Chen, Q. A multilevel attention network with sub-instructions for continuous vision-and-language navigation: Z. He et al. *Applied Intelligence* **2025**, *55*, 657.
146. Huang, B.; Zheng, Y.; Lan, C.; Sui, D.; Zhao, X.; Zhang, X.; Xiao, M.; Yu, D. Action-Aware Visual-Textual Alignment for Long-Instruction Vision-and-Language Navigation. *ACM Transactions on Multimedia Computing, Communications and Applications* **2025**, *21*, 1–22.
147. Wang, S.; Wang, Y.; Lian, G.; Wang, Y.; Chen, M.; Wang, K.; Zhang, B.; Su, Z.; Zhou, Y.; Li, W.; et al. ProgressThink: Semantic Progress Reasoning for Vision-Language Navigation. *arXiv preprint arXiv:2511.17097* **2025**.

 doi:10.20944/preprints202606.2231.v1

148. Chen, K.; An, D.; Huang, Y.; Xu, R.; Su, Y.; Ling, Y.; Reid, I.; Wang, L. Constraint-aware zero-shot vision-language navigation in continuous environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
149. Yin, H.; Wei, H.; Xu, X.; Guo, W.; Zhou, J.; Lu, J. GC-VLN: Instruction as graph constraints for training-free vision-and-language navigation. *arXiv preprint arXiv:2509.10454* **2025**.
150. Gao, Y.; Wang, Z.; Han, P.; Jing, L.; Wang, D.; Zhao, B. Exploring spatial representation to enhance LLM reasoning in aerial vision-language navigation. *arXiv preprint arXiv:2410.08500* **2024**.
151. Zhou, L.; Xue, R.; Luo, X. Structured Instruction Parsing and Scene Alignment For UAV Vision-Language Navigation. In Proceedings of the 2025 IEEE International Conference on Image Processing (ICIP). IEEE, 2025, pp. 2600–2605.
152. Zhang, W.; Gao, C.; Yu, S.; Peng, R.; Zhao, B.; Zhang, Q.; Cui, J.; Chen, X.; Li, Y. Citynavagent: Aerial vision-and-language navigation with hierarchical semantic planning and global memory. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 31292–31309.
153. Ma, T.; Zhang, Y.; Wang, Z.; Kordjamshidi, P. Breaking down and building up: Mixture of skill-based vision-and-language navigation agents. *arXiv preprint arXiv:2508.07642* **2025**.
154. Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; Xia, F. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14455–14465.
155. Bai, Q.; Chen, Z.; Luo, L.; Du, H.; Lei, Y.; Jiao, Z. Endowing embodied agents with spatial reasoning capabilities for vision-and-language navigation. *arXiv preprint arXiv:2504.08806* **2025**.
156. Du, Y.; Fu, T.; Chen, Z.; Li, B.; Su, S.; Zhao, Z.; Wang, C. Vl-nav: real-time vision-language navigation with spatial reasoning. *arXiv preprint arXiv:2502.00931* **2025**.
157. Liu, F.; Li, G.; Zou, L.; Chen, Y.; Cheng, P. DroneNav: Unified text-visual representation and structured spatial reasoning for robust UAV vision-and-language navigation. *Neurocomputing* **2026**, p. 133492.
158. Yue, L.; Fan, Y.; Lian, S.; Zhao, Y.; Yu, J.; Xie, L.; Zhang, F. Spatial-VLN: Zero-Shot Vision-and-Language Navigation With Explicit Spatial Perception and Exploration. *arXiv preprint arXiv:2601.12766* **2026**.
159. Qiao, Y.; Lyu, W.; Wang, H.; Wang, Z.; Li, Z.; Zhang, Y.; Tan, M.; Wu, Q. Open-nav: Exploring zero-shot vision-and-language navigation in continuous environment with open-source llms. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 6710–6717.
160. Jiang, Z.; Wang, X. SpatialGPT: Zero-Shot Vision-and-Language Navigation via Spatial CoT over Structured Spatial Memory. In Proceedings of the Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems, 2025, pp. 423–435.
161. Liu, C.; Zhou, Z.; Zhang, J.; Zhang, M.; Huang, S.; Duan, H. Msnav: Zero-shot vision-and-language navigation with dynamic memory and llm spatial reasoning. In Proceedings of the ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2026, pp. 20112–20116.
