Can VLM understand affordance in VLN

Guoli Huang *

June 15, 2025

*Supervised by Prof. Damith c. Ranasinghe

1 Introduction

Vision-Language Navigation (VLN) is a core research topic in the field of embodied AI; it requires the agent to navigate in a 3D environment based on natural language instructions and visual observations. The inherent challenge of VLN lies in bridging the gap between language comprehension, visual perception, and sequential decision-making within complex and unseen environments [1, 2]. This task integrates multiple disciplines such as artificial intelligence, natural language processing, computer vision, and robotics.

Fundamentally, the VLN task can be framed as a Partially Observable Markov Decision Process (POMDP), formally defined by the tuple $(S, A, T, R, \Omega, O, \gamma)$. This is because the agent, at any given moment, has only an incomplete view of the world and must make sequential decisions under this perceptual uncertainty [3]. Although these early methods laid the foundation, they faced limitations in generalization, long-term planning, and complex scene understanding. For example, the Behavioral Cloning method is prone to error accumulation due to the covariate shift issue [4, 5].

The limitations of these early methods highlighted the need for a standardized and realistic benchmark. To meet this demand, the R2R dataset proposed by Anderson et al. significantly advanced the field by providing a benchmark for vision-and-language navigation in real building scans [1]. On this new benchmark, sequence-to-sequence (Seq2Seq) models, a popular paradigm at the time for sequence modeling tasks, were naturally applied and achieved some initial success. However, these models performed poorly in generalizing to unseen environments, leading researchers to explore more sophisticated memory, planning, and representation techniques [6].

Consequently, the frontier of research has shifted from achieving basic instruction following in simulated environments to robust generalization in unknown settings, long-term task execution, and real-world applications such as continuous environments (VLN-CE) and unmanned aerial vehicles (UAVs) [7]. To address the aforementioned generalization challenges, recent research has increasingly focused on enhancing the decision-making of the agent by incorporating a deeper understanding of the environment. This review will delve into progress in this direction, with a specific focus on the integration of physical priors and, more pointedly, the concept of affordances.

1.1 The Role of Physical Priors

A persistent challenge in VLN is that agents often perceive the world through a monocular camera with 2D features. This compression of depth information sacrifices the completeness of scene structures in space. To overcome this, many studies focus on representation methods that convey more complete physical information. One significant advancement is the shift towards volumetric representations, such as the Volumetric Environment Representation (VER), which aggregates multiview 2D features into a unified 3D voxel space to capture comprehensive geometric and semantic information, including spatial attributes such as occupancy [8].

Beyond simply representing the physical world, another line of research focuses on how an agent can use this understanding to guide its behavior [7]. This integration of physical and geometric priors is advanced in two main ways. An agent can implicitly learn to navigate by mimicking expert data that inherently adhere to physical rules. Alternatively, the behavior of the agent can be explicitly limited by its physical understanding, for example, by planning paths only within areas identified as navigable by a representation such as VER [8, 9].

1.2 A Focus on Affordance

This report focuses on a promising direction in physical priors: affordances. Affordances refer to the potential actions that an environment or objects within it may offer or influence for an agent

[10]. In the context of navigation, this means identifying accessible paths, understanding how interactive objects affect planning instructions, or recognizing object properties that facilitate or restrict actions [11].

Although many successful VLN agents implicitly learn affordances during training (for example, recognizing that certain objects are impassable), we argue that explicitly modeling and reasoning about affordances is an emerging and highly promising field [1,12]. It can help agents shift from a static understanding of "what the environment is" to a dynamic, interactive comprehension of "what I can do with it." This shift is crucial for tackling increasingly complex VLN tasks, especially those requiring interaction with the environment to alter its state. The core argument of this paper is that affordances serve as another critical layer of information that bridges abstract language and visual perception with concrete physical actions.

2 A Review of Affordance Theory and Applications

2.1 Affordance Theory

2.1.1 Ecological Foundations: Gibson's Theory

The concept of affordance was first introduced by James J. Gibson in his groundbreaking research on ecological psychology. In his 1979 book "The Ecological Approach to Visual Perception," Gibson defined affordance as "what the environment offers the animal, what it provides or furnishes, either for good or ill. [13]" This definition reveals the essence of affordance—they are attributes that exist in the environment relative to the action capabilities of a specific actor (animal or observer) [14]. For example, when we perceive stairs, we do not focus on their height in centimeters but rather on whether they are "step-able. [15]" This means that a surface may provide "stand" for an adult but not for an infant, depending on their body sizes, structure, and movement capabilities.

