Vision-and-Language Navigation: A Comprehensive Review of Tasks, Methods, and Challenges

Anqi Song[1, a], Aobing Yin[2, b], Chenghe Kong[3, c], Yuhuan Xie[4, d]

[1] No.9, Wenyuan Road, Yadong New District, Nanjing 210023, China

[2] Imperial College London,The faculty Building, Imperial College London, Exhibition Road, South Kensington, London, SW7 2AZ., UK

[3] University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA 15260, USA

[4] Wuhan University of Technology, 122 Luoshi Road, Hongshan District, Wuhan 430070, China

[a] 352135410@qq.com, [b] jaykdk66@gmail.com, [c] 694502018@qq.com, [d] 2423235414@qq.com

Abstract

Vision-and-Language Navigation (VLN) is a core challenge in embodied AI, which aims to develop agents capable of understanding natural language instructions and navigating autonomously in visual environments. This survey systematically reviews the task paradigms and cutting-edge progress in the VLN field. We first propose a four-quadrant taxonomy based on environment, interaction, and instruction modality (Indoor, Outdoor, Interactive, and Multimodal-instruction Navigation), using this framework to deeply analyze the core characteristics, evaluation metrics, and technical challenges of various representative datasets. Furthermore, we provide a detailed review of mainstream technical methods, including classical modular paradigms, end-to-end learning (reinforcement learning and imitation learning), pre-training and transfer learning strategies, as well as advanced methods based on memory and graph structures, discussing their respective advantages, disadvantages, and applicable scenarios. Finally, we summarize the current challenges faced by the field, such as simulation-to-reality transfer, long-horizon planning, interactive reasoning, and embodied learning, and prospect future research directions. This survey aims to provide researchers with a clear technological panorama, promoting the development of VLN technology towards more general, robust, and practical applications.

Keywords

Vision-and-Language Navigation, embodied AI, multimodal learning, deep learning, survey.

1. Introduction

With the rapid development of deep learning and multimodal fusion technologies, enabling agents to understand natural language and interact physically like humans has become an important frontier in artificial intelligence. Vision-and-Language Navigation (VLN) [1], as a representative task in this direction, aims to empower agents with the ability to navigate complex visual environments based on intricate natural language instructions. This research is not only crucial for developing applications like home service robots, autonomous driving, and navigation assistance systems for the visually impaired but also serves as a key testbed for exploring the cognitive mechanisms linking vision, language, and action. Although significant progress has been made in VLN research in recent years, it still faces numerous severe challenges. Firstly, agents require deep cross-modal understanding capabilities to accurately align abstract linguistic concepts (e.g., "go through the living room and find the nightstand on

the left side of the bedroom") with high-dimensional visual perceptual information. Secondly, the complexity of real environments, partial observability, and the need for generalization to unseen scenarios demand effective long-horizon planning and decision-making from agents. Furthermore, instruction ambiguity often requires clarification through multi-turn human-agent interaction, and the inefficiency of purely linguistic instructions in certain scenarios has spurred the development of navigation paradigms incorporating multimodal signals such as gestures and sketches.

To address these challenges, researchers have constructed diverse task datasets (e.g., R2R [1], REVERIE [9], ALFRED [7]) and proposed a wide variety of technical methods. From early modular systems reliant on environmental priors to imitation and reinforcement learning strategies based on end-to-end deep learning, and further to recent paradigms leveraging large-scale pre-trained models and explicit structured memory, the development path of VLN technology demonstrates an evolution from the separation of "perception-planning-control" to multimodal joint representation and decision-making.

This survey aims to provide a systematic organize and summary of the VLN field. In Chapter II, we will first propose a multidimensional classification framework and detail the core features and technical assumptions of the four major categories of tasks and datasets-Indoor, Outdoor, Interactive, and Multimodal-instruction Navigation-within this framework. In Chapter III, we will delve into the key technical methods of VLN, including modular paradigms, end-to-end learning, pre-training and transfer learning, as well as advanced methods based on memory and graph structures. In Chapter IV, we will discuss the core challenges currently facing the field and prospect future research directions. Finally, Chapter Vconcludes the paper.

