Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions

Jing Gu[1] **Eliana Stefani**[1] **Qi Wu**[2] **Jesse Thomason**[3] **Xin Eric Wang**[1]
[1]University of California, Santa Cruz
[2]The University of Adelaide [3]University of Southern California
{jgu110,estefani,xwang366}@ucsc.edu
qi.wu01@adelaide.edu.au, jessetho@usc.edu

Abstract

A long-term goal of AI research is to build intelligent agents that can communicate with humans in natural language, perceive the environment, and perform real-world tasks. Vision-and-Language Navigation (VLN) is a fundamental and interdisciplinary research topic towards this goal, and receives increasing attention from natural language processing, computer vision, robotics, and machine learning communities. In this paper, we review contemporary studies in the emerging field of VLN, covering tasks, evaluation metrics, methods, etc. Through structured analysis of current progress and challenges, we highlight the limitations of current VLN and opportunities for future work. This paper serves as a thorough reference for the VLN research community.[1]

1 Introduction

Humans communicate with each other using natural language to issue tasks and request help. An agent that can understand human language and navigate intelligently would significantly benefit human society, both personally and professionally. Such an agent can be spoken to in natural language, and would autonomously execute tasks such as household chores indoors, repetitive delivery work outdoors, or work in hazardous conditions following human commands (bridge inspection; fire-fighting). Scientifically, developing such an agent explores how an artificial agent interprets natural language from humans, perceives its visual environment, and utilizes that information to navigate to complete a task successfully.

Vision-and-Language Navigation (VLN) (Anderson et al., 2018b; Chen et al., 2019; Thomason et al., 2019b) is an emerging research field that aims to build such an embodied agent that can communicate with humans in natural language and navigate in real 3D environments. VLN extends visual navigation in both simulated (Zhu et al., 2017; Mirowski, 2019) and real environments (Mirowski et al., 2018) with natural language communication. As illustrated in Figure 1, VLN is a task that involves the oracle (frequently a human), the agent, and the environment. The agent and the oracle communicate in natural language. The agent may ask for guidance and the oracle could respond. The agent navigates and interacts with the environment to complete the task according to the instructions received and the environment observed. Meanwhile, the oracle observes the environment and agent status, and may interact with the environment to help the agent.

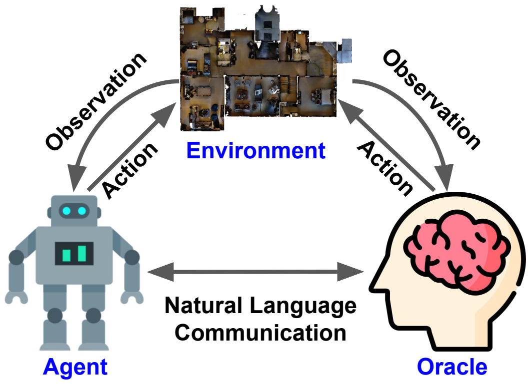

Figure 1: The agent and oracle discuss the VLN task in natural language. Both observe and interact with the navigable environment to accomplish a task.

Since the development and release of works such as Room-to-Room (R2R) (Anderson et al., 2018b), many VLN datasets have been introduced. Regarding the degree of communication, researchers create benchmarks where the agent is required to passively understand one instruction before navigation, to benchmarks where agents converse with the oracle in free-form dialog. Regarding the task objective, the requirements for the agent range from strictly following the route described in the ini-

[1]We also release a Github repo to keep track of advances in VLN: https://github.com/eric-ai-lab/awesome-vision-language-navigation

tial instruction to actively exploring the environment and interacting with objects. In a slight abuse of terminology, we refer to benchmarks that involve object interaction together with substantial sub-problems of navigation and localization, such as ALFRED (Shridhar et al., 2020), as VLN benchmarks.

Many challenges exist in VLN tasks. First, VLN faces a complex environment and requires effective understanding and alignment of information from different modalities. Second, VLN agents require a reasoning strategy for the navigation process. Data scarcity is also an obstacle. Lastly, the generalization of a model trained in seen environments to unseen environments is also essential. We categorize the solutions according to the respective challenges. (1) *Representation learning* methods help understand information from different modalities. (2) *Action strategy learning* aims to make reasonable decisions based on gathered information. (3) *Data-centric learning* methods effectively utilize the data and address data challenges such as data scarcity. (4) *Prior exploration* helps the model familiarize itself with the test environment, improving its ability to generalize.

We make three primary contributions. (1) We systematically categorize current VLN benchmarks from *communication complexity* and *task objective* perspectives, with each category focusing on a different type of VLN task. (2) We hierarchically classify current solutions and the papers within the scope. (3) We discuss potential opportunities and identify future directions.

2 Tasks and Datasets

The ability for an agent to interpret natural language instructions (and in some instances, request feedback during navigation) is what makes VLN unique from visual navigation (Bonin-Font et al., 2008). In Table 2, we mainly categorize current datasets on two axes, *Communication Complexity* and *Task Objective*.

Communication Complexity defines the level at which the agent may converse with the oracle, and we differentiate three levels: In the first level, the agent is only required to understand an *Initial Instruction* before navigation starts. In the second level, the agent sends a signal for help whenever it is unsure, utilizing the *Guidance* from the oracle. In the third level, the agent with *Dialogue* ability asks questions in the form of natural language during the navigation and understands further oracle guidance.

Task Objective defines *how* the agent attains its goal based on the *initial instructions* from the oracle. In the first objective type, *Fine-grained Navigation*, the agent can find the target according to a detailed step-by-step route description. In the second type, *Coarse-grained Navigation*, the agent is required to find a distant target goal with a coarse navigation description, requiring the agent to reason a path in a navigable environment and possibly elicit additional oracle help. Tasks in the previous two types only require the agent to navigate to complete the mission. In the third type, *Navigation and Object Interaction*, besides reasoning a path, the agent also needs to interact with objects in the environment to achieve the goal since the object might be hidden or need to change physical states.[2] As with coarse-grained navigation, some object interaction tasks can require additional supervision via dialogue with the oracle.

2.1 Initial Instruction

In many VLN benchmarks, the agent is given a natural language instruction for the whole navigation process, such as *"Go upstairs and pass the table in the living room. Turn left and go through the door in the middle."*

Fine-grained Navigation An agent needs to strictly follow the natural language instruction to reach the target goal. Anderson et al. (2018b) create the R2R dataset based on the Matterport3D simulator (Chang et al., 2017). An embodied agent in R2R moves through a house in the simulator traversing edges on a navigation graph, jumping to adjacent nodes containing panoramic views. R2R is extended to create other VLN benchmarks. Room-for-Room joins paths in R2R to longer trajectories (Jain et al., 2019). Yan et al. (2020) collect XL-R2R to extend R2R with Chinese instructions. RxR (Ku et al., 2020) contains instructions from English, Hindi, and Telegu. The dataset has more samples and the instructions in it are time-aligned to the virtual poses of the instruction. The English split of RxR is further extended to build Landmark-RxR (He et al., 2021) by incorporating landmark information.

In most current datasets, agents traverse a navigation graph at predefined viewpoints. To facil-

[2] Navigation and Object Interaction includes both fine-grained and coarse-grained instructions, which ideally should be split further. But given that there are only few datasets in this category, we keep the current categorization in Table 2.