Gibson's affordance theory is built upon several core principles, which together form the foundation of his ecological perspective [13, 14, 16–18]:

Direct Perception: The information in the environment (such as optical flow fields, texture gradients, etc.) is sufficient for the actor to directly gather information about affordances without the need for cognitive mediation, internal representation construction, or inference processes. What the actor directly perceives are the meaning and possibilities of action in the environment, not just physical properties. For example, an extended, solid, horizontal surface is perceived as "supportive for walking." This perception is immediate and guides behavior.

Animal-Environment (Mutuality): Animals and their environment form an integrated and inseparable system. Affordances are neither purely objective attributes of the environment nor purely subjective constructs by the actor; they exist in the mutual relationship between animals and their surroundings. Gibson stated: "An affordance points both to the environment and to the observer." It is both a fact of the environment and a fact of behavior. It means that the definition of the environment is inseparable from animals, just as the survival and behavior of animals are inseparable from their environment. Together, they define the existence of affordances.

Affordance, Capability, Intention: Whether an object is "graspable" hinges on the size and shape of the actor's hand. However, for an existing affordance, whether an actor perceives it, notices it, or utilizes it may be influenced by their current needs, intentions, or goals. Gibson posited that affordance, as an objective existence (relative to a specific actor's capabilities), is stable and does not change with the immediate shifts in the needs or intentions of the observer. For instance, a hammer always possesses the "hammerable" affordance for someone capable of using it, regardless of whether they currently need to drive nails. It is the needs of the actor

(such as repairing furniture) that make this existing affordance salient and selected for use.

The articulation and development of Gibson's theory was an ongoing process. He first introduced the concept of affordances in 1966 work "The Senses Considered as Perceptual Systems," but it wasn't until 1979 book "The Ecological Approach to Visual Perception" that the theory of affordances was more elaborated and developed. This evolutionary approach and presentation style, while making this theory highly suggestive, may also have led to some confusion and various interpretations among early researchers in their understanding and application of his concepts [14, 19, 20].

2.1.2 Operationalizing Affordances in Robotics

The action-centric perspective of affordance theory is valuable for autonomous agents that interact with the physical world. Consequently, the field of robotics has long been engaged in the process of interpreting, adapting, and formalizing Gibson's concepts to meet the practical needs of robot perception, decision making, and action [21]. In robotics, affordance typically refers to the actionable possibilities that objects or environments offer based on the specific capabilities of a robot (sensors, effectors, kinematics, and dynamics) and the current state (posture and task objectives) [22]. This transformation often involves learning affordances from experiential data or modeling and reasoning about them [23].

Research on robot affordances focuses on the following key characteristics: they are associated with specific objects, parts of objects, or environmental regions; they are closely related to robot effectors (such as grippers, wheels) and their physical and control capabilities; and they emphasize the connection with task completion and functional attributes, allowing the robot to reason about object substitution based on functional equivalence [24].

The core concept of "direct perception" in theory faces the challenge of being translated into computable models for robotics [21]. Most robotic systems pragmatically adopt computational models that rely on internal representations (such as scene graphs, object models, state-action-effect mappings) and learned mappings from perceptual inputs to action possibilities established through algorithms (like deep learning and reinforcement learning) [22].

The need for a computable framework prompted significant formalization efforts. A key early example is the work of Sahin et al., who approached the problem from a robot control perspective. They defined affordance as an "acquired relation" that exists between specific anticipated effects and particular combinations of an environmental entity and agent action. [25]. This perspective emphasizes that robots learn to understand affordances through interaction with the environment. Such formalization efforts often lead to structured representations for ease of computation. One common approach is to model an affordance as a tuple containing key elements such as (actor, object, action, effect).

This process of translating an abstract psychological theory into a practical operational framework is not unique to robotics. A parallel and highly influential adaptation occurred in the field of human-computer interaction (hci), where the work of Donald Norman has had a profound impact. Norman introduced the concept into the field of design, proposing important related ideas such as "perceived affordances" and "signifiers" [26, 27].