162. Zhou, X.; Xiao, T.; Liu, L.; Wang, Y.; Chen, M.; Meng, X.; Wang, X.; Feng, W.; Sui, W.; Su, Z. FSR-VLN: Fast and Slow Reasoning for Vision-Language Navigation with Hierarchical Multi-modal Scene Graph. *arXiv preprint arXiv:2509.13733* **2025**.
163. Zhang, J.; Li, Z.; Wang, S.; Shi, X.; Wei, Z.; Wu, Q. SpatialNav: Leveraging Spatial Scene Graphs for Zero-Shot Vision-and-Language Navigation. *arXiv preprint arXiv:2601.06806* **2026**.
164. Li, X.; Wang, Z.; Yang, J.; Wang, Y.; Jiang, S. Kerm: Knowledge enhanced reasoning for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2583–2592.
165. Zhou, G.; Hong, Y.; Wu, Q. Navgpt: Explicit reasoning in vision-and-language navigation with large language models. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2024, Vol. 38, pp. 7641–7649.
166. Lyu, K.; Wu, K.; Li, P.; Hu, X.; Si, Q.; Miao, C.; Yang, N.; Wang, Z.; Xiao, L.; Hu, L.; et al. HiMemVLN: Enhancing Reliability of Open-Source Zero-Shot Vision-and-Language Navigation with Hierarchical Memory System. *arXiv preprint arXiv:2603.14807* **2026**.
167. Lin, B.; Nie, Y.; Wei, Z.; Chen, J.; Ma, S.; Han, J.; Xu, H.; Chang, X.; Liang, X. Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2025**.
168. Wang, S.; Wang, Y.; Li, W.; Cai, X.; Wang, Y.; Chen, M.; Wang, K.; Su, Z.; Li, D.; Fan, Z. Aux-think: Exploring reasoning strategies for data-efficient vision-language navigation. *arXiv preprint arXiv:2505.11886* **2025**.

 doi:10.20944/preprints202606.2231.v1

169. Zuo, J.; Mu, L.; Jiang, F.; Ma, C.; Xu, M.; Qi, Y. FantasyVLN: Unified Multimodal Chain-of-Thought Reasoning for Vision-Language Navigation. *arXiv preprint arXiv:2601.13976* **2026**.
170. Ding, X.; Wei, J.; Yang, Y.; Jiang, S.; Zhang, Q.; Wu, H.; Jia, F.; Mi, L.; Yan, Y.; Wang, W.; et al. AdaNav: Adaptive Reasoning with Uncertainty for Vision-Language Navigation. *arXiv preprint arXiv:2509.24387* **2025**.
171. Xue, W.; Li, M.; Wu, X.; Tang, J.; Yang, D.; Zhang, L. ProFocus: Proactive Perception and Focused Reasoning in Vision-and-Language Navigation. *arXiv preprint arXiv:2603.05530* **2026**.
172. Li, X.; Lyu, F.; Wu, H.; Liu, M.; Liu, J.N.; Liu, G. Stop Wandering: Efficient Vision-Language Navigation via Metacognitive Reasoning. *arXiv preprint arXiv:2604.02318* **2026**.
173. Xin, Z.; Li, W.; Jiang, Y.; Wang, B.; Cong, R.; Qin, J.; Huang, S. DecoVLN: Decoupling Observation, Reasoning, and Correction for Vision-and-Language Navigation. *arXiv preprint arXiv:2603.13133* **2026**.
174. Fang, X.; Fang, W.; Wang, C. Hierarchical semantic-augmented navigation: Optimal transport and graph-driven reasoning for vision-language navigation. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
175. Lin, B.; Nie, Y.; Zai, K.L.; Wei, Z.; Han, M.; Xu, R.; Niu, M.; Han, J.; Zhang, H.; Lin, L.; et al. EvolveNav: Empowering LLM-Based Vision-Language Navigation via Self-Improving Embodied Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2026**.
176. Wang, H.; Liang, W.; Van Gool, L.; Wang, W. Dreamwalker: Mental planning for continuous vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 10873–10883.
177. Bar, A.; Zhou, G.; Tran, D.; Darrell, T.; LeCun, Y. Navigation world models. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 15791–15801.
178. Pan, Y.; Xu, Y.; Liu, Z.; Wang, H. Planning from imagination: Episodic simulation and episodic memory for vision-and-language navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 6345–6353.
179. Perincherry, A.; Krantz, J.; Lee, S. Do visual imaginations improve vision-and-language navigation agents? In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 3846–3855.
180. Huang, Y.; Wu, M.; Li, R.; Tu, Z. Vista: Generative visual imagination for vision-and-language navigation. *arXiv preprint arXiv:2505.07868* **2025**.