Comm Complexity	Task Objective		
	Fine-grained Navigation	Coarse-grained Navigation	Nav + Object Interaction
Initial Instruction(s)	Room-to-Room (Anderson et al., 2018b), Room-for-Room (Jain et al., 2019), Room-Across-Room (Ku et al., 2020), XL-R2R (Yan et al., 2020), Landmark-RxR (He et al., 2021), VLNCE (Krantz et al., 2020), TOUCHDOWN (Chen et al., 2019), StreetLearn (Mirowski et al., 2019), StreetNav (Hermann et al., 2020), Talk2Nav (Vasudevan et al., 2021), LANI (Misra et al., 2018)	RoomNav (Wu et al., 2018), EmbodiedQA (Das et al., 2018), REVERIE (Qi et al., 2020b), SOON (Zhu et al., 2021a)	IQA (Gordon et al., 2018), CHAI (Misra et al., 2018), ALFRED (Shridhar et al., 2020)
Oracle Guidance	Just Ask (Chi et al., 2020)	VNLA (Nguyen et al., 2019), HANNA (Nguyen and Daumé III, 2019), CEREALBAR (Suhr et al., 2019)	None
Dialogue	None	CVDN (Thomason et al., 2019b), RobotSlang (Banerjee et al., 2020), Talk the Walk (de Vries et al., 2018)	TEACh (Padmakumar et al., 2021), Minecraft Collaborative Building (Narayan-Chen et al., 2019), DialFRED (Gao et al., 2022)

Table 1: Vision-and-Language Navigation benchmarks organized by **Communication Complexity** versus **Task Objective**. Please refer to Appendix for more details about the datasets and the commonly used underlying simulators.

itate transfer learning to real agents, VLN tasks should provide a continuous action space and a freely navigable environment. To this end, Krantz et al. (2020) reconstruct the navigation graph based R2R trajectories in continuous environments and create VLNCE. Irshad et al. (2021) propose Robo-VLN task where the agent operates in a continuous action space over long-horizon trajectories.

Outdoor environments are usually more complex and contain more objects than indoor environments. In TOUCHDOWN (Chen et al., 2019), an agent follows instructions to navigate a streetview rendered simulation of New York City to find a hidden object. Most photo-realistic outdoor VLN datasets including TOUCHDOWN (Chen et al., 2019), StreetLearn (Mirowski et al., 2019; Mehta et al., 2020), StreetNav(Hermann et al., 2020), and Talk2Nav (Vasudevan et al., 2021) are proposed based on Google Street View.

Some work uses natural language to guide drones. LANI (Misra et al., 2018) is a 3D synthetic navigation environment, where an agent navigates between landmarks following natural language instructions. Current datasets on drone navigation usually fall in a synthetic environment such as Unity3D (Blukis et al., 2018, 2019).

Coarse-grained Navigation In real life, detailed information about the route may not be available since it may be unknown to the human instructor (oracle). Usually, instructions are more concise and contain merely information of the target goal.

RoomNav (Wu et al., 2018) requires agent navigate according to instruction "*go to X*", where X is a predefined room or object.

In Embodied QA (Das et al., 2018), the agent navigates through the environment to find answer for a given question. The instructions in REVERIE (Qi et al., 2020b) are annotated by humans, and thus more complicated and diverse. The agent navigates through the rooms and differentiates the object against multiple competing candidates. In SOON (Zhu et al., 2021a), an agent receives a long, complex coarse-to-fine instruction which gradually narrows down the search scope.

Navigation+Object Interaction For some tasks, the target object might be hidden (e.g., the spoon in a drawer), or need to change status (e.g., a sliced apple is requested but only a whole apple is available). In these scenarios, it is necessary to interact with the objects to accomplish the task (e.g., opening the drawer or cutting the apple). Interactive Question Answering (IQA) requires the agent to navigate and sometimes to interact with objects to answer a given question. Based on indoor scenes in AI2-THOR (Kolve et al., 2017), Shridhar et al. (2020) propose the ALFRED dataset, where agents are provided with both coarse-grained and fine-grained instructions complete household tasks in an interactive visual environment. CHAI (Misra et al., 2018) requires the agent to navigate and simply interact with the environments.

2.2 Oracle Guidance

Agents in Guidance VLN tasks may receive further natural language guidance from the oracle during navigation. For example, if the agent is unsure of the next step (e.g., entering the kitchen), it can send a [help] signal, and the oracle would assist by responding *"go left"* (Nguyen et al., 2019).

Fine-grained Navigation The initial fine-grained navigation instruction may still be ambiguous in a complex environment. Guidance from the oracle could clarify possible confusion. Chi et al. (2020) introduce Just Ask—a task where an agent could ask oracle for help during navigation.

Coarse-grained Navigation With only a coarse-grained instruction given at the beginning, the agent tends to be more confused and spends more time exploring. Further guidance resolves this ambiguity. VNLA (Nguyen et al., 2019) and HANNA (Nguyen and Daumé III, 2019) both train an agent to navigate indoors to find objects. The agent could request help from the oracle, which responds by providing a subtask which helps the agent make progress. While oracle in VNLA uses predefined script to respond, the oracle in HANNA uses a neural network to generate natural language responses. CEREALBAR (Suhr et al., 2019) is a collaborative task between a leader and a follower. Both agents move in a virtual game environment to collect valid sets of cards.

Navigation+Object Interaction While VLN is still in its youth, there are no VLN datasets in support of Guidance and Object Interaction.

2.3 Human Dialogue

It is human-friendly to use natural language to request help (Banerjee et al., 2020; Thomason et al., 2019b). For example, when the agent is not sure about what fruit the human wants, it could ask *"What fruit do you want, the banana in the refrigerator or the apple on the table?"*, and the human response would provide clear navigation direction.

Fine-grained Navigation No datasets are in the scope of this category. Currently, route-detailed instruction with possible guidance could help the agent achieve relatively good performance in most simulated environments. We expect datasets to be developed for this category for super long horizon navigation tasks in complex environments especially with rich dynamics where dialog is necessary to clear confusions.

Coarse-grained Navigation CVDN (Thomason et al., 2019b) is a dataset of human-human dialogues. Besides interpreting a natural language instruction and deciding on the following action, the VLN agent also needs to ask questions in natural language for guidance. The oracle, with knowledge of the best next steps, needs to understand and correctly answer said questions.

Dialogue is important in complex outdoor environments. de Vries et al. (2018) introduce the Talk the Walk dataset, where the guide has knowledge from a map and guides the tourist to a destination, but does not know the tourist's location; while the tourist navigates a 2D grid via discrete actions.

Navigation+Object Interaction Minecraft Collaborative Building (Narayan-Chen et al., 2019) studies how an agent places blocks into a building by communicating with the oracle. TEACh (Padmakumar et al., 2021) is a dataset that studies object interaction and navigation with free-form dialog. The follower converses with the commander and interacts with the environment to complete various house tasks such as making coffee. DialFRED (Gao et al., 2022) extends ALFRED (Shridhar et al., 2020) dataset by allowing the agent to actively ask questions.

3 Evaluation

Goal-oriented Metrics mainly consider the agent's proximity to the goal. The most intuitive is *Success Rate (SR)*, which measures how frequently an agent completes the task within a certain distance of the goal. *Goal Progress* (Thomason et al., 2019b) measures the reduction in remaining distance to the target goal. *Path Length (PL)* measures the total length of the navigation path. *Shortest-Path Distance (SPD)* measures the mean distance between the agent's final location and the goal. Since a longer path length is undesirable (increases duration and wear-and-tear on actual robots), *Success weighted by Path Length (SPL)* (Anderson et al., 2018a) balances both Success Rate and Path Length. Similarly, *Success weighted by Edit Distance (SED)* (Chen et al., 2019) compares the expert's actions/trajectory to the agent's actions/trajectory, also balancing SR and PL. *Oracle Navigation Error (ONE)* takes the shortest distance from any node in the path rather than just the last node, and *Oracle Success Rate (OSR)* measures whether any node in the path is within a threshold from the target location.

Path-fidelity Metrics evaluate to what extent an agent follows the desired path. Some tasks require the agent not only to find the goal location but also to follow specific path. Fidelity measures the matches between the action sequence in the expert demonstration and the action sequence in the agent trajectory. *Coverage weighted by LS (CLS)* (Jain et al., 2019) is the product of the *Path Coverage (PC)* and *Length Score (LS)* with respect to the reference path. It measures how closely an agent's trajectory follows the reference path. *Normalized Dynamic Time Warping (nDTW)* (Ilharco et al., 2019) softly penalizes deviations from the reference path to calculate the match between two paths. *Success weighted by normalized Dynamic Time Warping (SDTW)* (Ilharco et al., 2019) further constrains nDTW to only successful episodes to capture both success and fidelity.