2.2 Affordance in Visual Language Navigation

2.2.1 The Comprehension Gap in VLN

Vision-Language Navigation requires embodied intelligent agents to perform navigation tasks in complex 3D environments based on natural language instructions. However, natural language instructions are often abstract and generalized, while the actions an agent can execute are low-level, concrete physical operations. This mismatch between high-level abstract instructions and

low-level specific actions leads to the so-called "comprehension gap." For example, in tasks like REVERIE [28], an instruction might be a generalized description such as "Bring me a spoon," rather than detailed step-by-step directions.

Affordance reasoning serves as a crucial supplementary mechanism, anchoring the abstract reasoning of large language models (LLMs) and vision-language models (VLMs) to the physical interaction possibilities within the environment. This constrains the output of the model to better align with the physical reality. In the context of VLN, it bridges perception of the agent, comprehension of language, and potential action choices, endowing referential objects in language instructions with an "actionable" dimension. For instance, by identifying affordances in the environment, such as walkable surfaces, the agent can map abstract language instructions (e.g., "Find the target") to specific physical actions (e.g., "Walk along this path" or "Grab this object").

Meanwhile, complex VLN tasks, such as transitioning from discrete to continuous environment navigation or evolving from simple to advanced ambiguous instructions, further widen the comprehension gap, thereby raising the demand for affordance understanding and even reasoning capabilities.

2.2.2 Classification and Role of Affordances in VLN

To bridge this comprehension gap, agents need to perceive and reason about affordances at different levels within the environment. Although affordance is a unified theory, in the research practice of VLN, it is often divided into several interrelated categories for ease of analysis and modeling.

The most fundamental concept is **Object Affordance**, which refers to the direct interactive possibilities provided by the object itself. For example, a microwave oven affords "being opened" due to its structure, while a cup affords "being picked up" or "being grasped" because of its shape. However, translating this concept into a computable model presents a core challenge: how to avoid oversimplifying and static definitions of affordance to meet the needs of different tasks. In early research, an object is often assigned only its most prominent affordance. For example, the most intuitive label for a chair is "is-sittable." However, this single-label approach overlooks the fact that in more complex task scenarios, the same chair could also afford "being pushed" to clear a path or "being used as a temporary stand" to place items. This neglect of the diverse, context-dependent affordances of objects is a key reason why early models struggled with handling broader contexts. DISCO proposes a new paradigm that explicitly includes affordance labels like "can be picked up" or "can be opened" in its scene representation to guide subsequent action planning [29].

Navigation Affordance refers to the possibilities for movement and observation provided by the environment when the field of view is extended to the entire surroundings. It encompasses two core concepts.

Traversability, which refers to the perception of areas or paths in the environment that allow an agent to move, serving as a fundamental prerequisite for navigation tasks. Traditional methods select nodes from predefined navigation graphs for high-level task planning while neglecting low-level motion control in continuous environments. In contrast, AO-Planner directly uses navigation affordance as a medium for LLM planning. It leverages the Grounded SAM model to segment visible ground in the field of view of the agent based on text prompts ("ground") and defines this as the "navigation affordance region." Within this region, candidate points are scattered to generate images with visual markers, and the task of LLM is to select waypoints from the 2D image [30].

Visibility refers to the understanding of new perspectives provided by specific locations in the environment. In unknown environments, an agent not only needs to know where it can

go but also must decide where it should go. VL-Nav introduces a heuristic visual-language spatial reasoning mechanism for selecting among candidate target points. These candidate points are divided into two categories: instance points based on detected relevant objects and frontier points representing opportunities to explore unknown areas. It proposes a curiosity-driven exploration mechanism to embody visibility affordance. When selecting target points, the semantic relevance between candidate points and language instructions is considered, along with a "curiosity score." This score guides the prioritization of exploring unknown areas most likely to contain the target, rather than blindly moving toward the nearest known object [31].