2. Vision-and-Language Navigation: Task Space and Dataset Landscape

2.1. A Multi-Dimensional Taxonomy for VLN

Vision-and-Language Navigation (VLN) research focuses on developing autonomous agents that can interpret natural language instructions and navigate through complex, visually perceived environments to achieve specified goals. The rapid proliferation of diverse tasks and benchmarks in this field necessitates a structured framework for comprehension and comparison. We thus adopt a four-way taxonomy, categorizing VLN tasks along three core design axes: (1) the type of environment (Indoor vs. Outdoor), (2) the nature of agent-environment interaction (Non-Interactive vs. Interactive), and (3) the modality of the instruction (Language-Only vs. Multimodal). This taxonomy elucidates the fundamental assumptions inherent to each benchmark regarding environmental structure, action space, and human-agent communication, providing a consistent scaffold for analyzing and contrasting tasks, datasets, and evaluation protocols.

2.2. Indoor Navigation

Indoor navigation serves as the foundational setting for VLN research, with the Room-to-Room (R2R) dataset [1] being its most prominent instantiation. R2R pairs human-authored navigation instructions with shortest-path trajectories between points in real 3D building scans from Matterport3D. Its key innovation was the introduction of *val-seen*and *val-unseen*splits, rigorously testing an agent's ability to generalize to entirely novel building layouts not encountered during training. The task is built upon the Matterport3D simulator, which provides a navigable graph where nodes represent discrete panoramic viewpoints. Standard evaluation metrics include Success Rate (SR) (whether the agent stops within a threshold, e.g., 3 meters, of the goal) and Navigation Error (NE) (the final geodesic distance to the goal). SPL [2], a metric that combines success with path efficiency, and Oracle Success Rate (success if the agent ever visited the goal) are also widely adopted. Methodologically, agents typically process

a panoramic view by discretizing it into multiple single images (e.g., 12 headings × 3 elevations). A crucial component is an explicit STOP action, requiring the agent to autonomously decide when the instruction has been completed. The core challenges in this setting are precise cross-modal grounding, long-horizon execution under partial observability, and most critically, robust generalization to unseen environments. Notable methodological innovations under the R2R setup include.

2.3. Outdoor Navigation

Outdoor VLN moves from confined building interiors to expansive, unstructured urban and geographic environments. This shift introduces significant new challenges: vastly longer trajectories, more complex instructions rich with geographic landmarks, and highly dynamic visual conditions with occlusion, lighting changes, and clutter. Touchdown [7] formalizes outdoor VLN as a two-stage task within Google Street View: first, follow navigation instructions; second, upon stopping, resolve a Spatial Description Resolution (SDR) task by identifying a described target location in the final panorama. Its instructions and paths are substantially longer than R2R (mean ∼108 tokens and ∼35 panoramas vs. ∼29 tokens and ∼6 panoramas in R2R), intensifying the demands on long-horizon reasoning and spatial understanding. Linguistic analysis reveals a rich use of both egocentric ("left/right") and allocentric ("north/south") spatial relations. Talk2Nav [8] addresses long-range urban navigation by constructing a detailed city-level graph. Its key innovation is an annotation pipeline that decomposes monolithic instructions into segmented sub-phrases corresponding to landmarks and local directions. This structure enables novel agent architectures with dual attention mechanisms (over landmarks and directions) and explicit spatial memory, facilitating more robust reasoning over very long routes.

Outdoor VLN benchmarks like Touchdown and Talk2Nav feature longer, more complex paths and instructions than their indoor counterparts, forcing advancements in long-horizon planning, robust visual grounding under real-world visual clutter, and sophisticated spatial reasoning. Annotation innovations, such as instruction segmentation, are critical for supporting these advancements.