4 VLN Methods

As shown in Figure 2, we categorize existing methods into *Representation Learning*, *Action Strategy Learning*, *Data-centric Learning*, and *Prior Exploration*. Representation learning methods help agent understand relations between these modalities since VLN involves multiple modalities, including vision, language, and action. Moreover, VLN is a complex reasoning task where mission results depend on the accumulating steps, and better action strategies help the decision-making process. Additionally, VLN tasks face challenges within their training data. One severe problem is scarcity. Collecting training data for VLN is expensive and time-consuming, and the existing VLN datasets are relatively small with respect to the complexity of VLN tasks. Therefore, data-centric methods help to utilize the existing data and create more training data. Prior exploration helps adapt agents to previously unseen environments, improving their ability to generalize, decreasing the performance gap between seen versus unseen environments.

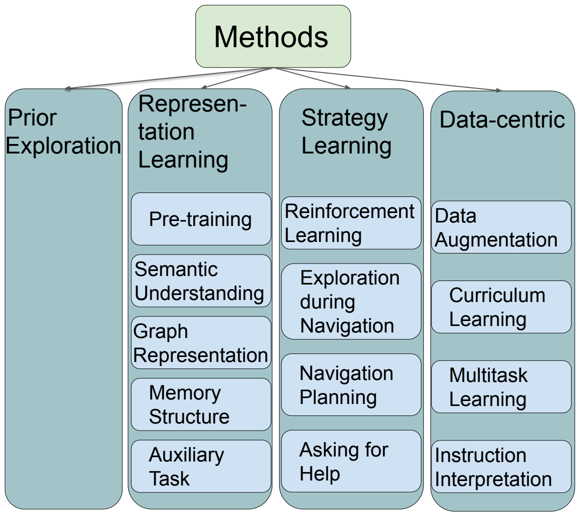

Figure 2: Categories of VLN methods. Methods may not be mutually exclusive to an individual category.

4.1 Representation Learning

Representation learning helps the agent understand how the words in the instruction relate to the perceived features in the environment.

4.1.1 Pretraining

Vision or Language Using a pretrained model to initialize a vision or text encoder provides agents with single-modality knowledge. pretrained vision models may use a ResNet (He et al., 2016) or Vision Transformers (Dosovitskiy et al., 2020). Other navigation tasks (Wijmans et al., 2019b) may also provide visual initialization (Krantz et al., 2020). Large pretrained language models such as BERT (Devlin et al., 2019) and GPT (Radford et al., 2019) can encode language and improve instruction understanding (Li et al., 2019), which can be further pretrained with VLN instructions (Pashevich et al., 2021) before fine-tuning in VLN task.

Vision and Language Vision-and-language pretrained models provide good joint representation for text and vision. A common practice is to initialize a VLN agent (Kim et al., 2021) with a pretrained model such as ViLBERT (Lu et al., 2019). The agent may be further trained with VLN-specific features such as objects and rooms (Qi et al., 2021).

VLN Downstream tasks benefit from being closely related to the pretraining task. Researchers also explored pretraining on the VLN domain directly. VLN-BERT (Majumdar et al., 2020) pretrains navigation models to measure the compatibility between paths and instructions, which formats VLN as a path selection problem. PREVALENT (Hao et al., 2020) is trained from scratch on image-text-action triplets to learn textual representations in VLN tasks. The output embedding from the [CLS] token in BERT-based pretraining models could be leveraged in a recurrent fashion to represent history state (Hong et al., 2021; Moudgil et al., 2021). Airbert (Guhur et al., 2021) achieve good performance on few-shot setting after pretraining on a large-scale in-domain dataset.

4.1.2 Semantic Understanding

Semantic understanding of VLN tasks incorporates knowledge about important features in VLN. In addition to the raw features, high-level semantic representations also improve performance in unseen environments.

Intra-Modality Visual or textual modalities can be decomposed into many features, which matter differently in VLN. The overall visual features extracted by a neural model may actually hurt the performance in some cases (Thomason et al., 2019a; Hu et al., 2019; Zhang et al., 2020b). Therefore, it is important to find the feature(s) that best improve performance. High-level features such as visual appearance, route structure, and detected objects outperform the low level visual features extracted by CNN (Hu et al., 2019). Different types of tokens within the instruction also function differently (Zhu et al., 2021c). Extracting these tokens and encoding the object tokens and directions tokens are crucial (Qi et al., 2020a; Zhu et al., 2021c).

Inter-Modality Semantic connections between different modalities: actions, scenes, observed objects, direction clues, and objects mentioned in instructions can be extracted and then softly aligned with attention mechanism (Qi et al., 2020a; Gao et al., 2021). The soft alignment also highlights relevant parts of the instruction with respect to the current step (Landi et al., 2019; Zhang et al., 2020a).

4.1.3 Graph Representation

Building graph to incorporate structured information from instruction and environment observation provides explicit semantic relation to guide the navigation. The graph neural network may encode the relation between text and vision to better interpret the context information (Hong et al., 2020a; Deng et al., 2020). The graph could record the location information during the navigation, which can used to predict the most likely trajectory (Anderson et al., 2019a) or probability distribution over action space (Deng et al., 2020). When connected with prior exploration, an overview graph about the navigable environment (Chen et al., 2021a) can be built to improve navigation interpretation.

4.1.4 Memory-augmented Model

Information accumulates as the agent navigates, which is not efficient to utilize directly. Memory structure helps the agent effectively leverage the navigation history. Some solutions leverage memory modules such as LSTMs or recurrently utilize informative states (Hong et al., 2021), which can be relatively easily implemented, but may struggle to remember features at the beginning of the path as path length increases. Another solution is to build a separate memory model to store the relevant information (Zhu et al., 2020c; Lin et al., 2021; Nguyen and Daumé III, 2019). Notably, by hierarchically encoding a single view, a panorama, and then all panoramas in history, HAMT (Chen et al., 2021b) successfully utilized the full navigation history for decision-making.

4.1.5 Auxiliary Tasks

Auxiliary tasks help the agent better understand the environment and its own status without extra labels. From the machine learning perspective, an auxiliary task is usually achieved in the form of an additional loss function. The auxiliary task could, for example, explain its previous actions, or predict information about future decisions (Zhu et al., 2020a). Auxiliary tasks could also involve the current mission such as current task accomplishment, and vision & instruction alignment (Ma et al., 2019a; Zhu et al., 2020a). Notably, auxiliary tasks are effective when adapting pretrained representations for VLN (Huang et al., 2019b).

4.2 Action Strategy Learning

With many possible action choices and complicated environment, action strategy learning provides a variety of methods to help the agent decide on those best actions.

4.2.1 Reinforcement Learning

VLN is a sequential decision-making problem and can naturally be modeled as a Markov decision process. So Reinforcement Learning (RL) methods are proposed to learn better policy for VLN tasks. A critical challenge for RL methods is that VLN agents only receive the success signal at the end of the episode, so it is difficult to know which actions to attribute success to, and which to penalize. To address the ill-posed feedback issue, Wang et al. (2019, 2020c) propose RCM model to enforces cross-modal grounding both locally and globally, with goal-oriented extrinsic reward and instruction-fidelity intrinsic reward. He et al. (2021) propose to utilize the local alignment between the instruction and critical landmarks as the reward. Evaluation metrics such as CLS (Jain et al., 2019) or nDTW (Ilharco et al., 2019) can also provide informative reward signal (Landi et al., 2020), and natural language may also provide suggestions for reward (Fu et al., 2019).

To model the dynamics in the environment, Wang et al. (2018) leverage model-based reinforcement learning to predict the next state and improve the generalization in unseen environment. Zhang et al. (2020a) find recursively alternating the learning schemes of imitation and reinforcement learning improve the performance.

4.2.2 Exploration during Navigation

Exploring and gathering environmental information while navigating provides a better understanding of the state space. Student-forcing is a frequently used strategy, where the agent keeps navigating based on sampled actions and is supervised by the shortest-path action (Anderson et al., 2018b).