Effect Affordance refers to the potential of certain interaction points in the environment to trigger state changes or achieve specific goals. It not only describes "what can be done" but also further answers "what happens after doing it." This is crucial for interactive navigation in "cluttered" and realistic environments, where preset optimal paths may be blocked by obstacles. Wang et al. proposed an "Effect-Oriented Affordance Map," a multi-channel top-down grid map that extends traditional navigation frameworks to adapt to dynamic environments. Its effect affordance channel stores normalized continuous values to quantify the time cost required to remove obstacles at the current location. A set of affordance functions are used to learn and predict the effects of interactions, i.e., the time cost of removing different obstacles. The training data for these functions is generated through a self-experience-driven approach: the system records the actual time steps taken by the agent to remove obstacles in the simulator and uses this value as the "effect" ground truth label for the interaction, training the model in a supervised learning manner. The interactive strategy of the agent improves upon the classic Fast Marching Method, planning paths with the highest time efficiency rather than merely the shortest geometric distance, balancing between detouring and removing obstacles [32].

2.2.3 Affordance Learning and Representation in VLN

The core of how embodied intelligent agents learn to perceive and represent action possibilities lies in the deepening "Symbol Grounding" problem. Gibson's theory, a pre-computational philosophical concept, must link abstract notions like "is-sittable" to entities that robots can perceive or compute to be useful for robotics. Early attempts grounded affordances in geometric primitives; subsequent deep learning models like AffordanceNet grounded labels such as "graspable" in pixel distributions within RGB images, achieving data-driven perceptual grounding [33]; while more modern frameworks like SayCan simultaneously ground natural language instructions in the linguistic patterns of large language models and the embodied experience (i.e., the value functions of skills) [34]. In the survey proposed by Yang et al., affordances can be defined using Markov Decision Processes (MDPs) as a subset of the state-action space, where "intentions" transform into computable entities about the relationships between states, actions, and outcomes [35].

In the early stages of affordance learning and certain specific applications, the mainstream approach involved explicit representations. Systems employed dedicated modules to predict and output interpretable affordance information, most commonly in the form of pixel-level segmentation masks that could be utilized by downstream planning modules. A landmark work in this field is AffordanceNet, proposed by Do et al. It is an end-to-end deep learning model capable of simultaneously performing object detection and corresponding affordance segmentation from a single RGB image input. Its architecture is based on a shared CNN backbone, branching into two parallel paths for object detection and affordance detection, jointly optimized through a multi-task loss function. This formalized affordance learning as a computer vision task—an extension of instance segmentation [33]. Later, DISCO was designed for complex mobile manipulation tasks, with its core being the generation of dynamic and differentiable scene semantic representations. From a first-person perspective RGB image, a

perception system predicts depth, instance-level, and pixel-level affordance maps. These 2D results are projected onto a global 3D point cloud and further integrated into a 2D grid map from a bird-eye view. Crucially, this global map is optimized online via backpropagation at each time step. DISCO leverages these explicit, dynamic affordance maps to implement an efficient two-level control strategy: coarse-grained control uses the global map for long-range navigation planning, while fine-grained control performs precise interaction actions based on local observations when approaching the target [29]. The most significant advantage of explicit representations lies in their interpretability, yet they have fundamental limitations: first, large-scale, finely annotated affordance datasets are scarce; second, their capabilities are constrained by a predefined set of labels, making generalization difficult.

Implicit representation refers to the fact that affordance is no longer output as an independent module, but is implicitly encoded as prior knowledge in the network weights or value functions of an end-to-end model. SayCan, "Do As I Can, Not As I Say," aims to address the issue where LLMs possess vast semantic knowledge but lack grounding in the physical world [34]. Its core grounding mechanism involves scoring a series of low-level skills available to the robot, where the final score for each skill is the product of two probabilities.

1. Language probability p(text—instruction) (The "Say"): First query the LLM to evaluate the reasonableness of each skill's textual description as the next step under the current high-level instruction, which provides task grounding.

2. Affordance probability p(success—state) (The "Can"): Each skill is associated with a learned value function (or affordance function) that estimates the probability of executing the skill in the current state, which provides world grounding.

The question shifts from "Is that object graspable?" to "Is my 'grasping' skill capable of succeeding in the current state?" This agent-centric perspective aligns more closely with Gibson's theory than the explicit object-centric model. The primary advantage of implicit representation lies in its flexibility and scalability. The system is no longer constrained by predefined sets of affordance labels; as long as a new skill can be learned by the robot, it can be integrated into the system. Its main drawback is the potential sacrifice of interpretability. While it addresses the issue of fixed "affordance labels," it introduces the need for a well-defined vocabulary of "basic skills" [36].