181. Lian, G.; Wang, S.; Wang, Y.; Wang, Y.; Chen, M.; Wang, K.; Zhang, B.; Su, Z.; Li, D.; Fan, Z. MapDream: Task-Driven Map Learning for Vision-Language Navigation. *arXiv preprint arXiv:2602.00222* **2026**.
182. Dai, G.; Wang, S.; Zhao, H.; Zhu, B.; Sun, Q.; Shu, X. ThinkMatter: Panoramic-Aware Instructional Semantics for Monocular Vision-and-Language Navigation. *IEEE Transactions on Image Processing* **2026**.
183. Liu, F.; Xie, S.; Luo, M.; Chu, Z.; Hu, J.; Wu, X.; Xu, M. NavForesee: A Unified Vision-Language World Model for Hierarchical Planning and Dual-Horizon Navigation Prediction. *arXiv preprint arXiv:2512.01550* **2025**.
184. Hu, J.; Chen, J.; Bai, H.; Luo, M.; Xie, S.; Chen, Z.; Liu, F.; Chu, Z.; Xue, X.; Ren, B.; et al. AstraNav-World: World Model for Foresight Control and Consistency. *arXiv preprint arXiv:2512.21714* **2025**.
185. Huang, C.; Tang, L.; Zhan, Z.; Yu, L.; Zeng, R.; Liu, Z.; Wang, Z.; Li, J. UNeMo: Collaborative Visual-Language Reasoning and Navigation via a Multimodal World Model. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 18315–18323.
186. Fan, Z.; Lyu, W.; Song, W.; Zhao, L.; Yang, Y.; Wang, X.; He, J.; Huang, L.; Liu, H.; Sun, B.; et al. PROSPECT: Unified Streaming Vision-Language Navigation via Semantic–Spatial Fusion and Latent Predictive Representation. *arXiv preprint arXiv:2603.03739* **2026**.
187. Chen, H.; Jiang, S.; Su, T.; Gao, C.; Chen, X.; Li, Y.; Chen, Z. WorldMAP: Bootstrapping Vision-Language Navigation Trajectory Prediction with Generative World Models. *arXiv preprint arXiv:2604.07957* **2026**.
188. Wu, K.; Li, P.; Lyu, K.; Zhao, L.; He, Q.; Wang, J.; Liu, J. Dual-Anchoring: Addressing State Drift in Vision-Language Navigation. *arXiv preprint arXiv:2604.17473* **2026**.
189. Liu, R.; Wu, S.; Lin, D.; Zhang, W. CVLN-Think: Causal Inference with Counterfactual Style Adaptation for Continuous Vision-and-Language Navigation. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2025, pp. 15299–15305.
190. Ye, S.; Ge, Y.; Zheng, K.; Gao, S.; Yu, S.; Kurian, G.; Indupuru, S.; Tan, Y.L.; Zhu, C.; Xiang, J.; et al. World Action Models are Zero-shot Policies, 2026, [arXiv:cs.RO/2602.15922].
191. Yuan, T.; Dong, Z.; Liu, Y.; Zhao, H. Fast-WAM: Do World Action Models Need Test-time Future Imagination? *arXiv preprint arXiv:2603.16666* **2026**.

 doi:10.20944/preprints202606.2231.v1

192. Ye, A.; Wang, B.; Ni, C.; Huang, G.; Zhao, G.; Li, H.; Li, H.; Li, J.; Lv, J.; Liu, J.; et al. GigaWorld-Policy: An Efficient Action-Centered World–Action Model. *arXiv preprint arXiv:2603.17240* **2026**.
193. Wang, L.; Zheng, Y.; Chen, Q.; Li, S.; Zhang, Y.; Xing, Z.; Zhang, Q.; Li, X.; Qian, D.; Yang, P.; et al. Latent-WAM: Latent World Action Modeling for End-to-End Autonomous Driving. *arXiv preprint arXiv:2603.24581* **2026**.
194. Zhou, Y.; Wang, X.; Shao, H.; Wang, L.; Zhao, G.; Shao, J.; Zhu, J.; Yu, T.; Zhu, Z.; Huang, G.; et al. DriveDreamer-Policy: A Geometry-Grounded World-Action Model for Unified Generation and Planning. *arXiv preprint arXiv:2604.01765* **2026**.
195. Yang, H.; Long, Y.; Yu, Z.; Yang, Z.; Wang, M.; Xu, J.; Wang, Y.; Yu, Z.; Cai, W.; Kang, L.; et al. NavSpace: How Navigation Agents Follow Spatial Intelligence Instructions. *arXiv preprint arXiv:2510.08173* **2025**.