2.4. Interactive Navigation

Interactive Navigation represents a paradigm shift from passive path-following to active interaction. Tasks in this category require the agent to not only navigate but also manipulate objects to change the environment's state (e.g., pick up an apple, slice it, then place it on a counter). This makes planning and evaluation fundamentally different, necessitating metrics that credit partial task completion. ALFRED [9] is a benchmark for long-horizon, compositional tasks set in the AI2-THOR simulator. It provides expert demonstrations of tasks involving sequences of navigation and manipulation actions. Key metrics include Task Success(complete task execution) and Goal-Condition Success (GC), which credits the correct state of each sub-goal object, even if the full sequence wasn't completed. Path-weighted metrics penalize inefficiency. Despite dense supervision, initial models showed extremely low performance on unseen scenes, highlighting the immense difficulty of generalized embodied instruction following. TEACh [10] studies dialogue-driven embodied task completion. It involves two agents: a Commander (with oracle knowledge of the task) and a Follower (which acts in the environment). They coordinate through natural language dialogue to complete multi-step household goals. The dataset captures over 3,000 human-human dialogue sessions, enabling research into dialogue understanding, grounding, and recovery from errors under realistic ambiguity. Interactive benchmarks like ALFRED and TEACh push agents beyond waypoint tracking to state-changing, compositional tasks. They introduce partial-credit metrics (GC) and demonstrate the utility of dialogue for disambiguation and recovery. However, current systems remain brittle, with low success rates in unseen environments, particularly in grounding fine-

grained object states and affordances. The complexity of integrating robust dialogue with precise manipulation remains a major hurdle.

2.5. Multimodal-Instruction Navigation

This emerging category seeks to move beyond the exclusivity of textual instructions by incorporating non-verbal communicative channels like gestures and sketches, which are natural and efficient components of human navigation communication. **GesNavi** [11] provides a dataset for **gesture-guided navigation** in a simulated urban environment. Each instruction pairs text with a gestured object (an intermediate anchor), explicitly requiring multi-hop reasoning between the language and the gesture to find the final target. Analysis shows humans tend to gesture towards closer, larger objects, revealing pragmatic patterns that can inform model design. A limitation is its reliance on simulation and a single gestured object per instruction. **Sketch-Based Navigation** [12] explores the use of hand-drawn sketches as a language-agnostic channel for conveying navigation intent. Sketches excel at efficiently communicating path topology, spatial relationships, and landmarks, especially when verbal descriptions are ambiguous or require fine-grained spatial detail. They complement text by reducing linguistic ambiguity and cognitive load. Multimodal-instruction navigation incorporates gestures and sketches to provide potent disambiguation signals and enable more natural human-agent communication. These datasets facilitate research into multi-hop, cross-modal reasoning. Current challenges include the **sim-to-real gap** in gesture perception, limited scope of gestures (often a single referent), and the complexity of annotating aligned multi-modal data, which complicates benchmarking.

The VLN task space has evolved from its foundation in indoor instruction following (R2R) to encompass more complex outdoor, interactive, and multimodal paradigms. This evolution has been driven by the desire for greater realism, which in turn introduces progressively harder challenges: from geometric generalization in unseen buildings, to long-horizon reasoning in cities, to compositional planning with object manipulation, and finally, to grounding instructions that fuse multiple communicative modalities.

The following table provides a concise comparative overview of the core datasets discussed across these categories:

Table 1. overview of the core datasets

Dataset	Primary Category	Environment	Key Innovation / Task	Instruction Length (Avg.)	Path Length (Avg.)	Core Metric (s)
R2R[1]	Indoor	Matterport3D Homes	Basic instruction following	~29 words	~6 panoramas	SR, SPL, NE
REVERIE [5]	Indoor	Matterport3D Homes	Remote object grounding	~18 words	-	RGS, SR, SPL
CVDN [6]	Indoor (Interactive)	Matterport3D Homes	Navigation from dialogue history	~82 words/dialog	~25 steps	Goal Progress
Touchdown [7]	Outdoor	Google Street View	Navigation + Spatial Description Resolution	~108 tokens	~35 panoramas	SR (Nav & SDR)
ALFRED [9]	Interactive	AI2-THOR (Indoor)	Navigation + Manipulation	-	-	Task Success, GC
GesNavi [11]	Multimodal	Simulated Urban	Gesture-guided navigation	~13 words	-	SR

3. Key Methods for Vision-Language Navigation

3.1. Classical Modular Paradigms

Early VLN systems adopted a modular design, decomposing the problem into sequential sub-tasks: mapping the environment, grounding the instruction, planning a global path, and executing local control. The primary advantage of this approach is its interpretability and the ability to incorporate explicit spatial reasoning and safety constraints. However, it often requires significant engineering and is prone to error propagation between modules.

A representative framework involves several key components:

Mapping & Localization: Building a persistent representation of the explored environment, often a topological or metric-semantic map.

Instruction Parsing & Grounding: Decomposing the natural language instruction into a sequence of sub-goals and associating spatial references with entities in the map.