Meanwhile, in the evolution of foundational large models, there has also emerged a prompt-driven approach—affordances as communicable instructions. The AO-Planner inputs images with marked points as Visual Affordances Prompting (VAP) to the LLM. The LLM's task is constrained to selecting sequences from these guaranteed reachable waypoints to form a path, avoiding the generation of unrealistic hallucinations by the LLM [30].

Transformer-based VIMA employs multimodal prompts to flexibly interweave textual and visual information. Here, affordances are no longer something the robot must detect but can be directly specified and communicated in task prompts. For example, a user can issue a task with the instruction "Arrange the blocks to look like a scene image," where the target image directly provides the affordances for the final placement state. The question shifts from "How to discover affordances?" to "How to most effectively communicate the desired affordances?" [37].

2.2.4 Applications and Challenges of VLMs in Affordance-based VLN

Large language/vision models (LLMs/VLMs) hold significant potential for affordance understanding due to their robust commonsense reasoning capabilities [38–42]. However, this potential is constrained by a core issue in their application to VLN: mainstream VLN datasets generally lack explicit affordance labels. For instance, the classic R2R dataset is structured as

(instruction, trajectory) pairs, originally designed to train agents to follow paths rather than to understand or annotate object interaction possibilities.

Current exploratory research primarily employs two approaches to leverage affordance in the absence of prior labels:

1. Dynamic zero-shot generation: The AO-Planner does not pre-annotate the dataset but instead dynamically segments traversable ground regions during inference using tools like Grounded SAM, serving as instant affordance labels to guide LLM-based planning [30].

2. Hierarchical control with implicit reasoning: The NaVILA framework adopts a hierarchical control structure, where a high-level VLM performs scene reasoning and generates mid-level language instructions (e.g., "move forward"), implicitly inferring affordances like "forward is traversable," while a low-level policy executes specific actions [43].

However, as VLN tasks grow more complex, the limitations of these methods in data density and generalization will become increasingly apparent. In fact, cutting-edge datasets like GSA-VLN (focused on long-term scene adaptation) and HA-VLN (incorporating human-agent interaction), while raising task difficulty, were not strictly designed for affordance research. This reveals a vast research gap in achieving affordance understanding within VLN tasks [6, 44].

3 Experimental Framework: Quantifying VLM Affordance Understanding

3.1 Experimental design

The previous sections systematically reviewed the theories, applications, and challenges of affordances in the field of embodied intelligence. However, there remains a lack of quantitative evaluation criteria for a core question: "To what extent do existing VLMs truly understand different levels of affordance?" To address this, this chapter proposes a comprehensive experimental evaluation framework.

This framework is not intended to introduce a new model, but rather to systematically explore the capabilities and limits of existing VLMs through a series of carefully designed tasks. Based on the preceding analysis, our experimental design follows a specific trajectory: it examines both the implicit affordance knowledge demonstrated by VLMs without external assistance and the performance gains when explicit affordance information (such as visual masks) is provided, ultimately assessing reasoning ability of the model regarding the advanced concept of effectual affordances. To achieve this goal, the framework first requires the construction of a new evaluation benchmark that supports diverse interactions.

3.2 Setup: Environment, Agent, and Task

Simulated Environment: The experiment is conducted in a large-scale procedurally generated simulation environment, includes 100,000 generated indoor floor plans to ensure scene diversity, and supports rich object interactions and dynamic changes [45].

Agent: We employ a single first-person-view agent whose movement is grid-based and discrete, with an action space that includes move forward, backward, turn left, and right. The agent can interact with objects within a 0.25-meter range directly in front of it, with actions such as picking up, opening, closing, and moving.

Task Design: The tasks range from basic perception to dynamic planning and are divided into three levels.

1. **Object Affordance** This task evaluates the fundamental understanding of the static and dynamic properties of an object according to Gibson theory. The model answers a series of questions about a specified object, covering its inherent attributes ("Can this object be moved?") and dynamic attributes ("Is this door currently open or closed?", "Can I pick up this object now?").