196. Zhao, X.; Liu, C.; Ji, R.; Zhang, Z.; Zhu, M.; Song, L.; Ren, Z.; Qingliang, L.; Gao, Y.; Du, Z.; et al. CoT-VLNBench: A Benchmark for Visual Chain-of-Thought Reasoning in Vision-Language-Navigation Robots. *Proceedings of the AAAI Conference on Artificial Intelligence* **2026**, *40*, 36573–36581. https://doi.org/10.1609/aaai.v40i43.40980.
197. Guo, W.; Xu, X.; Liu, Y.; Li, X.; Yin, H.; Chen, H.; Zheng, W.; Feng, J.; Zhou, J.; Lu, J. AwareVLN: Reasoning with Self-awareness for Vision-Language Navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 4065–4075.
198. Gao, P.; Wang, P.; Wang, F.; Fujita, H.; Aljuaid, H.; Shang, J.L. DeepVLN: Vision-and-Language Navigation via Deep Reasoning and Collaborative Mechanisms Based on Large Language Models. *IEEE Journal of Selected Topics in Signal Processing* **2026**, *20*, 47–62. https://doi.org/10.1109/JSTSP.2026.3652407.
199. Zhang, Z.; Li, Z.; Rahmati, B.; Yang, R.H.; Ma, Y.; Rasouli, A.; Pakdamansavoji, S.; Wu, Y.; Zhang, L.; Cao, T.; et al. Do World Action Models Generalize Better than VLAs? A Robustness Study, 2026, [arXiv:cs.RO/2603.22078].
200. Li, X.; Li, C.; Xia, Q.; Bisk, Y.; Celikyilmaz, A.; Gao, J.; Smith, N.A.; Choi, Y. Robust navigation with language pretraining and stochastic sampling. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 1494–1499.
201. Ma, C.Y.; Wu, Z.; AlRegib, G.; Xiong, C.; Kira, Z. The regretful agent: Heuristic-aided navigation through progress estimation. In Proceedings of the Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2019, pp. 6732–6740.
202. Ke, L.; Li, X.; Bisk, Y.; Holtzman, A.; Gan, Z.; Liu, J.; Gao, J.; Choi, Y.; Srinivasa, S. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6741–6749.
203. Zhu, F.; Zhu, Y.; Chang, X.; Liang, X. Vision-language navigation with self-supervised auxiliary reasoning tasks. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10012–10022.
204. Liu, R.; Wang, W.; Yang, Y. Vision-language navigation with energy-based policy. *Advances in Neural Information Processing Systems* **2024**, *37*, 108208–108230.
205. Cheng, A.C.; Ji, Y.; Yang, Z.; Gongye, Z.; Zou, X.; Kautz, J.; Bıyık, E.; Yin, H.; Liu, S.; Wang, X. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453* **2024**.
206. Zhu, W.; Zhang, Z.; Wang, X.; Pan, H.; Wang, T.; Geng, T.; Xu, R.; Zheng, F. NaVIDA: Vision-Language Navigation with Inverse Dynamics Augmentation. *arXiv preprint arXiv:2601.18188* **2026**.
207. Li, P.; Wu, K.; Xu, S.; Li, F.; Li, H.; Zhao, L.; Lyu, K.; Chen, L.; Yang, Z.X.; Zheng, N. SpaAct: Spatially-Activated Transition Learning with Curriculum Adaptation for Vision-Language Navigation. *arXiv preprint arXiv:2604.27620* **2026**.
208. Hao, H.; Chen, L.; Han, M.; Li, C.; An, D.; Yang, Y.; Li, Z.; Chang, X. LatentPilot: Scene-Aware Vision-and-Language Navigation by Dreaming Ahead with Latent Visual Reasoning. *arXiv preprint arXiv:2603.29165* **2026**.
209. Castro, M.G.; Rajagopal, S.; Gorbatov, D.; Schmittle, M.; Baijal, R.; Zhang, O.; Scalise, R.; Talia, S.; Romig, E.; de Melo, C.; et al. VAMOS: A Hierarchical Vision-Language-Action Model for Capability-Modulated and Steerable Navigation. *arXiv preprint arXiv:2510.20818* **2025**.
210. Cai, W.; Peng, J.; Yang, Y.; Zhang, Y.; Wei, M.; Wang, H.; Chen, Y.; Wang, T.; Pang, J. Navdp: Learning sim-to-real navigation diffusion policy with privileged information guidance. *arXiv preprint arXiv:2505.08712* **2025**.