Global Planning: Generating a feasible path from the agent's current location to the final goal, consistent with the parsed instruction. This can be formulated as a graph search problem (e.g., A*) with constraints derived from the language.

Local Control: Executing low-level actions (e.g., move_forward, turn_left) to follow the planned path, often using a separate controller.

Methods like TPT (Topological Planning and Transformation) exemplify this paradigm. TPT first constructs a topological graph of the environment. A cross-modal planner then generates a sequence of waypoints conditioned on the instruction, and a low-level controller executes actions to reach these waypoints, allowing for backtracking if the plan becomes invalid. This decoupling mitigates the exposure bias common in end-to-end models.

For interactive tasks like those in ALFRED, the modular stack becomes more complex, involving parsing instructions into a sequence of skills (nav_to, pickup, heat, place), grounding these skills to objects in a pre-built scene graph, and executing them through a library of predefined controllers. This explicit decomposition enhances transparency and controllability at the cost of requiring extensive engineering and being difficult to adapt to entirely new instructions or environments.

In summary, modular paradigms offer interpretability and strong global planning but are often brittle and struggle to generalize due to their reliance on hand-designed components and the compounding of errors between modules.

3.2. End-to-End Learning Methods

To overcome the limitations of modular systems, end-to-end learning approaches leverage deep neural networks to map directly from multi-modal inputs (visual observations and instructions) to actions, jointly learning perception, grounding, and decision-making.

3.2.1. Imitation Learning (IL)

Imitation Learning, particularly Behavior Cloning (BC), treats VLN as a supervised sequence-to-sequence problem. The agent is trained to predict the next action given the current observation and the instruction, using expert trajectories. While simple and data-efficient, BC suffers from **covariate shift**-small deviations at test time compound, causing the agent to veer off the training data distribution.

A pivotal innovation to address data scarcity and covariate shift is the **Speaker-Follower** model [3]. It introduces a dual process: a *Follower*agent learns to navigate, while a *Speaker*agent is trained to generate instructions for trajectories. The Speaker enables **back-translation** (synthesizing new instruction-path pairs) and **pragmatic reasoning** at test time, where the Follower re-ranks candidate paths based on their likelihood under the Speaker.

3.2.2. Reinforcement Learning (RL)

Reinforcement Learning frames navigation as a Markov Decision Process, optimizing a policy to maximize cumulative reward. The primary challenge is the **sparsity of rewards**; a success signal is only provided upon completing the task, making learning over long horizons inefficient.

A common solution is **reward shaping**, which provides intermediate rewards to guide the agent (e.g., rewarding a decrease in the geodesic distance to the goal). The **RCM (Reinforced Cross-Modal Matching)** method provides a more nuanced approach by learning an intrinsic reward signal. RCM uses a "matching critic" that evaluates how well a trajectory aligns with the instruction, providing dense feedback that promotes cross-modal grounding rather than mere geometric progress. It is often combined with **Self-Imitation Learning (SIL)**, where the agent reinforces its own successful past trajectories, improving generalization to unseen environments.

3.2.3. Hybrid and Interactive Training

Hybrid methods combine the strengths of IL and RL. DAgger is a prominent algorithm that addresses covariate shift by having the agent roll out its current policy and then query an expert for corrective actions on visited states, aggregating these into a new training set.

Another line of work explores model-based RL, where an agent learns a predictive world model of the environment. This model can be used for internal simulation ("look-ahead"), aiding long-horizon planning and improving sample efficiency by reducing the reliance on expensive real interactions.

3.3. Pre-training and Transfer Learning Paradigms

Acknowledging the data-scarce nature of VLN, pre-training on large-scale external corpora has become a dominant strategy for learning robust, transferable visio-linguistic representations.

VLN-BERT exemplifies this paradigm. It involves a two-stage process:

Generic Visio-Linguistic Pre-training: A transformer model is first pre-trained on massive image-text pairs (e.g., Conceptual Captions) using objectives like masked language modeling and image-text matching. This teaches the model fundamental cross-modal alignment.

VLN-specific Fine-tuning: The pre-trained model is then adapted to navigation, often by encoding panoramic observations and instructions and fine-tuning for trajectory ranking or action prediction.