There is a tradeoff between exploration versus exploitation: with more exploration, the agent sees better performance at the cost of a longer path and longer duration, so the model needs to determine when and how deep to explore (Wang et al., 2020a). After having gathered the local information, the agent needs to decide which step to choose, or whether to backtrack (Ke et al., 2019). Notably, Koh et al. (2021) designed Pathdreamer, a visual world model to synthesize visual observation future viewpoints without actually looking ahead.

4.2.3 Navigation Planning

Planing future navigation steps leads to a better action strategy. From the visual side, predicting the waypoints (Krantz et al., 2021), next state and reward (Wang et al., 2018), generate future observation (Koh et al., 2021) or incorporating neighbor views (An et al., 2021) has proven effective. Recognizing and stopping at the correct location also reduces navigation costs (Xiang et al., 2020). The natural language instruction also contains landmarks and direction clues to plan detailed steps. Anderson et al. (2019b) predict the forthcoming events based on the instruction, which is used to predict actions with a semantic spatial map. (Kurita and Cho, 2020) formulates VLN as a generative approach where a language model is used to compute the distribution over all possible instructions. The instruction may also be used to tag navigation and interaction milestones which the agent needs to complete step by step (Raychaudhuri et al., 2021; Song et al., 2022).

4.2.4 Asking for Help

An intelligent agent asks for help when uncertain about the next action (Nguyen et al., 2021b). Action probabilities or a separately trained model (Chi et al., 2020; Zhu et al., 2021e; Nguyen et al., 2021a) can be leveraged to decide whether to ask for help. Using natural language to converse with the oracle covers a wider problem scope than sending a signal. Both rule-based methods (Padmakumar et al., 2021) and neural-based methods (Roman et al., 2020; Nguyen et al., 2021a) have been developed to build navigation agents with dialog ability. Meanwhile, for tasks (Thomason et al., 2019b; Padmakumar et al., 2021) that do not provide an oracle agent to answer question in natural language, researchers also need to build a rule-based (Padmakumar et al., 2021) or neural-based (Roman et al., 2020) oracle. DialFRED (Gao et al., 2022) uses a language model as an oracle to answer questions.

4.3 Data-centric Learning

Compared with previously discussed works that focus on building a better VLN agent structure, data-centric methods most effectively utilize the existing data, or create synthetic data.

4.3.1 Data Augmentation

Trajectory-Instruction Augmentation Augmented path-instruction pairs could be used in VLN directly. Currently the common practice is to train a speaker module to generate instructions given a navigation path (Fried et al., 2018). This generated data have varying quality (Zhao et al., 2021; Huang et al., 2019a). Therefore an alignment scorer (Huang et al., 2019b) or adversarial discriminator (Fu et al., 2020) can select high-quality pairs for augmentation. Style transfer module may also improve instruction quality via adapting instructions from the source domain (Zhu et al., 2021d).

Environment Augmentation Generating more environment data not only helps generate more trajectories, but also alleviates the problem of overfitting in seen environments. Randomly masking the same visual feature across different viewpoints (Tan et al., 2019) or simply splitting the house scenes and re-mixing them (Liu et al., 2021) could create new environments, which could further be used to generate more trajectory-instruction pairs (Fried et al., 2018). Training data may also be augmented by replacing some visual features with counterfactual ones (Parvaneh et al., 2020).

4.3.2 Curriculum Learning

Curriculum learning (Bengio et al., 2009) gradually increases the task's difficulty during the training process. The instruction length could be a metric for task difficulty. BabyWalk (Zhu et al., 2020b) keep increasing training samples' instruction length during the training process. Attributes from the trajectory may also be used to rank task difficulty. Zhang et al. (2021) rearrange the R2R dataset using the number of rooms each path traverses. They found curriculum learning helps smooth the loss landscape and find a better local optima.

4.3.3 Multitask Learning

Different VLN tasks can benefit from each other by cross-task knowledge transfer. Wang et al. (2020d) propose an environment-agnostic multitask navigation model for both VLN and Navigation from Dialog History tasks (Thomason et al., 2019b). Chaplot et al. (2020) propose an attention module to train a multitask navigation agent to follow instructions and answer questions (Wijmans et al., 2019a).

4.3.4 Instruction Interpretation

A trajectory instruction interpreted multiple times in different ways may help the agent better understand its objective. LEO (Xia et al., 2020) leverages and encodes all the instructions with a shared set of parameters to enhance the textual understanding. LWIT (Nguyen et al., 2021c) interprets the instructions to make it clear to interact with what class of objects. Shorter, and more concise instructions provide clearer guidance for the agent compared to longer, semantically entangled instructions, thus Hong et al. (2020b) breaks long instructions into shorter ones, allowing the agent to track progress and focus on each atomic instruction individually.

4.4 Prior Exploration

Good performance in seen environments often cannot generalize to unseen environments (Hu et al., 2019; Parvaneh et al., 2020; Tan et al., 2019). Prior exploration methods allow the agent to observe and adapt to unseen environments,[3] bridging the performance gap between seen and unseen environments.

Wang et al. (2019) introduce a self-supervised imitation learning to learn from the agent's own past, good behaviors. The best navigation path determined to align the instruction the best by a matching critic will be used to update the agent. Tan et al. (2019) leverage the testing environments to sample and augment paths for adaptation. Fu et al. (2020) propose environment-based prior exploration, where the agent can only explore a particular environment where it is deployed. When utilizing graph, prior exploration may construct a map or overview about the unseen environment to provide explicit guidance for navigation (Chen et al., 2021a; Zhou et al., 2021).

5 Related Visual-and-Language Tasks

This paper focuses on Vision-and-Language Navigation tasks with an emphasis on photo-realistic environments. 2D map may also be a uesful virtual environment for navigation tasks (Vogel and Jurafsky, 2010; Chen and Mooney, 2011; Paz-Argaman and Tsarfaty, 2019). Synthetic environments may also be a substitute for realistic environment (MacMahon et al., 2006; Blukis et al., 2020). Tellex et al. (2011) propose to instantiate a probabilistic graphical model for natural language commands in robotic navigation and mobile manipulation process.

In VLN, an agent needs to follow the given instruction and even ask for assistants in human language. An agent in Visual Navigation tasks is usually not required to understand information from textual modality. Visual Navigation (Zhu et al., 2021b) is a problem of navigating an agent from the current location to find the goal target. Researchers have achieved success in both simulated environments (Zhu et al., 2017; Mirowski, 2019) and real environments (Mirowski et al., 2018).

6 Conclusion and Future Directions

In this paper, we discuss the importance of VLN agents as a part of society, how their tasks vary as a function of communication level versus task objective, and how different agents may be evaluated. We broadly review VLN methodologies and categorize them. This paper only discusses these issues broadly at an introductory level. In reviewing these

[3] Thus prior exploration methods are not directly comparable with other VLN methods.

papers, we can see the immense progress that has already been made, as well as directions that this research topic can be expanded on.

Current methods usually do not explicitly utilize external knowledge such as objects and general house descriptions in Wikipedia. Incorporating knowledge also improves the interpretability and trust of embodied AI. Moreover, currently several navigation agents learn which direction to move and with what to interact, but there is a last-mile problem of VLN—how to interact with objects. Anderson et al. (2018b) asked whether a robot could learn to *"Bring me a spoon"*; new research may ask how a robot can learn to *"Pick up a spoon"*. The environments also lack diversity: most interior terrestrial VLN data consists of American houses, but never warehouses or hospitals: the places where these agents may be of most use.

Below we detail additional future directions:

Collaborative VLN Current VLN benchmarks and methods predominantly focus on tasks where only one agent navigates, yet complicated real-world scenarios may require several robots collaborating. Multi-agent VLN tasks require development in swarm intelligence, information communication, and performance evaluation. MeetUp! (Ilinykh et al., 2019) is a two-player coordination game where players move in a visual environment to find each other. VLN studies the relationship between the human and the environment in Figure 1, yet here humans are oracles simply observing (but not acting on) the environment. Collaboration between humans and robots is crucial for them to work together as teams (e.g., as personal assistants or helping in construction). Future work may target at collaborative VLN between multiple agents or between human and agents.