 doi:10.20944/preprints202606.2231.v1

211. Sun, X.; Si, W.; Ni, W.; Li, Y.; Wu, D.; Xie, F.; Guan, R.; Xu, H.Y.; Ding, H.; Wu, Y.; et al. AutoFly: Vision-Language-Action Model for UAV Autonomous Navigation in the Wild. *arXiv preprint arXiv:2602.09657* **2026**.
212. Xu, P.; Deng, Z.; Deng, J.; Gu, Z.; Wan, S. AerialVLA: A Vision-Language-Action Model for UAV Navigation via Minimalist End-to-End Control. *arXiv preprint arXiv:2603.14363* **2026**.
213. Wang, X.; Xiong, W.; Wang, H.; Wang, W.Y. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In Proceedings of the Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 37–53.
214. Wang, X.; Huang, Q.; Celikyilmaz, A.; Gao, J.; Shen, D.; Wang, Y.F.; Wang, W.Y.; Zhang, L. Vision-language navigation policy learning and adaptation. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 4205–4216.
215. Chen, J.; Gao, C.; Meng, E.; Zhang, Q.; Liu, S. Reinforced structured state-evolution for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15450–15459.
216. Wang, J.; Wang, T.; Xu, L.; He, Z.; Sun, C. Discovering intrinsic subgoals for vision-and-language navigation via hierarchical reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**, *36*, 6516–6528.
217. Wang, J.; Wang, T.; Cai, W.; Xu, L.; Sun, C. Boosting efficient reinforcement learning for vision-and-language navigation with open-sourced llm. *IEEE Robotics and Automation Letters* **2024**, *10*, 612–619.
218. Liu, R.; Kong, P.; Wu, S.; Zhang, W. RewardVLN: AnImproved Agent Navigation Based On Visual-Instruction Alignment. In Proceedings of the 2024 International Conference on Advanced Robotics and Mechatronics (ICARM). IEEE, 2024, pp. 126–133.
219. Wang, Y.; Sun, Z.; Zhang, J.; Xian, Z.; Biyik, E.; Held, D.; Erickson, Z. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. *arXiv preprint arXiv:2402.03681* **2024**.
220. Zhang, Z.; Zhu, W.; Pan, H.; Wang, X.; Xu, R.; Sun, X.; Zheng, F. Activevln: Towards active exploration via multi-turn rl in vision-and-language navigation. *arXiv preprint arXiv:2509.12618* **2025**.
221. Ye, S.; Mao, S.; Cui, Y.; Yu, X.; Zhai, S.; Chen, W.; Zhou, S.; Xiong, R.; Wang, Y. ETP-R1: Evolving Topological Planning with Reinforcement Fine-tuning for Vision-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2512.20940* **2025**.
222. Li, J.; Wan, C.; Dong, S.; Ding, C.; Wang, Q.; Ma, Z.; Gong, Y. Trajectory-Diversity-Driven Robust Vision-and-Language Navigation. *arXiv preprint arXiv:2603.15370* **2026**.
223. Wang, Z.; Lin, Z.; Yang, Y.; Fu, H.; Ye, D. SeeNav-Agent: Enhancing Vision-Language Navigation with Visual Prompt and Step-Level Policy Optimization. *arXiv preprint arXiv:2512.02631* **2025**.
224. Huang, T.; Li, D.; Yang, R.; Zhang, Z.; Yang, Z.; Tang, H. Mobilevla-r1: Reinforcing vision-language-action for mobile robots. *arXiv preprint arXiv:2511.17889* **2025**.
225. Li, H.; Liu, R.; Fan, H.; Yang, Y. Let's Reward Step-by-Step: Step-Aware Contrastive Alignment for Vision-Language Navigation in Continuous Environments. *arXiv preprint arXiv:2603.09740* **2026**.
226. Liu, Q.; Huang, T.; Zhang, Z.; Tang, H. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884* **2025**.
227. Wang, S.; Luo, Y.; Chen, X.; Luo, A.; Li, D.; Liu, C.; Chen, S.; Zhang, Y.; Yu, J. VLingNav: Embodied Navigation with Adaptive Reasoning and Visual-Assisted Linguistic Memory. *arXiv preprint arXiv:2601.08665* **2026**.
228. Ross, S.; Bagnell, D. Efficient reductions for imitation learning. In Proceedings of the Proceedings of the thirteenth international conference on artificial intelligence and statistics. JMLR Workshop and Conference Proceedings, 2010, pp. 661–668.