This paradigm has been extended in creative ways to further reduce reliance on human annotations:

Self-Supervised Exploration (ProbES): Agents explore environments without instructions, using models like CLIP to automatically generate structured pseudo-instructions for their trajectories, creating a self-supervised pre-training dataset.

Video-Based Pre-training (YouTube-VLN): Mining house tour videos to create large-scale, albeit noisy, instruction-path pairs for pre-training, capturing realistic layouts and diverse language.

Beyond sequence-based pre-training, **spatially explicit pre-training** has emerged. **BEVBert** constructs a local metric map from ego-centric observations and pre-trains a model with cross-modal objectives over this unified spatial representation. This explicitly encourages the model to learn geometry and topology, leading to more robust long-horizon planning [13].

3.4. Memory and Graph-Structured Methods

To address the challenge of long-horizon planning and information retention, advanced architectures incorporate explicit, structured memory.

History-Aware Memory: The **Episodic Transformer (E.T.)** replaces recurrent state with a memory bank that stores the entire history of observations and actions. The model attends over

this full history, enabling better credit assignment and reasoning over past events. This is particularly effective for compositional tasks in **ALFRED**. **MemoNav**introduces a cognitively-inspired gating mechanism, separating Short-Term Memory (STM) and Long-Term Memory (LTM). An attention-based controller selectively retains salient information in STM and summarizes it into a global LTM node, reducing memory noise and improving efficiency.

Topological Graph Memory: These methods explicitly maintain the explored environment as a graph, where nodes represent visited states and edges represent navigable connections. This structure enables efficient global planning and re-localization. **DUET** constructs a online topological map and uses a dual-scale graph transformer: a coarse graph for long-range planning and a fine-grained graph for local grounding, achieving state-of-the-art performance on several benchmarks. **ETPNav** dynamically builds a topological graph in continuous environments, coupling a cross-modal planner with a low-level controller, effectively decoupling high-level planning from obstacle avoidance. **PRET** defines a differentiable graph over the explored area and scores candidate paths based on their visual-linguistic alignment with the instruction, enabling efficient and interpretable global planning.

3.5. Summary and Comparative Analysis

The methodological evolution in VLN reflects a shift from engineered modularity to learned, end-to-end models, and recently towards a hybrid philosophy that incorporates the strengths of both: leveraging the power of learning while retaining elements of structure for improved reasoning and generalization. Pre-training has become a foundational step, equipping models with essential visio-linguistic commonsense.

The following table provides a comparative overview of these core methodological families:

Table 2. Overview of these core methodological families

Method Category	Key Idea	Advantages	Limitations	Representative Models
Modular	Decompose task into sequential sub-modules	Interpretable, strong global planning, controllable	Error propagation, engineering heavy, poor generalization	TPT, IPPD
End-to-End (IL)	Supervised learning from expert demonstrations	Simple, data-efficient	Covariate shift, poor generalization	Behavior Cloning, Speaker-Follower [3]
End-to-End (RL)	Learn a policy to maximize reward	Optimizes for task completion, closed-loop	Sparse reward, sample inefficient	RCM+SIL, EnvDrop [4]
Pre-training	Transfer knowledge from large vision-language corpora	Improved generalization, data efficiency	Computationally expensive, domain gap	VLN-BERT, BEVBert
Structured Memory	Maintain explicit memory of history/environment	Improves long-horizon reasoning, reduces perceptual aliasing	Increased model complexity	Episodic Transformer, DUET

This landscape of methods provides a rich toolkit for tackling the diverse challenges posed by the different VLN tasks outlined in Chapter II. The optimal approach often depends on the specific task constraints, available data, and requirements for interpretability.

4. Current Challenges and Future Research Direction

Although vision-and-language navigation research has made significant progress in task paradigms and model innovation, a considerable gap remains between current technological levels and achieving human-like fluent and robust navigation capabilities. Agent performance often degrades significantly when faced with unknown environments, complex instructions, and multimodal interactions. This chapter will systematically explore the core challenges facing the field and prospect future research directions based on these.