2. **Navigation Affordance** This task assesses navigation planning ability in navigation, that is, understanding the effect affordance of "How should I act to reach the goal?" In an interference-free scenario, the model must select the best single-step movement direction for the instruction "Go to [target] (the door)" based on different combinations of three input modalities, vision only, vision + depth map and vision + visual markers.

3. **Effect Affordance** This task is the test of end-to-end capabilities of re-planning and obstacle handling in dynamic environment. On a preset navigation path, an obstacle (a movable chair or an immovable wall) is suddenly placed in front of the agent. The experiment evaluates whether the model can recognize the path blockage and adopt different coping strategies (detouring or moving the obstacle) based on whether the obstacle is movable.

3.3 Evaluation method

This evaluation will select a series of pre-trained VLMs that can be inferred offline in a certain domain for testing, including Gemma3 of different scales (such as 4b, 12b, 27b), Moondream, Qwen2.5-VL, Granite-3.2, Pixtral 12B, TARS 7B, Mistral Small series, Minicpm-o-2.6 and InternVL [46–51]. In order to systematically verify the role of affordance information, we will set four experimental conditions for each model: implicit affordance, where the model only relies on its end-to-end capabilities; visual affordance input, where visual cues are provided to the model, such as ground masks superimposed with "passable areas"; language affordance input, where high-level language cues are provided through Prompt; and combined input, where both visual and language affordance information are provided.

For different types of tasks, we will use the corresponding evaluation indicators for a precise quantification. For task 1 (object affordance recognition), the evaluation will mainly use accuracy and F1 score to measure the accuracy of the model's judgment of object attributes. For the more complex Tasks 2 and 3 (planning and replanning), we will use a series of standard indicators in the navigation field, including the task success rate (SR), the path length weighted success rate (SPL) that comprehensively considers efficiency, the first-step accuracy (FSA) that directly reflects the effectiveness of affordance information, and the plan modification rate (PMR) that measures the model's adaptability in dynamic environments.

The experimental results will be analyzed in depth from the following core dimensions to answer the key questions of this study. First, by comparing performance under different input conditions, we will quantify the role of affordance. Second, by analyzing the performance of the baseline model on various tasks, we will explore the boundaries of the current end-to-end capabilities. Third, we will compare model architectures to analyze the advantages and disadvantages of different VLMs in dealing with affordance issues. Finally, we will conduct an in-depth analysis of typical failure cases to summarize the specific reasons for model failure. This experimental framework also has anticipated limitations, including the Sim-to-Real Gap, the inherent inability of the task to cover all real-world complexities, and the performance related to prompt engineering.

4 Conclusion

Starting from the origin of affordance theory, this review systematically combs its evolution in the field of embodied intelligence: from a psychological concept to formalization and operationalization in robotics; from focusing on the affordance of a single object to understanding the affordance of navigation in the environment, and then to the affordance of effect to foresee the consequences of actions. We further analyze the application of affordance in visual language navigation tasks, and point out the dilemma and two coping paths faced by current research due to the lack of affordance labels in mainstream datasets.

Through an analysis of the existing literature, we observe that there are several trade-offs in this field. The first is the trade-off between explicit and implicit representations: explicit methods (such as generating affordance masks) have high interpretability but limited generalization ability; while implicit methods (such as encoding knowledge in value functions) are highly flexible but the model is opaque. The second is the trade-off between the power and dilemma of large models: VLM brings powerful common sense reasoning capabilities, but is also accompanied by the risk of "hallucination" and the problem of symbol grounding. These unresolved challenges and trade-offs highlight the need for a rigorous and quantitative evaluation of current model capabilities.

To this end, we propose a three-level progressive experimental evaluation framework to systematically explore the capabilities of existing VLMs at different levels of affordance. We believe that a standardized evaluation benchmark is a key steptoward ultimately building a general intelligent agent that can understand and interact with the physical world.

References

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

[2] P. Gao, P. Wang, F. Gao, F. Wang, and R. Yuan, "Vision-language navigation with embodied intelligence: A survey," *arXiv preprint arXiv:2402.14304*, 2024.

[3] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[4] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 661–668.

[5] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," *Advances in neural information processing systems*, vol. 31, 2018.

[6] H. Hong, Y. Qiao, S. Wang, J. Liu, and Q. Wu, "General scene adaptation for vision-and-language navigation," *arXiv preprint arXiv:2501.17403*, 2025.