229. Shi, H.; Deng, X.; Li, Z.; Chen, G.; Wang, Y.; Nie, L. DAgger Diffusion Navigation: DAgger Boosted Diffusion Policy for Vision-Language Navigation. *arXiv preprint arXiv:2508.09444* **2025**.
230. He, G.; Liu, Z.; Xu, K.; Xu, L.; Qiao, T.; Yu, W.; Wu, C.; Xie, W. Nipping the Drift in the Bud: Retrospective Rectification for Robust Vision-Language Navigation. *arXiv preprint arXiv:2602.06356* **2026**.
231. Liang, X.; Ma, L.; Guo, S.; Han, J.; Xu, H.; Ma, S.; Liang, X. Cornav: Autonomous agent with self-corrected planning for zero-shot vision-and-language navigation. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 12538–12559.
232. Long, Y.; Li, X.; Cai, W.; Dong, H. Discuss before moving: Visual language navigation via multi-expert discussions. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2024, pp. 17380–17387.

 doi:10.20944/preprints202606.2231.v1

233. Xin, Z.; Li, W.; Jiang, Y.; Huang, Z.; Wang, B.; Li, P.; Zhu, J.; Qin, J.; Huang, S. AgentVLN: Towards Agentic Vision-and-Language Navigation. *arXiv preprint arXiv:2603.17670* **2026**.
234. Yu, Z.; Long, Y.; Yang, Z.; Zeng, C.; Fan, H.; Zhang, J.; Dong, H. Correctnav: Self-correction flywheel empowers vision-language-action navigation model. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 18737–18745.
235. Dong, X.; Zhao, H.; Gao, J.; Li, H.; Ma, X.; Zhou, Y.; Chen, F.; Liu, J. Se-vln: A self-evolving vision-language navigation framework based on multimodal large language models. *arXiv preprint arXiv:2507.13152* **2025**.
236. Zhong, Y.; Zhang, Z.; Zhang, R.; Huang, L.; Gao, H.; Wang, S.; Li, D.; Han, R.; Guo, J.; Peng, S.; et al. Run, Ruminate, and Regulate: A Dual-process Thinking System for Vision-and-Language Navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2026, Vol. 40, pp. 18845–18854.
237. Huang, J.; Huang, J.; Yang, H.; Li, H.; Wang, Y. AERR-Nav: Adaptive Exploration-Recovery-Reminiscing Strategy for Zero-Shot Object Navigation. *arXiv preprint arXiv:2603.17712* **2026**.
238. Li, J.; Tan, H.; Bansal, M. Envedit: Environment editing for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15407–15417.
239. He, K.; Si, C.; Lu, Z.; Huang, Y.; Wang, L.; Wang, X. Frequency-enhanced data augmentation for vision-and-language navigation. *Advances in neural information processing systems* **2023**, *36*, 4351–4364.
240. Li, J.; Bansal, M. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. *Advances in neural information processing systems* **2023**, *36*, 21878–21894.
241. Wang, S.; Zhou, D.; Xie, L.; Xu, C.; Yan, Y.; Yin, E. Panogen++: Domain-adapted text-guided panoramic environment generation for vision-and-language navigation. *Neural Networks* **2025**, *187*, 107320.
242. Zhong, Y.; Zhang, R.; Zhang, Z.; Wang, S.; Fang, C.; Zhang, X.; Guo, J.; Peng, S.; Huang, D.; Yan, Y.; et al. World-Consistent Data Generation for Vision-and-Language Navigation. *arXiv preprint arXiv:2412.06413* **2024**.
243. Kamath, A.; Anderson, P.; Wang, S.; Koh, J.Y.; Ku, A.; Waters, A.; Yang, Y.; Baldridge, J.; Parekh, Z. A new path: Scaling vision-and-language navigation with synthetic instructions and imitation learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 10813–10823.
244. Wang, Z.; Li, J.; Hong, Y.; Li, S.; Li, K.; Yu, S.; Wang, Y.; Qiao, Y.; Wang, Y.; Bansal, M.; et al. Bootstrapping language-guided navigation learning with self-refining data flywheel. *arXiv preprint arXiv:2412.08467* **2024**.
245. Wang, Z.; Zhu, Y.; Lee, G.H.; Fan, Y. Navrag: Generating user demand instructions for embodied navigation through retrieval-augmented llm. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 8430–8440.