Current VLN technology first faces Environmental and Embodied Challenges. Among these, the Simulation-to-Reality (Sim-to-Real) Gap is the primary bottleneck hindering its move towards practical application. Mainstream research heavily relies on simulators like Matterport3D and AI2-THOR. The rendered images from these simulators exhibit domain differences compared to the real world in terms of texture, lighting, and dynamic objects, causing the visual representations trained in simulators to suffer from sharply reduced generalization ability in real scenarios [6]. Additionally, the simplification of physical dynamics (e.g., friction, precise grasping) in simulators and the provision of perfect perceptual information (noise-free depth, segmentation) make policies trained in these environments ill-equipped to handle the noise, occlusion, and calibration errors present in real sensors. On the other hand, most navigation benchmarks (e.g., R2R) are built upon discrete panoramic graph nodes, and their action space (e.g., turn, move forward) vastly differs from the continuous velocity control and fine-grained low-level manipulation (e.g., grasping and placing in ALFRED) required by real robots, posing higher demands on the model's fine control capabilities.

Secondly, Task and Cognitive Complexity presents severe tests for the agent's reasoning abilities. Long-Horizon and Compositional Task Planning is one of the core difficulties. Faced with lengthy, multi-subgoal instructions, models often struggle with the credit assignment problem, finding it difficult to attribute final success or failure to early key decisions. Simultaneously, once deviating from the correct path during testing, "exposure bias" can lead to error accumulation, and agents generally lack effective error recovery mechanisms. Even with long-context Transformers, models may still forget crucial information from the early parts of the instruction or lose sight of the ultimate goal. Furthermore, as revealed by tasks like CVDN and TEACh, Interactive and Dialogical Reasoning capabilities are crucial. Agents not only need to learn whento ask actively (When to Ask) to clarify ambiguities but also must generate the most informative questions (What to Ask), all while continuously tracking the multi-turn dialog state throughout this process. This places extremely high demands on the model's contextual understanding and generation capabilities.

Thirdly, limitations in Generalization and Evaluation Systems make it difficult to measure true progress. Although the "unseen environments" setting has become a standard evaluation practice, generalization challenges extend far beyond this. Models need to handle novel linguistic expressions, unknown object categories, and drastically different environmental layout styles not present in the training data; their Domain Generalization ability remains very weak. At the same time, the limitations of existing Evaluation Metrics (e.g., SPL, SR), although widely used, are becoming increasingly apparent. They primarily measure the success and efficiency of the final outcome, unable to assess whether the agent truly understands the instruction semantics or has merely learned dataset biases. In interactive tasks, existing metrics still insufficiently capture partial success and lack human evaluation standards capable of measuring the "reasonableness" of behavior.

Finally, System and Application Practical Constraints are obstacles that must be overcome for technological deployment. The computational overhead of many advanced models (e.g., large Transformers, complex memory structures) is immense, making it difficult to meet the millisecond-level response requirements for Real-Time Performance and Computational Efficiency in physical robots. More importantly, when VLN agents are deployed in the real world, Safety and Ethical Issues become unavoidable. Their behavior must comply with social norms and be able to predict and avoid potential harm to people or the environment. Additionally, potential linguistic and cultural biases present in the training data can be absorbed and amplified by the models, leading to unfair performance across different user groups-an issue that urgently needs addressing.

Looking forward, several research directions hold significant potential. The foremost is Building Higher-Fidelity Simulation and Data Ecosystems, developing simulation platforms with more accurate physics, more realistic visuals, and incorporating dynamic elements, while leveraging generative AI to synthesize large-scale, high-quality instruction-trajectory data. Secondly, there is a need to Develop More Powerful World Models and Reasoning Architectures, exploring models capable of counterfactual reasoning and long-range temporal dependency modeling, enabling agents to possess inner simulation capabilities. Promoting the Integration of "Large Models + Robotics" is a highly promising direction, fully utilizing the commonsense knowledge of Large Language Models (LLMs) and Vision-Language Models (VLMs) for task planning and code generation, while delegating low-level control to specialized modules. Furthermore, exploring Embodied Multimodal Learning Paradigms that construct internal representations of space and physics through autonomous learning and environmental interaction will reduce dependence on fully supervised data. Finally, designing Fairer, More Comprehensive, and Efficient Evaluation Benchmarks is also crucial, requiring the establishment of unified benchmarks covering cross-task and cross-language generalization capabilities, and developing evaluation protocols that can more deeply reflect the agent's understanding.