Simulation to Reality There is a performance loss when transferred to real-life robot navigation (Anderson et al., 2020). Real robots function in continuous space, but most simulators only allow agents to "hop" through a pre-defined navigation graph which is unrealistic for three reasons (Krantz et al., 2020). Navigation graphs assume: (1) perfect localization—in the real world it is a noisy estimate; (2) oracle navigation—real robots cannot "teleport" to a new node; (3) known topology—in reality an agent may not have access to a preset list of navigable nodes. Continuous implementations of realistic environments may contain patches of the images, be blurred, or have parallax errors, making them unrealistic. A simulation that is based on both a 3D model and realistic imagery could improve the match between virtual sensors (in simulation) and real sensors. Lastly, most simulators assume a static environment only changed by the agent. This does not account for other dynamics such as people walking or objects moving, nor does it account for lighting conditions through the day. VLN environments with probabilistic transition functions may also narrow the gap between simulation and reality.

Ethics & Privacy During both training and inference, VLN agents may observe and store sensitive information that can get leaked or misused. Effective navigation with privacy protection is crucially important. Relevant areas such as federated learning (Konečnỳ et al., 2016) or differential privacy (Dwork et al., 2006) could also be studied in VLN domain to preserve the privacy of training and inference environments.

Multicultural VLN VLN lacks diversity in 3D environments: most outdoor VLN datasets use Google Street View recorded in major American cities, but lacks data in developing countries. Agents trained on American data face potential generalization problems in other city or housing layouts. Future work should explore more diverse environments across multiple cultures and regions. Multilingual VLN datasets (Yan et al., 2020; Ku et al., 2020) could be good resources to study multicultural differences from the linguistic perspective.

Acknowledgement

We thank anonymous reviewers, Juncheng Li, Yue Fan, Tongzhou Jiang for their feedback.

References

Dong An, Yuankai Qi, Yan Huang, Qi Wu, Liang Wang, and Tieniu Tan. 2021. Neighbor-view enhanced model for vision and language navigation. *arXiv preprint arXiv:2107.07201*.

Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. 2018a. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*.

Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. 2019a. Chasing ghosts: Instruction following as bayesian state tracking. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Peter Anderson, Ayush Shrivastava, Devi Parikh, Dhruv Batra, and Stefan Lee. 2019b. Chasing ghosts: Instruction following as bayesian state tracking. *Advances in Neural Information Processing Systems*, 32:371–381.

Peter Anderson, Ayush Shrivastava, Joanne Truong, Arjun Majumdar, Devi Parikh, Dhruv Batra, and Stefan Lee. 2020. Sim-to-real transfer for vision-and-language navigation. In *Conference on Robot Learning (CoRL)*.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018b. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shurjo Banerjee, Jesse Thomason, and Jason J. Corso. 2020. The RobotSlang Benchmark: Dialog-guided robot localization and navigation. In *Conference on Robot Learning (CoRL)*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Valts Blukis, Nataly Brukhim, Andrew Bennett, Ross A. Knepper, and Yoav Artzi. 2018. Following high-level navigation instructions on a simulated quadcopter with imitation learning. In *Robotics: Science and Systems (RSS)*.

Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, and Yoav Artzi. 2019. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *Conference on Robot Learning (CoRL)*.

Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A. Knepper, and Yoav Artzi. 2020. Learning to map natural language instructions to physical quadcopter control using simulated flight. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 1415–1438. PMLR.

Francisco Bonin-Font, Alberto Ortiz, and Gabriel Oliver. 2008. Visual navigation for mobile robots: A survey. *Journal of intelligent and robotic systems*, 53(3):263–296.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*.

Devendra Singh Chaplot, Lisa Lee, Ruslan Salakhutdinov, Devi Parikh, and Dhruv Batra. 2020. Embodied multimodal multitask learning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*. International Joint Conferences on Artificial Intelligence Organization.

David Chen and Raymond Mooney. 2011. Learning to interpret natural language navigation instructions from observations. In *AAAI Conference on Artificial Intelligence*.

Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. 2019. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12530–12539.

Kevin Chen, Junshen K Chen, Jo Chuang, Marynel Vázquez, and Silvio Savarese. 2021a. Topological planning with transformers for vision-and-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11276–11286.

Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. 2021b. History aware multimodal transformer for vision-and-language navigation. *arXiv preprint arXiv:2110.13309*.

Ta-Chung Chi, Minmin Shen, Mihail Eric, Seokhwan Kim, and Dilek Hakkani-tur. 2020. Just ask: An interactive learning framework for vision and language navigation. In *AAAI Conference on Artificial Intelligence*.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue.

Zhiwei Deng, Karthik Narasimhan, and Olga Russakovsky. 2020. Evolving graphical planner: Contextual global planning for vision-and-language navigation. *Advances in Neural Information Processing Systems*, 2020-December.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer.

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

Justin Fu, Anoop Korattikara, Sergey Levine, and Sergio Guadarrama. 2019. From language to goals: Inverse reinforcement learning for vision-based instruction following. *arXiv preprint arXiv:1902.07742*.

Tsu-Jui Fu, Xin Eric Wang, Matthew Peterson, Scott Grafton, Miguel Eckstein, and William Yang Wang. 2020. Counterfactual vision-and-language navigation via adversarial path sampler. In *European Conference on Computer Vision (ECCV)*.

Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. 2021. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *arXiv preprint arXiv:2202.13330*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098.

Pierre-Louis Guhur, Makarand Tapaswi, Shizhe Chen, Ivan Laptev, and Cordelia Schmid. 2021. Airbert: In-domain pretraining for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1634–1643.

Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. 2020. Towards learning a generic agent for vision-and-language navigation via pretraining. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Keji He, Yan Huang, Qi Wu, Jianhua Yang, Dong An, Shuanglin Sima, and Liang Wang. 2021. Landmark-rxr: Solving vision-and-language navigation with fine-grained alignment supervision. In *NeurIPS*.

Karl Moritz Hermann, Mateusz Malinowski, Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, and Raia Hadsell. 2020. Learning to follow directions in street view. In *AAAI Conference on Artificial Intelligence*.

Yicong Hong, Cristian Rodriguez, Yuankai Qi, Qi Wu, and Stephen Gould. 2020a. Language and visual entity relationship graph for agent navigation. *Advances in Neural Information Processing Systems*, 33:7685–7696.

Yicong Hong, Cristian Rodriguez, Qi Wu, and Stephen Gould. 2020b. Sub-instruction aware vision-and-language navigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3360–3376, Online. Association for Computational Linguistics.

Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1643–1653.

Ronghang Hu, Daniel Fried, Anna Rohrbach, Dan Klein, Trevor Darrell, and Kate Saenko. 2019. Are you looking? grounding to multiple modalities in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6551–6557, Florence, Italy. Association for Computational Linguistics.

Haoshuo Huang, Vihan Jain, Harsh Mehta, Jason Baldridge, and Eugene Ie. 2019a. Multi-modal discriminative model for vision-and-language navigation. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 40–49.

Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhaes, Jason Baldridge, and Eugene Ie. 2019b. Transferable representation learning in vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Gabriel Ilharco, Vihan Jain, Alexander Ku, Eugene Ie, and Jason Baldridge. 2019. General evaluation for instruction conditioned navigation using dynamic time warping. *arXiv preprint arXiv:1907.05446*.

Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. Meetup! a corpus of joint activity dialogues in a visual environment. *arXiv preprint arXiv:1907.05084*.

Muhammad Zubair Irshad, Chih-Yao Ma, and Zsolt Kira. 2021. Hierarchical cross-modal agent for robotics vision-and-language navigation. *arXiv preprint arXiv:2104.10674*.

Vihan Jain, Gabriel Magalhaes, Alexander Ku, Ashish Vaswani, Eugene Ie, and Jason Baldridge. 2019.

Stay on the path: Instruction fidelity in vision-and-language navigation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1862–1872, Florence, Italy. Association for Computational Linguistics.

Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. 2019. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hyounghun Kim, Jialu Li, and Mohit Bansal. 2021. Ndh-full: Learning and evaluating navigational agents on full-length dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6432–6442.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. 2021. Pathdreamer: A world model for indoor navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14738–14748.

Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. 2017. AI2-THOR: An Interactive 3D Environment for Visual AI. *arXiv*.

Jakub Konečný, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

Jacob Krantz, Aaron Gokaslan, Dhruv Batra, Stefan Lee, and Oleksandr Maksymets. 2021. Waypoint models for instruction-guided navigation in continuous environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15162–15171.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Alexander Ku, Peter Anderson, Roma Patel, Eugene Ie, and Jason Baldridge. 2020. Room-Across-Room: Multilingual vision-and-language navigation with dense spatiotemporal grounding. In *Conference on Empirical Methods for Natural Language Processing (EMNLP)*.

Shuhei Kurita and Kyunghyun Cho. 2020. Generative language-grounded policy in vision-and-language navigation with bayes' rule. In *International Conference on Learning Representations*.

Federico Landi, Lorenzo Baraldi, Marcella Cornia, Massimiliano Corsini, and Rita Cucchiara. 2020. Perceive, transform, and act: Multi-modal attention networks for vision-and-language navigation.

Federico Landi, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Embodied vision-and-language navigation with dynamic convolutional filters. In *Proceedings of the British Machine Vision Conference*.

Xiujun Li, Chunyuan Li, Qiaolin Xia, Yonatan Bisk, Asli Celikyilmaz, Jianfeng Gao, Noah Smith, and Yejin Choi. 2019. Robust navigation with language pretraining and stochastic sampling. In *Empirical Methods in Natural Language Processing (EMNLP)*.

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

Chong Liu, Fengda Zhu, Xiaojun Chang, Xiaodan Liang, Zongyuan Ge, and Yi-Dong Shen. 2021. Vision-language navigation with random environmental mixup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1644–1654.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chih-Yao Ma, Jiasen Lu, Zuxuan Wu, Ghassan AlRegib, Zsolt Kira, Richard Socher, and Caiming Xiong. 2019a. Self-monitoring navigation agent via auxiliary progress estimation. In *International Conference on Learning Representations (ICLR)*.

Chih-Yao Ma, Zuxuan Wu, Ghassan AlRegib, Caiming Xiong, and Zsolt Kira. 2019b. The regretful agent: Heuristic-aided navigation through progress estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Matt MacMahon, Brian Stankiewicz, and Benjamin Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. *Def*, 2(6):4.

Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Manolis Savva*, Abhishek Kadian*, Oleksandr Maksymets*, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. 2019. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Harsh Mehta, Yoav Artzi, Jason Baldridge, Eugene Ie, and Piotr Mirowski. 2020. Retouchdown: Releasing touchdown on StreetLearn as a public resource for language grounding tasks in street view. In *Proceedings of the Third International Workshop on Spatial Language Understanding*, pages 56–62, Online. Association for Computational Linguistics.

Piotr Mirowski. 2019. Learning to navigate. In *1st International Workshop on Multimodal Understanding and Learning for Embodied Applications*, MULEA '19, page 25, New York, NY, USA. Association for Computing Machinery.

Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, et al. 2019. The streetlearn environment and dataset. *arXiv preprint arXiv:1903.01292*.

Piotr Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. 2018. Learning to navigate in cities without a map. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2424–2435, Red Hook, NY, USA. Curran Associates Inc.

Dipendra Misra, Andrew Bennett, Valts Blukis, Eyvind Niklasson, Max Shatkhin, and Yoav Artzi. 2018. Mapping instructions to actions in 3d environments with visual goal prediction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2667–2678.

Abhinav Moudgil, Arjun Majumdar, Harsh Agrawal, Stefan Lee, and Dhruv Batra. 2021. Soat: A scene- and object-aware transformer for vision-and-language navigation. In *NeurIPS*.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy. Association for Computational Linguistics.

Khanh Nguyen, Yonatan Bisk, and Hal Daumé III au2. 2021a. Learning when and what to ask: a hierarchical reinforcement learning framework.

Khanh Nguyen, Yonatan Bisk, and Hal Daumé. 2021b. A framework for learning to request rich and contextually useful information from humans.

Khanh Nguyen and Hal Daumé III. 2019. Help, anna! visual navigation with natural multimodal assistance via retrospective curiosity-encouraging imitation learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 684–695, Hong Kong, China. Association for Computational Linguistics.

Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2019. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. 2021c. Look wide and interpret twice: Improving performance on interactive instruction-following tasks. *arXiv preprint arXiv:2106.00596*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramithu, Gokhan Tur, and Dilek Hakkani-Tur. 2021. Teach: Task-driven embodied agents that chat.

Amin Parvaneh, Ehsan Abbasnejad, Damien Teney, Qinfeng Shi, and Anton van den Hengel. 2020. Counterfactual vision-and-language navigation: Unravelling the unseen. In *NeurIPS*.

Alexander Pashevich, Cordelia Schmid, and Chen Sun. 2021. Episodic transformer for vision-and-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15942–15952.

Tzuf Paz-Argaman and Reut Tsarfaty. 2019. Run through the streets: A new dataset and baseline models for realistic urban navigation. *arXiv preprint arXiv:1909.08970*.

Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. 2021. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1655–1664.

Yuankai Qi, Zizheng Pan, Shengping Zhang, Anton van den Hengel, and Qi Wu. 2020a. Object-and-action aware model for visual language navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 303–317. Springer.

Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. 2020b. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sonia Raychaudhuri, Saim Wani, Shivansh Patel, Unnat Jain, and Angel Chang. 2021. Language-aligned waypoint (law) supervision for vision-and-language navigation in continuous environments. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4018–4028.

Homero Roman Roman, Yonatan Bisk, Jesse Thomason, Asli Celikyilmaz, and Jianfeng Gao. 2020. RMM: A recursive mental model for dialog navigation. In *Findings of Empirical Methods in Natural Language Processing (EMNLP Findings)*.

Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chan Hee Song, Jihyung Kil, Tai-Yu Pan, Brian M Sadler, Wei-Lun Chao, and Yu Su. 2022. One step at a time: Long-horizon vision-and-language navigation with milestones. *arXiv preprint arXiv:2202.07028*.

Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. 2017. Semantic scene completion from a single depth image. *CVPR*.

Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Alane Suhr, Claudia Yan, Jack Schluger, Stanley Yu, Hadi Khader, Marwa Mouallem, Iris Zhang, and Yoav Artzi. 2019. Executing instructions in situated collaborative interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2119–2130, Hong Kong, China. Association for Computational Linguistics.

Q. Sun, Y. Zhuang, Z. Chen, Y. Fu, and X. Xue. 2021. Depth-guided adain and shift attention network for vision-and-language navigation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.

Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. 2021. Habitat 2.0: Training home assistants to rearrange their habitat. *arXiv preprint arXiv:2106.14405*.

Hao Tan, Licheng Yu, and Mohit Bansal. 2019. Learning to navigate unseen environments: Back translation with environmental dropout. In *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)*, pages 2610–2621, Minneapolis, Minnesota. Association for Computational Linguistics.

Sinan Tan, Mengmeng Ge, Di Guo, Huaping Liu, and Fuchun Sun. 2022. Self-supervised 3d semantic representation learning for vision-and-language navigation. *arXiv preprint arXiv:2201.10788*.

Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew Walter, Ashis Banerjee, Seth Teller, and Nicholas Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI Conference on Artificial Intelligence*.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019a. Shifting the baseline: Single modality performance on visual navigation & QA. In *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)*, pages 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019b. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. 2021. Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory. *International Journal of Computer Vision*, 129(1):246–266.

Adam Vogel and Dan Jurafsky. 2010. Learning to follow navigational directions. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 806–814.

Hanqing Wang, Wenguan Wang, Wei Liang, Caiming Xiong, and Jianbing Shen. 2021. Structured scene memory for vision-language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8455–8464.

Hanqing Wang, Wenguan Wang, Tianmin Shu, Wei Liang, and Jianbing Shen. 2020a. Active visual information gathering for vision-language navigation. In *European Conference on Computer Vision (ECCV)*.