246. Zheng, Y.; Zhang, L.; Sun, Y.; Shen, Y.; Zhao, S. CaneSpeaker: An LLM-Assisted Speaker for Generating Human-Like Navigation Instructions. *ACM Transactions on Multimedia Computing, Communications and Applications* **2026**, *22*, 1–26.
247. Han, M.; Ma, L.; Zhumakhanova, K.; Radionova, E.; Zhang, J.; Chang, X.; Liang, X.; Laptev, I. Roomtour3d: Geometry-aware video-instruction tuning for embodied navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 27586–27596.
248. Wei, M.; Wan, C.; Peng, J.; Yu, X.; Yang, Y.; Feng, D.; Cai, W.; Zhu, C.; Wang, T.; Pang, J.; et al. Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. *arXiv preprint arXiv:2512.08186* **2025**.
249. Zhang, W.; Ma, C.; Wu, Q.; Yang, X. Language-guided navigation via cross-modal grounding and alternate adversarial learning. *IEEE Transactions on Circuits and Systems for Video Technology* **2020**, *31*, 3469–3481.
250. Wang, X.E.; Jain, V.; Ie, E.; Wang, W.Y.; Kozareva, Z.; Ravi, S. Environment-agnostic multitask learning for natural language grounded navigation. In Proceedings of the European conference on computer vision. Springer, 2020, pp. 413–430.
251. Liang, X.; Zhu, F.; Zhu, Y.; Lin, B.; Wang, B.; Liang, X. Contrastive instruction-trajectory learning for vision-language navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 1592–1600.
252. Guhur, P.L.; Tapaswi, M.; Chen, S.; Laptev, I.; Schmid, C. Airbert: In-domain pretraining for vision-and-language navigation. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 1634–1643.

 doi:10.20944/preprints202606.2231.v1

253. Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525* **2024**.
254. Li, A.; Wang, Z.; Zhang, J.; Li, M.; Qi, Y.; Chen, Z.; Zhang, Z.; Wang, H. Urbanvla: A vision-language-action model for urban micromobility. *arXiv preprint arXiv:2510.23576* **2025**.
255. Huang, Z.; Zhang, Y.; Liu, J.; Song, R.; Tang, C.; Ma, J. TIC-VLA: A Think-in-Control Vision-Language-Action Model for Robot Navigation in Dynamic Environments. *arXiv preprint arXiv:2602.02459* **2026**.
256. Yin, H.; Xu, X.; Wu, Z.; Zhou, J.; Lu, J. Sg-nav: Online 3d scene graph prompting for llm-based zero-shot object navigation. *Advances in neural information processing systems* **2024**, *37*, 5285–5307.
257. Huang, X.; Zhao, S.; Wang, Y.; Lu, X.; Zhang, W.; Qu, R.; Li, W.; Wang, Y.; Wen, C. Msgnav: Unleashing the power of multi-modal 3d scene graph for zero-shot embodied navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 37154–37163.
258. Dorbala, V.S.; Sigurdsson, G.; Piramuthu, R.; Thomason, J.; Sukhatme, G.S. Clip-nav: Using clip for zero-shot vision-and-language navigation. *arXiv preprint arXiv:2211.16649* **2022**.
259. Zhang, W.; Zhang, J. Language-Driven Zero-Shot Object Navigation via Dynamic Probabilistic Strategy and Large Language Models. *IEEE Access* **2025**.
260. Majumdar, A.; Aggarwal, G.; Devnani, B.; Hoffman, J.; Batra, D. Zson: Zero-shot object-goal navigation using multimodal goal embeddings. *Advances in Neural Information Processing Systems* **2022**, *35*, 32340–32352.
261. Chen, J.; Lin, B.; Liu, X.; Ma, L.; Liang, X.; Wong, K.Y.K. Affordances-oriented planning using foundation models for continuous vision-language navigation. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 23568–23576.
262. Team, I.N. InternVLA-N1: An Open Dual-System Navigation Foundation Model with Learned Latent Plans, 2025.
263. Gao, C.; Peng, X.; Yan, M.; Wang, H.; Yang, L.; Ren, H.; Li, H.; Liu, S. Adaptive zone-aware hierarchical planner for vision-language navigation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14911–14920.
264. Song, X.; Chen, W.; Liu, Y.; Chen, W.; Li, G.; Lin, L. Towards long-horizon vision-language navigation: Platform, benchmark and method. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025, pp. 12078–12088.