[7] P. Saxena, N. Raghuvanshi, and N. Goveas, "Uav-vln: End-to-end vision language guided navigation for uavs," *arXiv preprint arXiv:2504.21432*, 2025.

[8] R. Liu, W. Wang, and Y. Yang, "Volumetric environment representation for vision-language navigation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 317–16 328.

[9] ——, "Vision-language navigation with energy-based policy," *arXiv preprint arXiv:2410.14250*, 2024.

[10] J. J. Gibson, "The theory of affordances:(1979)," in *The people, place, and space reader*. Routledge, 2014, pp. 56–60.

[11] B. Lin, Y. Nie, Z. Wei, J. Chen, S. Ma, J. Han, H. Xu, X. Chang, and X. Liang, "Navcot: Boosting llm-based vision-and-language navigation via learning disentangled reasoning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[12] Q. Gu, A. Kuwajerwala, S. Morin, K. M. Jatavallabhula, B. Sen, A. Agarwal, C. Rivera, W. Paul, K. Ellis, R. Chellappa *et al.*, "Conceptgraphs: Open-vocabulary 3d scene graphs for perception and planning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5021–5028.

[13] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[14] J. McGrenere and W. Ho, "Affordances: Clarifying and evolving a concept," in *Graphics interface*, vol. 2000, no. 1, 2000, pp. 179–186.

[15] W. H. Warren, "Perceiving affordances: visual guidance of stair climbing." *Journal of experimental psychology: Human perception and performance*, vol. 10, no. 5, p. 683, 1984.

[16] T. A. Stoffregen, "Affordances as properties of the animal-environment system," in *How Shall Affordances Be Refined?* Routledge, 2018, pp. 115–134.

[17] L. Lobo, M. Heras-Escribano, and D. Travieso, "The history and philosophy of ecological psychology," *Frontiers in Psychology*, vol. 9, p. 2228, 2018.

[18] A. J. Wells, "Gibson's affordances and turing's theory of computation," *Ecological psychology*, vol. 14, no. 3, pp. 140–180, 2002.

[19] J. J. Gibson, "The senses considered as perceptual systems." 1966.

[20] C. Blewett and W. Hugo, "Actant affordances: a brief history of affordance theory and a latourian extension for education technology research," *Critical Studies in Teaching and Learning*, vol. 4, no. 1, pp. 55–76, 2016.

[21] P. Ardón, E. Pairet, K. S. Lohan, S. Ramamoorthy, and R. Petrick, "Building affordance relations for robotic agents-a review," *arXiv preprint arXiv:2105.06706*, 2021.

[22] E. Renaudo, P. Zech, R. Chatila, and M. Khamassi, "Computational models of affordance for robotics," p. 1045355, 2022.

[23] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor, "Modeling affordances using bayesian networks," in *2007 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2007, pp. 4102–4107.

[24] M. Andries, R. O. Chavez-Garcia, R. Chatila, A. Giusti, and L. M. Gambardella, "Affordance equivalences in robotics: a formalism," *Frontiers in neurorobotics*, vol. 12, p. 26, 2018.

[25] E. Şahin, M. Cakmak, M. R. Doğar, E. Uğur, and G. Üçoluk, "To afford or not to afford: A new formalization of affordances toward affordance-based robot control," *Adaptive Behavior*, vol. 15, no. 4, pp. 447–472, 2007.

[26] D. A. Norman, *The psychology of everyday things.* Basic books, 1988.

[27] E. Tenner, "The design of everyday things by donald norman," *Technology and Culture*, vol. 56, no. 3, pp. 785–787, 2015.

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

[29] X. Xu, S. Luo, Y. Yang, Y.-L. Li, and C. Lu, "Disco: Embodied navigation and interaction via differentiable scene semantics and dual-level control," in *European Conference on Computer Vision*. Springer, 2024, pp. 108–125.

[30] J. Chen, B. Lin, X. Liu, L. Ma, X. Liang, and K.-Y. K. Wong, "Affordances-oriented planning using foundation models for continuous vision-language navigation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 22, 2025, pp. 23 568–23 576.

[31] Y. Du, T. Fu, Z. Chen, B. Li, S. Su, Z. Zhao, and C. Wang, "Vl-nav: Real-time vision-language navigation with spatial reasoning," *arXiv preprint arXiv:2502.00931*, 2025.