5. Conclusion

After years of development, Vision-and-Language Navigation has evolved from simple instruction following in constrained indoor environments to a broad research field encompassing complex environments (indoor and outdoor), rich interactions (dialogue, manipulation), and multimodal instructions (gestures, sketches). This article synthesizes the complex task space through a unified four-quadrant taxonomy and systematically reviews the evolution of key technological paradigms, from modular design to end-to-end learning, and further to pre-training and structured memory.Despite significant progress, a vast gap remains between the most advanced current systems and human-level navigation capabilities. Future research is expected to make breakthroughs in the following directions: 1Generalization and Robustness: Achieving zero-shot or few-shot generalization to truly unseen scenes through large-scale unsupervised environmental exploration and cross-modal pre-training, reducing reliance on expensive human-annotated data. 2) Efficient Interaction and Reasoning: Developing more natural and efficient multi-turn interaction mechanisms, enabling agents to actively ask questions, confirm information, and perform causal and counterfactual reasoning to handle ambiguous instructions and unexpected situations. 3) From Navigation to Embodied Manipulation: Promoting the deep integration of pure navigation tasks with embodied manipulation, completing the full loop of "perception-navigation-interaction-reasoning," and ultimately enabling the execution of complex multi-step everyday tasks. 4) From Simulation to Reality: Overcoming the visual, dynamic, and interactive disparities between simulation and reality by building higher-fidelity simulators and developing sim-to-real transfer techniques,

eventually deploying systems on physical robots. As a key bridge connecting perception, language, and action, the development of VLN will undoubtedly greatly advance the realization of general embodied AI agents. We look forward to more interdisciplinary collaboration in the future to jointly tackle this challenging field.

References

[1] Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., & Van Den Hengel, A. (2018). Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3674-3683.

[2] Chen, H., Suhr, A., Misra, D., Snavely, N., & Artzi, Y. (2019). TOUCHDOWN: Natural language navigation and spatial reasoning in visual street environments. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12538-12547.

[3] Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., & Fox, D. (2020). ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[4] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W. Y., Shen, C., & Van Den Hengel, A. (2020). REVERIE: Remote embodied visual referring expression in real indoor environments. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[5] Jain, A., Misu, T., Yamada, K., & Yanaka, H. (2024). GesNavi: Gesture-guided outdoor vision-and-language navigation. Proceedings of the 18th Conference of the European Chapter of the ACL: Student Research Workshop (EACL-SRW), 290-295.

[6] Krantz, J., Wijmans, E., Majumdar, A., Batra, D., & Lee, S. (2020). Beyond the nav-graph: Vision-and-language navigation in continuous environments. European Conference on Computer Vision (ECCV), 104-120.

[7] Fried, D., Hu, R., Cirik, V., Rohrbach, A., Andreas, J., Morency, L.-P., ... & Darrell, T. (2018). Speaker-Follower models for vision-and-language navigation. Advances in Neural Information Processing Systems (NeurIPS).

[8] Tan, H., Yu, L., & Bansal, M. (2019). Learning to navigate unseen environments: Back translation with environmental dropout. Proceedings of NAACL-HLT, 2610–2621.

[9] Vasudevan, A. B., Dai, D., & Van Gool, L. (2021). Talk2Nav: Long-range vision-and-language navigation with dual attention and spatial memory. International Journal of Computer Vision, 129(1), 246–266.

[10] Thomason, J., Murray, M., Cakmak, M., & Zettlemoyer, L. (2020). Vision-and-Dialog Navigation. Proceedings of the Conference on Robot Learning (CoRL), in Proceedings of Machine Learning Research, 100, 394–406.

[11] Jain, A., Misu, T., Yamada, K., & Yanaka, H. (2024). GesNavi: Gesture-guided outdoor vision-and-language navigation. Proceedings of the 18th Conference of the European Chapter of the ACL: Student Research Workshop (EACL-SRW), 290–295.

[12] Ahmad, H., Usama, S. M., Hussain, W., & Anjum, M. L. (2021). A sketch is worth a thousand navigational instructions. Autonomous Robots, 45(2), 313–333.

[13] Krantz, J., Wijmans, E., Majumdar, A., Batra, D., & Lee, S. (2020). Beyond the nav-graph: Vision-and-language navigation in continuous environments. In European Conference on Computer Vision (ECCV) (pp. 104–120). Cham: Springer.