Hu Wang, Qi Wu, and Chunhua Shen. 2020b. Soft expert reward learning for vision-and-language navigation. In *European Conference on Computer Vision (ECCV'20)*.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the CVF/IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA. CVF/IEEE.

Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2020c. Vision-language navigation policy learning and adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4205–4216.

Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. 2018. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. *Proceedings of the European Conference on Computer Vision (ECCV 2018)*.

Xin Eric Wang, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. 2020d. Environment-agnostic multitask learning for natural language grounded navigation. In *European Conference on Computer Vision (ECCV'20)*.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019a. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Erik Wijmans, Abhishek Kadian, Ari Morcos, Stefan Lee, Irfan Essa, Devi Parikh, Manolis Savva, and Dhruv Batra. 2019b. Dd-ppo: Learning near-perfect pointgoal navigators from 2.5 billion frames. In *International Conference on Learning Representations*.

Yi Wu, Yuxin Wu, Georgia Gkioxari, and Yuandong Tian. 2018. Building generalizable agents with a realistic and rich 3d environment.

Fei Xia, Amir R. Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. 2018. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qiaolin Xia, Xiujun Li, Chunyuan Li, Yonatan Bisk, Zhifang Sui, Jianfeng Gao, Yejin Choi, and Noah A. Smith. 2020. Multi-view learning for vision-and-language navigation.

Jiannan Xiang, Xin Wang, and William Yang Wang. 2020. Learning to stop: A simple yet effective approach to urban vision-language navigation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 699–707.

An Yan, Xin Eric Wang, Jiangtao Feng, Lei Li, and William Yang Wang. 2020. Cross-lingual vision-language navigation.

Jiwen Zhang, Zhongyu Wei, Jianqing Fan, and Jiajie Peng. 2021. Curriculum learning for vision-and-language navigation. In *NeurIPS*.

Weixia Zhang, Chao Ma, Qi Wu, and Xiaokang Yang. 2020a. Language-guided navigation via cross-modal grounding and alternate adversarial learning. *IEEE Transactions on Circuits and Systems for Video Technology*.

Yubo Zhang, Hao Tan, and Mohit Bansal. 2020b. Diagnosing the environment bias in vision-and-language navigation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 890–897. International Joint Conferences on Artificial Intelligence Organization. Main track.

Ming Zhao, Peter Anderson, Vihan Jain, Su Wang, Alexander Ku, Jason Baldridge, and Eugene Ie. 2021. On the evaluation of vision-and-language navigation instructions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1302–1316.

Xinzhe Zhou, Wei Liu, and Yadong Mu. 2021. Rethinking the spatial route prior in vision-and-language navigation.

Fengda Zhu, Xiwen Liang, Yi Zhu, Qizhi Yu, Xiaojun Chang, and Xiaodan Liang. 2021a. Soon: Scenario oriented object navigation with graph-based exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12689–12699.

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

Fengda Zhu, Yi Zhu, Vincent Lee, Xiaodan Liang, and Xiaojun Chang. 2021b. Deep learning for embodied vision navigation: A survey. *arXiv preprint arXiv:2108.04097*.

Wang Zhu, Hexiang Hu, Jiacheng Chen, Zhiwei Deng, Vihan Jain, Eugene Ie, and Fei Sha. 2020b. BabyWalk: Going farther in vision-and-language navigation by taking baby steps. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2539–2556. Association for Computational Linguistics.

Wanrong Zhu, Yuankai Qi, Pradyumna Narayana, Kazoo Sone, Sugato Basu, Xin Eric Wang, Qi Wu, Miguel Eckstein, and William Yang Wang. 2021c. Diagnosing vision-and-language navigation: What really matters.

Wanrong Zhu, Xin Wang, Tsu-Jui Fu, An Yan, Pradyumna Narayana, Kazoo Sone, Sugato Basu, and William Yang Wang. 2021d. Multimodal text style transfer for outdoor vision-and-language navigation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1207–1221.

Yi Zhu, Yue Weng, Fengda Zhu, Xiaodan Liang, Qixiang Ye, Yutong Lu, and Jianbin Jiao. 2021e. Self-motivated communication agent for real-world vision-dialog navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1594–1603.

Yi Zhu, Fengda Zhu, Zhaohuan Zhan, Bingqian Lin, Jianbin Jiao, Xiaojun Chang, and Xiaodan Liang. 2020c. Vision-dialog navigation by exploring cross-modal memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10730–10739.

Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. 2017. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364. IEEE.

A Dataset Details

Here in Table 2, we introduce more information about the datasets. Compared with the number of the datasets, the simulators are limited. More specifically, most indoor datasets are based on Matterport3D and most outdoor datasets are based on Google Street View. Also, more datasets are about indoor environments rather than outdoor environments. Outdoor environments are usually more complex and contain more objects compared with indoor environments.

B Simulator

The virtual features of the dataset are deeply connected with the simulator in which datasets are built. Here we summarize simulators frequently used during the VLN dataset creation process.

House3D (Wu et al., 2018) is a realistic virtual 3D environment built based on the SUNCG (Song et al., 2017) dataset. An agent in the environment has access to first-person view RGB images, together with semantic/instance masks and depth information.

Matterport3D (Anderson et al., 2018b) simulator is a large-scale visual reinforcement learning simulation environment for research on embodied AI based on the Matterport3D dataset (Chang et al., 2017). Matterport3D contains various indoor scenes, including houses, apartments, hotels, offices, and churches. An agent can navigate between viewpoints along a pre-defined graph. Most indoors VLN datasets such as R2R and its variants are based on the Matterport3D simulator.

Habitat (Manolis Savva* et al., 2019; Szot et al., 2021) is a 3D simulation platform for training embodied AI in 3D physics-enabled scenarios. Compared with other simulation environments, Habitat 2.0 (Szot et al., 2021) shows strength in system response speed. Habitat has the following datasets built-in: Matterport3D (Chang et al., 2017), Gibson (Xia et al., 2018), and Replica (Straub et al., 2019). AI2-THOR (Kolve et al., 2017) is a near photo-realistic 3D indoor simulation environment, where agents could navigate and interact with objects. Based on the object interaction function, it helps to build a dataset that requires object interaction, such as ALFRED (Shridhar et al., 2020).

Gibson (Xia et al., 2018) is a real-world perception interactive environment with complex semantics. Each viewpoint has a set of RGB panoramas with global camera poses and reconstructed 3D meshes. Matterport3D dataset (Chang et al., 2017) is also integrated into the Gibson simulator.

House3D (Wu et al., 2018) converts SUNCG's static environment into a virtual environment, where the agent can navigate with physical constraints (e.g. it cannot pass through walls or objects).

LANI (Misra et al., 2018) is a 3D simulator built in Unity3D platform. The environment in LANI is a fenced, square, grass field containing randomly placed landmarks. An agent needs to navigate between landmarks following the natural language instruction. Drone navigation tasks (Blukis et al., 2018, 2019) are also built based on LANI.

Currently, most datasets and simulators focus on indoors navigable scenes partly because of the difficulty of building an outdoor photo-realistic 3D simulator out of the increased complexity. Google Street View [4], an online API that is integrated with Google Maps, is composed of billions of realistic street-level panoramas. It has been frequently used to create outdoor VLN tasks since the development of TOUCHDOWN (Chen et al., 2019).

C Room-to-Room Leaderboard

Room-to-Room (R2R) (Anderson et al., 2018b) is the benchmark used most frequently for evaluating different methods. Here we collect all the reported performance metrics in the corresponding papers and the official R2R leaderboard[5]. Since beam search explores more routes, and since prior exploration has additional observations in the test environment, their performance can not be directly compared with other methods.