[32] X. Wang, Y. Liu, X. Song, Y. Liu, S. Zhang, and S. Jiang, "An interactive navigation method with effect-oriented affordance," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 16 446–16 456.

[33] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5882–5889.

[34] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," *arXiv preprint arXiv:2204.01691*, 2022.

[35] X. Yang, Z. Ji, J. Wu, and Y.-K. Lai, "Recent advances of deep robotic affordance learning: a reinforcement learning perspective," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 15, no. 3, pp. 1139–1149, 2023.

[36] E. Tong, A. Opipari, S. Lewis, Z. Zeng, and O. C. Jenkins, "Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding," *arXiv preprint arXiv:2404.11000*, 2024.

[37] Y. Jiang, A. Gupta, Z. Zhang, G. Wang, Y. Dou, Y. Chen, L. Fei-Fei, A. Anandkumar, Y. Zhu, and L. J. Fan, "Vima: General robot manipulation with multimodal prompts," *arXiv preprint arXiv:2210.03094*, 2022.

[38] Y. Qin, A. Sun, Y. Hong, B. Wang, and R. Zhang, "Navigatediff: Visual predictors are zero-shot navigation assistants," *arXiv preprint arXiv:2502.13894*, 2025.

[39] D. Goetting, H. G. Singh, and A. Loquercio, "End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering," *arXiv preprint arXiv:2411.05755*, 2024.

[40] D. Song, J. Liang, A. Payandeh, A. H. Raj, X. Xiao, and D. Manocha, "Vlm-social-nav: Socially aware robot navigation through scoring using vision-language models," *IEEE Robotics and Automation Letters*, 2024.

[41] Z. Yin, C. Cheng *et al.*, "Navigation with vlm framework: Go to any language," *arXiv preprint arXiv:2410.02787*, 2024.

[42] J. Chen, B. Lin, R. Xu, Z. Chai, X. Liang, and K.-Y. K. Wong, "Mapgpt: Map-guided prompting with adaptive path planning for vision-and-language navigation," *arXiv preprint arXiv:2401.07314*, 2024.

[43] A.-C. Cheng, Y. Ji, Z. Yang, Z. Gongye, X. Zou, J. Kautz, E. Bıyık, H. Yin, S. Liu, and X. Wang, "Navila: Legged robot vision-language-action model for navigation," *arXiv preprint arXiv:2412.04453*, 2024.

[44] H. Li, M. Li, Z.-Q. Cheng, Y. Dong, Y. Zhou, J.-Y. He, Q. Dai, T. Mitamura, and A. G. Hauptmann, "Human-aware vision-and-language navigation: Bridging simulation to reality with dynamic human interactions," *arXiv preprint arXiv:2406.19236*, 2024.

[45] M. Deitke, E. VanderBilt, A. Herrasti, L. Weihs, J. Salvador, K. Ehsani, W. Han, E. Kolve, A. Farhadi, A. Kembhavi, and R. Mottaghi, "ProcTHOR: Large-Scale Embodied AI Using Procedural Generation," in *NeurIPS*, 2022, outstanding Paper Award.

[46] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love *et al.*, "Gemma: Open models based on gemini research and technology," *arXiv preprint arXiv:2403.08295*, 2024.

[47] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A frontier large vision-language model with versatile abilities," *arXiv preprint arXiv:2308.12966*, vol. 1, no. 2, p. 3, 2023.

[48] G. V. Team, L. Karlinsky, A. Arbelle, A. Daniels, A. Nassar, A. Alfassi, B. Wu, E. Schwartz, D. Joshi, J. Kondic *et al.*, "Granite vision: a lightweight, open-source multimodal model for enterprise intelligence," *arXiv preprint arXiv:2502.09927*, 2025.

[49] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand *et al.*, "Mixtral of experts," *arXiv preprint arXiv:2401.04088*, 2024.

[50] S. Hu, Y. Tu, X. Han, C. He, G. Cui, X. Long, Z. Zheng, Y. Fang, Y. Huang, W. Zhao *et al.*, "Minicpm: Unveiling the potential of small language models with scalable training strategies," *arXiv preprint arXiv:2404.06395*, 2024.

[51] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 185–24 198.