265. Han, Z.; Wang, X.; Liu, B.; Lyu, Q.; Shang, Z.; Dong, J.; Liu, L.; Han, Z. SeqWalker: Sequential-Horizon Vision-and-Language Navigation with Hierarchical Planning. *arXiv preprint arXiv:2601.04699* **2026**.
266. Dai, G.; Wang, S.; Wang, Z.; Xie, G.S.; Yang, Y.; Pan, J.; Sun, Q.; Shu, X. History to Future: Evolving Agent with Experience and Thought for Zero-shot Vision-and-Language Navigation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2026, pp. 15177–15187.
267. Wang, X.; Li, G.; Liu, Z.; Wang, Y.; Liu, L.; Han, Z. All-day multi-scenes lifelong vision-and-language navigation with tucker adaptation. *arXiv preprint arXiv:2603.14276* **2026**.
268. Jiang, Y.; Zhang, H.; Luo, X.; He, S. M^3E: Continual Vision-and-Language Navigation via Mixture of Macro and Micro Experts. In Proceedings of the The Fourteenth International Conference on Learning Representations.
269. Yao, X.; Gao, J.; Xu, C. Navmorph: A self-evolving world model for vision-and-language navigation in continuous environments. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2025, pp. 5536–5546.
270. Pei, J.; Liu, Y.; Pan, G.; Jiang, Y.; Liu, H.; Wang, X. OVAL: Open-Vocabulary Augmented Memory Model for Lifelong Object Goal Navigation. *arXiv preprint arXiv:2604.12872* **2026**.
271. Tian, H.; Meng, J.; Zheng, W.S.; Li, Y.M.; Yan, J.; Zhang, Y. Loc4plan: Locating before planning for outdoor vision and language navigation. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 4073–4081.
272. Elnoor, M.; Weerakoon, K.; Seneviratne, G.; Xian, R.; Guan, T.; Jaffar, M.K.M.; Rajagopal, V.; Manocha, D. VLM-GroNav: Robot Navigation Using Physically Grounded Vision-Language Models in Outdoor Environments. In Proceedings of the 2025 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2025, pp. 2391–2398.
273. Xu, Y.; Pan, Y.; Liu, Z.; Wang, H. Flame: Learning to navigate with multimodal llm in urban environments. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2025, Vol. 39, pp. 9005–9013.
274. Ning, Y.; Zhao, G.; Qin, Y.; Liu, S.; Liu, Y.; Lin, L.; Li, G. LookasideVLN: direction-aware aerial vision-and-language navigation. *arXiv preprint arXiv:2604.17190* **2026**.

 doi:10.20944/preprints202606.2231.v1

275. Taioli, F.; Rosa, S.; Castellini, A.; Natale, L.; Del Bue, A.; Farinelli, A.; Cristani, M.; Wang, Y. I2EDL: Interactive Instruction Error Detection and Localization. In Proceedings of the 2024 33rd IEEE International Conference on Robot and Human Interactive Communication (ROMAN). IEEE, 2024, pp. 1872–1877.
276. Li, C.; Tang, W.; Huang, Y.; Zhan, S.S.; Hu, M.; Jia, X.; Liu, Y. Shedding Light on VLN Robustness: A Black-box Framework for Indoor Lighting-based Adversarial Attack. *arXiv preprint arXiv:2511.13132* **2025**.
277. Zhang, H.; Xu, M.; Dhafer, A.; Yue, S.; Dong, H.; Hao, Z.D. Embodied Interpretability: Linking Causal Understanding to Generalization in Vision-Language-Action Models. *arXiv preprint arXiv:2605.00321* **2026**.
278. Wang, Z.; Li, X.; Yang, J.; Liu, Y.; Jiang, S. Sim-to-real transfer via 3d feature fields for vision-and-language navigation. *arXiv preprint arXiv:2406.09798* **2024**.
279. Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; Chen, E. A survey on multimodal large language models. *National Science Review* **2024**, *11*, nwae403.
280. Caffagni, D.; Cocchi, F.; Barsellotti, L.; Moratelli, N.; Sarto, S.; Baraldi, L.; Cornia, M.; Cucchiara, R. The revolution of multimodal large language models: A survey. *Findings of the association for computational linguistics: ACL 2024* **2024**, pp. 13590–13618.
281. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen technical report. *arXiv preprint arXiv:2309.16609* **2023**.
282. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* **2023**.
283. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International conference on machine learning. PMLR, 2023, pp. 19730–19742.
284. Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; Zhang, Y. Matterport3D: Learning from RGB-D Data in Indoor Environments. In Proceedings of the International Conference on 3D Vision (3DV), 2017.