[4] https://developers.google.com/maps/documentation/streetview/overview

[5] https://eval.ai/web/challenges/challenge-page/97/leaderboard/270

Name	Simulator	Language-Active	Environment
Room-to-Room (Anderson et al., 2018b)	Matterport3D	✗	Indoor
Room-for-Room (Jain et al., 2019)	Matterport3D	✗	Indoor
Room-Across-Room (Ku et al., 2020)	Matterport3D	✗	Indoor
Landmark-RxR (He et al., 2021)	Matterport3D	✗	Indoor
XL-R2R (Yan et al., 2020)	Matterport3D	✗	Indoor
VLNCE (Krantz et al., 2020)	Habitat	✗	Indoor
StreetLearn (Mirowski et al., 2019)	Google Street View	✗	Outdoor
StreetNav (Hermann et al., 2020)	Google Street View	✗	Outdoor
TOUCHDOWN (Chen et al., 2019)	Google Street View	✗	Outdoor
Talk2Nav (Vasudevan et al., 2021)	Google Street View	✗	Outdoor
LANI (Misra et al., 2018)	-	✗	Outdoor
RoomNav (Wu et al., 2018)	House3D	✗	Indoor
EmbodiedQA (Das et al., 2018)	House3D	✗	Indoor
REVERIE (Qi et al., 2020b)	Matterport3D	✗	Indoor
SOON (Zhu et al., 2021a)	Matterport3D	✗	Indoor
IQA (Gordon et al., 2018)	AI2-THOR	✗	Indoor
CHAI (Misra et al., 2018)	CHALET	✗	Indoor
ALFRED (Shridhar et al., 2020)	AI2-THOR	✗	Indoor
VNLA (Nguyen et al., 2019)	Matterport3D	✓	Indoor
HANNA (Nguyen and Daumé III, 2019)	Matterport3D	✓	Indoor
CEREALBAR (Suhr et al., 2019)	-	✓	Indoor
Just Ask (Chi et al., 2020)	Matterport3D	✓	Indoor
CVDN (Thomason et al., 2019b)	Matterport3D	✓	Indoor
RobotSlang (Banerjee et al., 2020)	-	✓	Indoor
Talk the Walk (de Vries et al., 2018)	-	✓	Outdoor
MC Collab (Narayan-Chen et al., 2019)	Minecraft	✓	Outdoor
TEACh (Padmakumar et al., 2021)	AI2-THOR	✓	Indoor
DialFRED (Gao et al., 2022)	AI2-THOR	✓	Indoor

Table 2: Vision-and-Language Navigation datasets. Language-Active means the agent needs to use natural language to request help, including both Guidance datasets and Dialog datasets in Table 1.

Simulator	Photo-realistic	3D
House3D (Wu et al., 2018)	✓	✓
Matterport3D (Chang et al., 2017)	✓	✓
Habitat (Manolis Savva* et al., 2019)	✓	✓
AI2-THOR (Kolve et al., 2017)	✗	✓
Gibson (Xia et al., 2018)	✓	✓
LANI (Misra et al., 2018)	✗	✓
*Google Street View	✓	✓

Table 3: Common simulators used to build VLN datasets. *Google Street View is online API, providing similar functionality as a simulator for building VLN datasets.

Leader-Board (Test Unseen)	Single Run					Prior Exploration					Beam Search				
Models	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑	TL↓	NE↓	OSR↑	SR↑	SPL↑
Random	9.89	9.79	0.18	0.13	0.12	-	-	-	-	-	-	-	-	-	-
Human	11.85	1.61	0.90	0.86	0.76	-	-	-	-	-	-	-			
Seq-to-Seq (Anderson et al., 2018b)	8.13	20.4	0.27	0.20	0.18	-	-	-	-	-	-	-	-	-	-
RPA (Wang et al., 2018)	9.15	7.53	0.32	0.25	0.23	-	-	-	-	-	-	-	-	-	-
Speaker-Follower (Fried et al., 2018)	14.82	6.62	0.44	0.35	0.28	-	-	-	-	-	1257.38	4.87	0.96	0.54	0.01
Chasing Ghosts (Anderson et al., 2019a)	10.03	7.83	0.42	0.33	0.30	-	-	-	-	-	-	-	-	-	-
Self-Monitoring (Ma et al., 2019a)	18.04	5.67	0.59	0.48	0.35	-	-	-	-	-	373.1	4.48	0.97	0.61	0.02
RCM !(Wang et al., 2019)	11.97	6.12	0.50	0.43	0.38	9.48	4.21	0.67	0.60	0.59	357.6	4.03	0.96	0.63	0.02
Regretful Agent (Ma et al., 2019b)	13.69	5.69	0.56	0.48	0.40	-	-	-	-	-	-	-	-	-	-
FAST (Ke et al., 2019)	22.08	5.14	0.64	0.54	0.41	-	-	-	-	-	196.5	4.29	0.90	0.61	0.03
ALTR (Huang et al., 2019b)	10.27	5.49	0.56	0.48	0.45	-	-	-	-	-	-		-	-	-
EnvDrop (Tan et al., 2019)	11.66	5.23	0.59	0.51	0.47	9.79	3.97	0.70	0.64	0.61	686.8	3.26	0.99	0.69	0.01
PRESS (Li et al., 2019)	10.52	4.53	0.63	0.57	0.53	-	-	-	-	-	-	-	-	-	-
PTA (Landi et al., 2020)	10.17	6.17	0.47	0.40	0.36	-	-	-	-	-	-	-			
EGP (Deng et al., 2020)	-	5.34	0.61	0.53	0.42	-	-	-	-	-	-	-	-	-	-
SERL (Wang et al., 2020b)	12.13	5.63	0.61	0.53	0.49	-	-	-	-	-	690.61	3.21	0.99	0.70	0.01
OAAM (Qi et al., 2020a)	10.40	-	0.61	0.53	0.50	-	-	-	-	-	-	-	-	-	-
CMG-AAL (Zhang et al., 2020a)	12.07	3.41	0.76	0.67	0.60	-	-	-	-	-		-	-	-	-
AuxRN (Zhu et al., 2020a)	-	5.15	0.62	0.55	0.51	10.43	3.69	0.75	0.68	0.65	40.85	3.24	0.81	0.71	0.21
RelGraph (Hong et al., 2020a)	10.29	4.75	0.61	0.55	0.52	-	-	-	-	-	-	-	-	-	-
PRRVALENT (Hao et al., 2020)	10.51	5.30	0.61	0.54	0.51	-	-	-	-	-	-	-	-	-	-
Active Exploration (Wang et al., 2020a)	21.03	4.34	0.71	0.60	0.43	9.85	3.30	0.77	0.70	0.68	176.2	3.07	0.94	0.70	0.05
VLN-BERT (Majumdar et al., 2020)	-	-	-	-	-	-	-	-	-	-	686.62	3.09	0.99	0.73	0.01
DASA (Sun et al., 2021)	10.06	5.11	-	0.54	0.52	-	-	-	-	-	-	-	-	-	-
ORIST (Qi et al., 2021)	11.31	5.10	-	0.57	0.52	-	-	-	-	-	-	-	-	-	-
NvEM (An et al., 2021)	12.98	4.37	0.66	0.58	0.54	-	-	-	-	-	-	-	-	-	-
SSM (Wang et al., 2021)	20.39	4.57	0.70	0.61	0.46	-	-	-	-	-	-	-	-	-	-
Recurrent VLN BERT (Hong et al., 2021)	12.35	4.09	0.70	0.63	0.57	-	-	-	-	-	-	-	-	-	-
SOAT (Moudgil et al., 2021)	12.26	-	4.49	58	53										
REM (Liu et al., 2021)	13.11	3.87	0.72	0.65	0.59	-	-	-	-	-	-	-	-	-	-
HAMT(Chen et al., 2021b)	12.27	3.93	0.72	0.65	0.60	-	-	-	-	-	-	-	-	-	-
Spatial Route Prior (Zhou et al., 2021)	-	-	-	-	-	-	-	-	-	-	625.27	3.55	0.99	0.74	0.01
Airbert (Guhur et al., 2021)	-	-	-	-	-	-	-	-	-	-	686.54	2.58	0.99	0.78	0.01
3DSR (Tan et al., 2022)	15.89	3.73	0.73	0.66	0.60	-	-	-	-	-	-	-	-	-	-

Table 4: Leaderboard of Room-to-Room benchmark as of March, 2022